Ecological Complexity xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Ecological Complexity

journal homepage: www.elsevier.com/locate/ecocom

Original Research Article

Circumventing structural uncertainty: A Bayesian perspective on nonlinear forecasting for ecology

Stephan B. Munch, Dr.[a,b,*], Valerie Poynor[c], Juan Lopez Arriaza[d]

[a] Fisheries Ecology Division, Southwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, 110 Shaffer Rd., Santa Cruz, CA 95060, United States
[b] Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, CA 95064, United States
[c] Department of Mathematics, California State University, Fullerton, CA 92831, United States
[d] Department of Applied Mathematics and Statistics, University of California, Santa Cruz, CA 95064, United States

ARTICLE INFO

Article history:
Received 9 March 2016
Received in revised form 24 August 2016
Accepted 30 August 2016
Available online xxx

Keywords:
Structural uncertainty
Nonlinear forecasting
Time-delay embedding
Gaussian process
Hierarchical model
Nonstationarity

ABSTRACT

As a consequence of the complexity of ecosystems and context-dependence of species interactions, structural uncertainty is pervasive in ecological modeling. This is particularly problematic when ecological models are used to make conservation and management plans whose outcomes may depend strongly on model formulation. Nonlinear time series approaches allow us to circumvent this issue by using the observed dynamics of the system to guide policy development. However, these methods typically require long time series from stationary systems, which are rarely available in ecological settings. Here we present a Bayesian approach to nonlinear forecasting based on Gaussian processes that readily integrates information from several short time series and allows for nonstationary dynamics. We demonstrate the utility of our modeling methods on simulated from a wide range of ecological scenarios. We expect that these models will extend the range of ecological systems to which nonlinear forecasting methods can be usefully applied.

Published by Elsevier B.V.

1. Introduction

Ecosystems involve a large and often unknown number of organisms and environmental factors. These components interact within and across species, age groups, genotypes, and environmental factors through time leading to systems that can be extremely complex. While it is possible to disentangle these sources of complexity for a handful of experimentally tractable, well-studied systems, identifying models for less-studied or intractable systems is a daunting task. Due to the system complexity, seemingly slight changes in model structure can lead to qualitatively different predictions (Wood and Thomas, 1999; Walters et al., 2016). This is particularly relevant when models are needed to inform conservation and management decisions (Lee et al., 1999; Wood and Thomas, 1999).

Alternatively, nonparametric time series methods allow us to study the dynamics of a system without having to specify a model. These methods originated in the 1980's and 90's based on Takens' (1981) theorem of time-delay embedding. Although initially restricted to single time series from an autonomous, deterministic system, these methods have since been generalized to multiple time series (Deyle and Sugihara, 2011) from non-autonomous systems with deterministic (Stark, 1999) and stochastic forcing (Stark et al., 2003). These methods have been of great use in physics (Buzug and Pfister, 1992), neurobiology (Kannathal et al., 2005), and econometrics (Mayfield and Mizrach, 1992) where long time series that are relatively free of observation noise are fairly common. Although a correctly specified parametric model is able to extract more information about the system (e.g. estimates of relevant parameters, reduced uncertainty), the insights gained from nonparametric methods tend to be robust to model misspecification.

Ecological applications of nonlinear forecasting were popular in the 1980's and 90's, including outstanding work by Sugihara (1994), Schaffer (1985), and Ellner and Turchin (1993); see Hastings et al. (1993) for a review. In the current literature, these methods seem to have been supplanted by more 'mechanistic' state-space models (see Patterson et al., 2008; and references therein) or linear models with time varying coefficients (e.g. Ives and Dakos, 2012).

* Corresponding author at: Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, CA 95064, United States.
E-mail address: steve.munch@noaa.gov (S.B. Munch).

http://dx.doi.org/10.1016/j.ecocom.2016.08.006
1476-945X/Published by Elsevier B.V.

The primary objections to using time-delay embedding in ecology seem to be that ecological time series are noisy, too short to define an attractor, and not stationary (Sugihara et al., 1990; Grenfell et al., 2001). Modern statistical methods, developed outside the time-delay embedding literature, may mitigate these objections. Hierarchical approaches allow information to be shared across data sets without assuming that they are identical (e.g., Shi et al., 2005; Bjornstad and Grenfell, 2001; Royle and Dorazio, 2008; Halstead et al., 2012). Nonstationary dynamics, in which the system drivers change through time, can be accommodated by allowing parameters to change as well (West and Harrison, 1997; Wikle, 2003; Ives and Dakos, 2012).

Here, we develop a Bayesian nonparametric framework for time-delay embedding that makes use of these modern statistical ideas. To do so, we use Gaussian process models to infer dynamics in delay coordinates. The chief advantages of the Gaussian process (GP) are its simple parameterization and ability to estimate with precision complicated nonlinear functions (O'Hagan, 1978). We then extend the GP framework to incorporate hierarchical inference from multiple time series and allow for nonstarionarity.

2. Methods and results

To begin, we briefly describe time delay embedding. We then introduce Gaussian processes as a tool for time-delay embedding and present a model specification that allows us to identify the relevant lags in the series. We extend this model to a hierarchical form that accommodates information from multiple related times series. Finally, we demonstrate how nonstationarity can be incorporated into the Gaussian process time-delay embedding model. These methods are then applied to a sequence of simulated data sets.

There is now a long history of applying Takens' theorem and time delay embedding in the ecological literature, see, e.g., Schaffer (1985), Ellner and Turchin (1993), and Sugihara (1994). However, most of the descriptions of the idea are steeped in the alien vernacular of topology. While there *are* deeper insights to be gained from the topological viewpoint, the practical upshot of Takens' theorem is that we are justified in modeling the dynamics of a single time series $y_t (t = 1, \ldots, T)$ as a function of its lags. That is, $y_t = f(y_{t-1}, \ldots y_{t-L})$ for some unknown function f and 'embedding dimension' L which is at least twice the dimension of the attractor (Takens, 1981). Here, a fixed time step of 1 is assumed in keeping with the majority of ecological time series applications. In settings where the data are continuously sampled through time, an appropriate time lag, Δ, must also be determined and the model is $y_t = f(y_{t-\Delta}, \ldots y_{t-L\Delta})$.

Various approaches to nonlinear forecasting can be thought of as approximating the unknown function f, including polynomials (e.g. Turchin and Ellner, 1995), support vector machines (e.g. Mukherjee et al., 1997), and neural networks (e.g. Bakker et al., 2000). A particularly useful way to approximate f is using locally-weighted multiple linear regression, as in Sugihara's S-Map (Sugihara 1994). Specifically, a locally linear model of the form $y_t = \sum_{i=1}^{L} \beta_{t,i} y_{t-i} + \varepsilon_t$ is fit to the time series by weighted least squares. We have highlighted this method in particular because it was precisely locally linear models that motivated O'Hagan (1978) to introduce Gaussian processes (GP) as priors for flexible regression modeling from a Bayesian point of view. Here, we use the tools of Bayesian GP regression to construct a hierarchical approach to nonlinear forecasting that allows integration of information from multiple time series and explicitly deals with nonstationarity. The main text lays out the model specification and the simulations used to test each model. Further details of prior specification and posterior inference are provided in the Appendices.

2.1. Gaussian process time-delay embedding

Assume we have a scalar time series $y_1, \ldots, y_T$, and the goal is to estimate the unknown function f that maps the history of y into the future. To simplify notation, we'll use $\mathbf{x}_t = \{y_{t-1}, \ldots, y_{t-L}\}$ to represent the 'delay-coordinate vector' so that we are attempting to fit a model of the form $y_t = f(x_t) + \varepsilon_t$ for $t \in \{L+1, \ldots, T\}$. The errors ε_t are explicitly included here to account for approximation errors as well as process noise. For convenience, we assume that ε_t is (at least approximately) normally distributed with mean 0 and variance V_ε.

The shape of the function f is unknown and we would like to estimate it from the available data. In a Bayesian context, we do so by assigning a prior to f and updating the distribution over f given the observed data. Since we are inferring a function, we need a prior on a space of functions and the natural place to look for these is the theory of stochastic process. The Gaussian process is particularly convenient to work with as a prior for uncertain regression functions (O'Hagan, 1978). GP models have been used widely in spatial statistics under the moniker Kriging (Cressie, 1993). In addition, they have been used in population modeling to estimate the form of density dependence (Munch et al., 2005), test for the presence of Allee effects (Sugeno and Munch, 2013), and as a tool to assess model misspecification (Thorson et al., 2014). Rasmussen and Williams (2006) is an excellent source for additional background on modeling with Gaussian processes.

The GP is a continuous generalization of the multivariate normal distribution and as such is completely defined in terms of a mean and covariance. However, because it is a distribution on a function space, the mean and covariance are functions as well, denoted by $\mu(w)$ and $\Sigma(w, w')$, respectively. Here w and w' denote two arbitrary 'inputs'. At a single input, the marginal distribution for $f(w)$ is Gaussian with mean $\mu(w)$ and variance $\Sigma(w, w)$. For any finite collection of input points, $w = \{w_1, \ldots, w_n\}^T$ (superscript T denotes transpose), the marginal distribution is multivariate normal with mean vector $\mu(w) = \{\mu(w_1), \ldots, \mu(w_n)\}^T$ and covariance matrix $\Sigma(w, w^T)$ {i.e. the covariance matrix is constructed by evaluating the covariance function at all pairs of inputs, such that the i,j^{th} element is $\Sigma(w_i, w_j)$}.

In the present application we set the mean function to zero, $\mu = 0$, to indicate that we do not have any *a priori* information on the shape of the function we want to infer. This is particularly the case for time-delay embedding where the 'true' function is bound to be something rather complicated. In other applications, such as modeling density dependence or population productivity, we can use standard parametric models as the prior mean and use the GP to infer model misspecification (see e.g. Thorson et al., 2014; Sugeno and Munch 2013).

Setting the mean function to zero means that the covariance function informs the shape of f by specifying how strongly correlated realizations of f are at different inputs. In general, the slower the correlation decays with increasing separation between inputs, the smoother realizations of f will be. There are many choices for the covariance function (see e.g. Rasmussen and Williams, 2006; Paciorek and Schervish, 2004). The squared exponential correlation function, $R(d) = exp[-d^2]$ where $d = w - w'$, is among the most widely used.

In the present application the 'inputs' are the delay coordinate vectors, $\mathbf{x}_t = \{y_{t-1}, \ldots, y_{t-L}\}$ and we need to specify the covariance between f evaluated at the delay coordinates for two different times, e.g. $f(\mathbf{x}_t)$ and $f(\mathbf{x}_s)$ for times t and s, respectively. We build

Please cite this article in press as: S.B. Munch, et al., Circumventing structural uncertainty: A Bayesian perspective on nonlinear forecasting for ecology, Ecol. Complex. (2016), http://dx.doi.org/10.1016/j.ecocom.2016.08.006

the covariance function for this L-dimensional input from the product of 1-dimensional lag-specific squared exponentials. Specifically, the covariance between $f(\mathbf{x}_t)$ and $f(\mathbf{x}_s)$ is given by $\Sigma(\mathbf{x_t}, \mathbf{x_s}) = \tau^2 \prod_{i=1}^{L} R(\phi_i | y_{t-i} - y_{s-i}|/r)$ where the factor $r = \max(y) - \min(y)$ scales the difference $(|y_{t-i} - y_{s-i}|/r)$ to stay in [0,1] and the ϕ_i's control the 'wiggliness' of f in the direction of the i^{th} time-lag (i.e. larger values of ϕ_i allow more local extrema on the interval [0,1]). The product is taken over all lags going from 1 to L. In practical applications, the maximum identifiable lag, L scales roughly as $\sqrt{T}$ where T is the length of the series (Chen and Tong, 1992) and will be less than 10 for all but the longest ecological time series. The parameter τ^2 controls the prior variance in f at a given point.

Note that when $\phi_i = 0$, the correlation in the i^{th} direction is 1. Hence f is constant in the i^{th} direction. Thus, to facilitate identification of a parsimonious model, we used a prior on ϕ that places most weight on $\phi_i = 0$. This approach has been taken in the machine learning literature where it is referred to as ARD, automatic relevance determination, (Neal, 1996). Here we use $p(\phi_i) = 2\exp[-\phi_i^2/\pi]/\pi$ which sets the expected number of local extrema to ~1 within the range of the data (See Appendix A for details). ARD typically employs a threshold value for ϕ_i to drop the i^{th} variable from the model, but using model selection criteria to choose among models with a range of embedding dimensions gives comparable results.

The fully specified GP model is then given by

$$\begin{aligned} p[y_t | f, x_t, V_\varepsilon] &\sim N\left(f(x_t), V_\varepsilon\right) \\ p[f | \tau^2, \phi] &\sim GP(0, \Sigma) \\ p[V_\varepsilon, \tau^2, \phi] & \end{aligned} \tag{1}$$

The final line represents prior specification for the variance and length scale parameters. These are detailed in Appendix A.

Full Bayesian inference for f and the hyperparameters can be obtained via MCMC or other methods (Rasmussen and Williams, 2006). However such a computationally intensive approach is impractical for our simulation study. Instead, we fix the hyperparameters at the MAP, maximum *a posteriori*, estimates (Rasmussen and Williams, 2006). To find the posterior mode, we use the R-prop algorithm developed for neural networks (Riedmeiller and Braun, 1993) because it is more stable for training GP models than other numerical optimizers (Nocedal and Wright, 1999; Blum and Reidmiller, 2013). Given the MAP estimates of the hyperparameters, the posterior for f is also a GP, $f|x, \mathbf{y}, \{\tau^2, \phi, V_\varepsilon\}_{\text{MAP}} \sim GP(m_c, \Sigma_c)$ where m_c and Σ_c are the posterior mean and covariance functions obtained using standard formulae for conditioning in multivariate normals (Appendix B provides details on updating).

In the context of dynamical modeling, it is often the case that we want to know what the equilibria are and to characterize their stability. Since f is uncertain we need to think instead about the distribution of plausible equilibria. Specifically, we seek the probability, $p(x^*)$, that the state x^* is a fixed point (i.e., $x^* = f(x^*, x^*, \ldots)$) and obtain this by evaluating the posterior density along the line where $x = f$. Given the hyperparameters, the distribution for f at the point x is just a normal density, hence $p(x^*)$ is simply

$$p(x^*) = \frac{q}{\sqrt{2\pi\Sigma_c(x^*, x^*)}} \exp\left\{ -\frac{1}{\Sigma_c(x^*, x^*)} [x^* - m_c(x^*)]^2 \right\} \tag{2}$$

where q is the normalization constant. Note that this is not a normal distribution in x^* because both m_c and Σ_c depend on x^*. This generalizes immediately to higher dimensions by evaluating $m_c(x^*, x^*, \ldots)$ and $\Sigma_c(x^*, x^*, \ldots)$.

To test the utility of the GP framework for identifying the embedding dimension, relevant lags, and plausible equilibria, we simulated data from models with $n_t = r n_{t-1} \exp[-n_{t-\Delta} + \varepsilon_t]$ where Δ ranged from 1 to 4 with corresponding values of $r = [8, 3.5, 2.5, 1.75]$ and $\varepsilon_t \sim N(0, 0.1)$. Importantly, the relevant lags are spelled out in this simulation model, making the evaluation of ARD lag selection totally unambiguous. This is in contrast to simulating a sequence of increasingly complex multi-species models, which provide no *a priori* way to determine which lags are

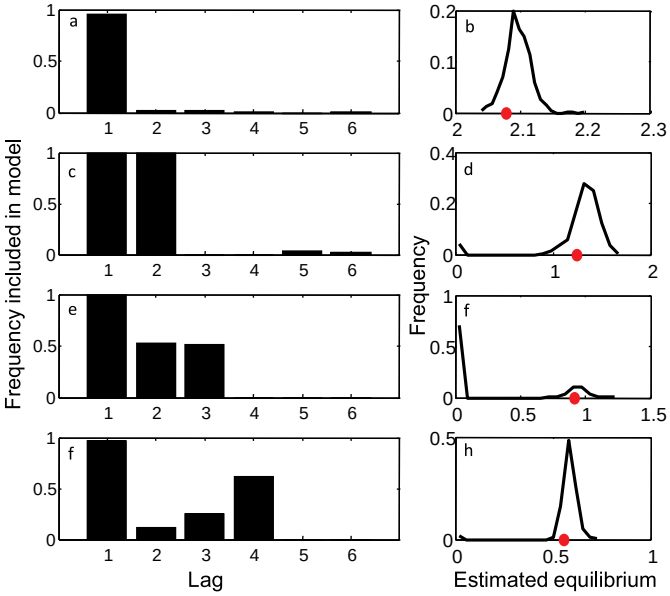

Fig. 1. Automatic relevance determination and putative equilibria. From top to bottom, rows present results for $\Delta = 1$–$\Delta = 4$. For each simulation, lag 1 is always relevant. Left: Bars indicate the frequency with which each lag coordinate was relevant in the fitted model (i.e. the fraction of simulations in which $\phi_i > 0.1$). Right: For each simulation, the posterior for plausible equilibria was calculated and the value with the maximum probability was recorded. Black lines show the distribution of these estimates across 1000 simulated data sets while the red dot is the true value for a fixed point for the deterministic skeleton. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Please cite this article in press as: S.B. Munch, et al., Circumventing structural uncertainty: A Bayesian perspective on nonlinear forecasting for ecology, Ecol. Complex. (2016), http://dx.doi.org/10.1016/j.ecocom.2016.08.006

most relevant except in the simplest cases. For each value of Δ, we simulated 1000 time series of length 100. For each simulated data set, we fit the GP model with $L = 6$ and used $\phi_i > 0.1$ as the threshold to determine if the i^{th} lag is relevant in the model. We then computed the distribution of plausible equilibria and selected the posterior mode as the best estimate for that data set.

Overall, the ARD approach seems to identify the relevant lags correctly a majority of the time (Fig. 1). The chief exception to this is when $\Delta = 3$, where lags 2 and 3 appear relevant with roughly equal frequency. The MAP estimates of the plausible equilibria are also generally quite close to the true value obtained from the deterministic skeleton (Fig. 1).

2.2. Hierarchical modeling

The most compelling reasons to formulate a Bayesian approach to time-delay embedding are the extensions that it invites. Here we extend the GP time-delay embedding model to accommodate multiple short time series. Previously, Hsieh et al. (2008) demonstrated that multiple short series can be concatenated to improve forecasts. However, this assumes that the attractor is identical for each series. Although we expect underlying similarities across series to be present (e.g. for the same species in several locations), we also expect population-specific features to exist. In order to account for this, we propose a hierarchical model structure that allows the degree of similarity between series to be determined from the data.

In the standard approach to hierarchical Bayesian models, information is shared across data sets by asserting that their parameters come from a common distribution (e.g., Shi et al., 2005; Bjornstad and Grenfell, 2001; Royle and Dorazio, 2008; Halstead et al., 2012). For example, we might improve estimation of the mean for under sampled populations by treating each population-specific mean as a random deviation from the cross-population mean.

In the present case we do something analogous, by proposing that the unknown map for the ith population, f_i, is a GP, drawn independently from a common, non-zero, mean function μ, i.e.$f_i \sim GP(\mu, \Sigma)$. Since the shape of μ is unknown, we model it with a GP as well, i.e. $\mu \sim GP(0, C)$. The covariance function C is defined analogously to Σ in (1) but with potentially different point-wise variance and length scale parameters, i.e., σ^2 and γ respectively.

With this specification, we can evaluate the 'correlation' between maps for two different populations. Note that this is not the correlation between time series, but instead measures the similarity between their reconstructed maps. Evaluated at the same state, the correlation is $Corr[f_i(x), f_j(x)] = C(x,x)/[C(x,x) + \Sigma(x,x)]$. The following parameterization simplifies this expression: Marginalizing over μ, the total point-wise variance in f is given by $\varpi^2 = \tau^2 + \sigma^2$. Introducing parameter $\rho \in [0,1]$ we can write $\sigma^2 = \rho\varpi^2$ for the among population variance in f and $\tau^2 = (1-\rho)\varpi^2$ for the within population variance. The correlation between maps,$Corr[f_i(x), f_j(x)]$, then reduces to ρ. All f' s are identical when $\rho = 1$ and independent when $\rho = 0$. In this way, information is shared across data sets without assuming that the dynamics are identical. The GP model for this case is given by

$$\begin{aligned} p[y_{it}|f_i, x_{it}, V_\varepsilon] &\sim N\left(f_i(x_t), V_{\varepsilon i}\right) \\ p[f_i|\theta] &\sim GP(\mu, \Sigma) \\ p[\mu|\theta] &\sim GP(0, C) \\ p[V_\varepsilon, \theta] & \end{aligned} \tag{3}$$

where $V_{\varepsilon i}$ is the population specific process variance and θ collects all of the other parameters. In this model we set a uniform prior, $U(0,1)$, on ρ.

To demonstrate the utility of this framework we simulated data from 12 models with 1–5 states (see Table 1a,b) and repeated these over 3 independent locations. These models represent a wide range of ecological phenomena including delayed regulation (e.g. Turchin, 1990), seasonal variation in productivity (e.g. Summers et al., 2000), density-dependent maturation (Neubert and Caswell, 2000), maternal effects (Ginzurg and Taneyhill, 1994), host-parasitoid interactions (Beddington et al., 1975), competition (e.g. Schoombie and Getz, 1998), contemporary evolution (Doebeli and de Jong, 1999), and migration (e.g. Gerber et al., 2002). When the models include multiple classes of individuals, we use time

Table 1a
Simulation models used to evaluate the GP embedding approach. For each model one parameter is varied to produce chaotic, limit cycle, and stable dynamics. For all models, except the delayed logistic, the noise term z_t is drawn independently from a normal distribution, $N(-s^2/2, s^2)$. Noise in the delayed logistic model must respect the requirement that population size N_t remain within [0,1] at all times, thus, we use a uniform distribution for the errors, $w_t \sim U[s(1-N_{t-\tau}), s(N_{t-\tau}-1)]$. In several of the models the noise term follows a '$\circ$',which indicates element-wise multiplication of the noise vector with the deterministic 'next-state' vector.

Scenario	Model	Parameter Values
Delayed regulation (Logistic)	$N_{t+1} = rN_t(1 + N_{t-\tau})(1 + w_t)$	$\tau = 1$ $r = \{2.1, 2.19, 1.54\}$
Delayed regulation (Hassel)	$N_{t+1} = rN_t(1 + N_{t-\tau})^{-b}e^{z_t}$	$r = 4, \tau = 1$ $b = \{3, 2, 1.75\}$
Seasonal productivity	$N_{t+1} = N_t e^{r[1+\alpha\sin(2\pi t/\theta)] - N_t + z_t}$	$r = 1.95$, $a = 3$ $\theta = \{4, 3, 8\}$
Density-dependent maturation	$\begin{bmatrix} A \\ J \end{bmatrix}_{t+1} = \begin{bmatrix} s_A & s_J g(A_t + J_t) \\ be^{z_t} & s_J[1 - g(A_t + J_t)] \end{bmatrix} \begin{bmatrix} A \\ J \end{bmatrix}_t$ $g(x) = G_{max}\exp(-\gamma x)$	$s_A = 0.1$, $s_J = 0.5$ $G_{max} = 0.9$, $\gamma = 0.01$ $b = \{35, 34, 24\}$
Host-parasitoid	$N_{t+1} = rN_t e^{-N_t - \gamma P_t + z_t}$ $P_{t+1} = \alpha N_t e^{-N_t + z_t}(1 - e^{-\gamma P_t})$	$\gamma = 0.5$, $\alpha = 2$ $r = \{7, 8, 5\}$
Competition (Hassell-Comins)	$N_{1,t+1} = r_1 N_{1,t} e^{z_{1,t}}[1 + (N_{1,t} + cN_{2,t})/\gamma]^{-\gamma}$ $N_{2,t+1} = r_2 N_{2,t} e^{z_{2,t}}[1 + (N_{2,t} + dN_{1,t})/\gamma]^{-\gamma}$	$c = 0.9$, $d = 0.8$ $\gamma = 20$, $r_1 = 1.5r_2$ $r_2 = \{16, 12, 5\}$
Competition (Shepherd)	$N_{1,t+1} = r_1 N_{1,t} e^{z_{1,t}}/\{1 + (r_1 - 1)[N_{1,t} + cN_{2,t})]^{\gamma}\}$ $N_{2,t+1} = r_2 N_{2,t} e^{z_{2,t}}/\{1 + (r_2 - 1)\left[N_{2,t} + dN_{1,t})\right]^{\gamma}$	$c = 0.2$, $d = 0.1$ $\gamma = 4$, $r_1 = 1.5r_2$ $r_2 = \{3, 2.4, 1.6\}$
Contemporary evolution	$\begin{bmatrix} N_{AA} \\ N_{Aa} \\ N_{aa} \end{bmatrix}_{t+1} = \begin{bmatrix} p_{AA} + p_{Aa} & p_{Aa}/4 & 0 \\ p_{aa} & p_{Aa}/2 + p_{aa} & p_{AA} \\ 0 & p_{Aa}/4 & p_{AA} + p_{Aa} \end{bmatrix} \begin{bmatrix} N_{AA}/(1 + ax^c) \\ N_{Aa}/(1 + bx^c) \\ N_{aa}/(1 + ax^c) \end{bmatrix} \circ e^{z_t}$ $x = \sum N, p_{ij} = N_{ij}/x$	$a = 4 * 8^{-c}$, $b = 0.9aC = \{4, 3.35, 2.5\}$

Table 1b
Continued. Simulation models with migration. In the 5-location migration models, the migration matrix, M, represents the exchange of individuals across sites. We used 3 different migration topologies representing a linear chain with reflective boundaries, a ring, and uniform dispersal. In each case the parameter μ' is the fraction of the resident population that does not migrate and the $(1-\mu')$ migrants are distributed equally over all accessible neighbors. The same parameters are used for all 5 sites. The noise term z_t is drawn independently from $N(-s^2/2, s^2)$ for each site.

Scenario	Model	Parameter Values
Migration (2 locations, Ricker)	$\begin{bmatrix} N_1 \\ N_2 \end{bmatrix}_{t+1} = \begin{bmatrix} m_{11} & (1-m_{22}) \\ (1-m_{11}) & m_{22} \end{bmatrix} \begin{bmatrix} N_1 e^{r_1 - N_1} \\ N_2 e^{r_2 - N_2} \end{bmatrix} \circ e^{z_t}$	$m_{11} = 0.35$, $m_{22} = 0.25$ $r_2 = 18$, $r_1 = \{15, 13, 7\}$
Migration (5 locations, Shepherd)	$\begin{bmatrix} N_1 \\ \vdots \\ N_5 \end{bmatrix}_{t+1} = M(\mu') \begin{bmatrix} rN_{1,t}/\{1+(r-1)N_{1,t}{}^{\gamma}\} \\ \vdots \\ rN_{5,t}/\{1+(r-1)N_{5,t}{}^{\gamma}\} \end{bmatrix}_t \circ e^{z_t}$	$\mu' = 0.9, \gamma = 4$ $r = \{2, 1.3, 1.2\}$ M − migration matrix, see caption

series from the same focal class for each location. Each model was simulated under 3 baseline parameter sets that generate fixed point, limit cycle, and chaotic dynamics respectively. Among populations, model parameters were allowed to vary from this baseline between 0 and 25%. In these simulations, we focused on short time series and evaluated the improvement in the out of sample forecast precision for series of 10, 15, and 20 points. Because the series are so short, the maximum lag (L) was set to 4. To quantify forecast precision, we produced 1-step ahead forecasts for the final 5 years of data for each location and computed the total forecast error as $SS = \sum_{i=1}^{3} \sum_{t=T-5}^{T} [y_{it} - f_i(x_{it})]^2$. To do so, we computed the MAP estimates of θ based on the first $T-6$ years of data, and conditioned on these to produce estimates for $T-5$ to T. For comparison, we repeated this calculation with each series modeled independently.

Fig. 2 illustrates the improvement over independent modeling for the density dependent maturation model. Overall, the hierarchical approach reduced forecast error in series of length 10 by $\sim 60\%$ on average, but this drops off quickly as T increases (Table 2). This is consistent with the observation that low dimensional dynamics (such as we simulated) can typically be reconstructed with $30-40$ points. Increasing process variance increased the fraction of simulations in which the hierarchical model outperformed the independent model, but diminished the average magnitude of error reduction. Interestingly, neither the dynamical regime nor the variance among populations had as strong an effect on the relative performance of the hierarchical model (Table 2).

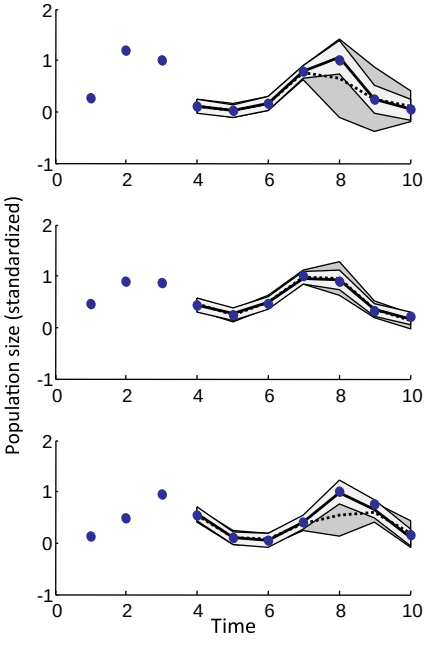

Fig. 2. An illustration of hierarchical GP forecasting. The time series are generated with the density-dependent maturation model (Table 1). Blue points in each panel represent time series for adult abundant in 3 independent realizations of the model with 10% variation among parameters. The solid black line (mean) and light grey region (±2s.d.) indicate 1-step ahead forecasts from the hierarchical model. For comparison, the dashed line and dark grey region indicate results for each population modeled independently. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Thus far, our simulations have focused on 1-step forecasts. However, multi-step forecasts are obviously desirable and it is important to evaluate their accuracy. In an independent set of simulations, we evaluated the utility of the GP framework for forecasts up to 10 steps into the future. To do so, we computed the MAP estimates of θ based on the first T years of data. The k-step ahead forecast model (for $k = 1, \ldots, 10$) was then obtained by iterating the estimated one-step ahead map, $E[f_i(x)|data, \theta_{MAP}]$, over k-steps. We then ran the simulation model forward k-steps and computed the scaled mean-square error as $S = \sum_{i=1}^{3} \{y_{it+k} - f_i^k(x_{it})\}^2 / var(y_i)$. We repeated this calculation with each series modeled independently for comparison. The same set of simulation models was used with the exception that the $T = 20$ treatment was omitted.

In many cases the GP framework produces useful forecasts up to 10 steps into the future (Fig. 3), though the forecast error and the rate at which it increases depend greatly on the dynamical regime, the process variance (s^2), and the length of the training data. Not surprisingly, forecasts in the chaotic regime degrade fastest while the simulations in the limit cycle regime are the best behaved. The $T = 10$ (Fig. 3a–c) and $T = 15$ (Fig. 3d–f) simulations differ substantially in both the total forecast error and the utility of the hierarchical approach. Forecast error is substantially greater when $T = 10$ as is the reduction in error due to the hierarchical approach. The effect of process variance is more obvious when $T = 15$: With $s^2 = 0$, forecast error is less than 50% up to 10 steps into the future for the chaotic simulations, and 10% for the limit

Table 2
Hierarchical GP performance. Error Red. is the proportional error reduction calculated as 1-mean(SS_{hier})/mean(SS_{ind}) and Freq. Imp. is the fraction of simulations in which the forecast error for the hierarchical model was less than for series modeled independently. These are reported for the main effects of dynamical regime, process variance, variation among populations, and the length of the time series. Performance metrics reported for each main effect are averaged over all models and other parameters.

	Dynamical Regime			Process Variance		
	Chaos	Limit Cycle	Stable	0	0.05	0.1
Error Red.	0.57	0.56	0.41	0.68	0.56	0.44
Freq. Imp.	0.84	0.81	0.73	0.76	0.81	0.81

	Variation among populations			Length of series		
	0	0.1	0.25	10	15	20
Error Red.	0.56	0.54	0.47	0.61	0.33	0.22
Freq. Imp.	0.82	0.8	0.75	0.87	0.78	0.73

Please cite this article in press as: S.B. Munch, et al., Circumventing structural uncertainty: A Bayesian perspective on nonlinear forecasting for ecology, Ecol. Complex. (2016), http://dx.doi.org/10.1016/j.ecocom.2016.08.006

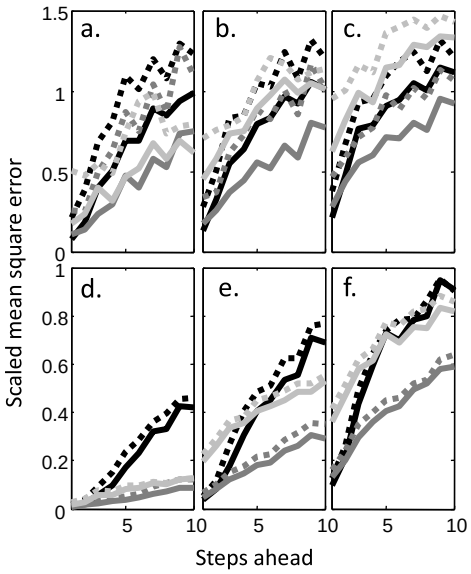

Fig. 3. Forecast error for hierarchical and independent GP models. Lines indicate the scaled mean-square error of prediction. A value of 1 indicates that the GP forecast is no better than the series mean and 0 indicates perfect prediction. Solid (dashed) lines represent results for the hierarchical (independent) model. The black, dark grey, and light grey lines correspond to simulations from the chaotic, limit cycle, and stable regimes respectively. The top and bottom rows correspond to T = 10 and T = 15 (note the change in vertical scale) while the process variance increases from left to right: (a,d):$s^2 = 0$, (b,e):$s^2 = 0.1$, (c,f):$s^2 = 0.2$.

cycles. But when the process variance was increased to $s^2 = 0.2$, the rate at which forecasts degrade increased roughly 6-fold.

2.3. Dynamic embedding for nonstationarity

It is widely recognized that for time-delay embedding to work well, long series are needed. Unfortunately, the longer the time period over which ecological data are collected, the more likely it is that some aspect of the system will change. Changes in climate, species introductions, and changes in management practices are common and recent studies suggest that nonstationary dynamics are present in a broad range of systems, e.g., multidecadal shifts in the dynamics of pacific ecosystems (Hare and Mantua, 2000). When using mechanistic models, a common approach to compensate for the apparent nonstationarity in a time series is to allow the parameters of the models to vary with time, either in a specified way or by letting them drift randomly (West and Harrison, 1997; Wikle, 2003; Ives and Dakos, 2012). This concept is the basis of the Dynamic Linear Model (DLM, West and Harrison, 1997) and has been used to identify regime shifts (Carpenter, 2003). In DLMs, the parameters of the model are assigned a random-walk prior, allowing the model to adapt to changes in the underlying ecosystem dynamics. For example, in the dynamic linear model $y_t = x_t^T b_t + \in_t$, the coefficient vector,b_t, changes according to the random walk $b_{t+1} = b_t + z_t$ where z_t is zero-mean multivariate normal noise. This specification is not intended to mechanistically represent dynamics in **b**, but provides the model with enough slack that variations in **b** through time can be inferred retrospectively.

The framework of GP time-delay embedding can be extended in an analogous fashion. More specifically, rather than letting the parameters of the system change over time, we allow f to vary. We do this by setting $f_{t+1} = f_t + \zeta_t$, where $\zeta_t \sim GP(0, W)$ and the prior for f at time 0 is $f_0 \sim GP(0, \Sigma)$. Here, W is the covariance function modeling the undirected changes in f. For simplicity, and to keep the number of additional parameters to a minimum, we set $W = \delta\Sigma$ where the scalar δ ranges from 0 to 1 and controls the rate at which f, and therefore the relationship between y_t and x_t changes through time. As $\delta \to 0$, f changes less through time and when $\delta = 0$ the model reduces to the basic GP time-delay embedding model (Section 1). We can gain some intuition for how f changes through time under this prior specification by looking at the correlation between f_t and f_0 : $Corr(f_t(x_t), f_0(x_s)) = (1+\delta t)^{-1/2} \prod_{i=1}^{L} R(\phi_i | y_{t-i} - y_{s-i}| / r)$. Thus, at a specific state x, the correlation decays with t as $(1+\delta t)^{-1/2}$ becoming effectively independent for long t. For a single population, this 'dynamic embedding' model is given by

$$\begin{aligned} p[y_t | f_t, x_t, V_\varepsilon] &\sim N\left(f_t(x_t), V_\varepsilon\right) \\ p[f_{t+1} | f_t, \boldsymbol{\theta}] &\sim GP(f_t, \delta\Sigma) \\ p[f_0 | \boldsymbol{\theta}] &\sim GP(0, \Sigma) \\ p[V_\varepsilon, \boldsymbol{\theta}] & \end{aligned} \tag{4}$$

In this model, we assign $1/\delta$ an exponential prior with mean $0.49/T$ so that the expected time for the correlation to drop to 0.1 is $\sim 2T$.

To evaluate the performance of the dynamic embedding model, we simulate data from three of the models described in Section 2: density-dependent maturation, migration (2 location, Ricker), and maternal effects. In this section, the goal is to determine our ability to forecast nonstationary dynamics; therefore, we allow the growth rate, migration rate and the maximum reproductive rate to vary through time in each of the models respectively (see Table 3.

Table 3
Dynamic embedding results for three ecological scenarios. For each ecological scenario, time series were simulated with one parameter (Driver) increasing linearly through time over the range indicated. This was repeated under three different parameter sets (see Table 1 for model definitions and parameter values) and three different levels of process stochasticity. Because of the driving variable, the dynamical regimes for the 3 parameter sets from Table 1 are no longer strictly chaotic, limit cycle, and stable. For each combination, we report the proportional reduction in 1-step forecast error for the dynamic embedding model compared to one without temporal drift. The numbers in parentheses represent the percentage of simulations that the dynamic embedding model outperformed the independent model.

Model Name	Driver	Set	$s^2 = 0.05$	$s^2 = 0.1$	$s^2 = 0.02$
Density dependent maturation	$G_{max} \in (0.25,0.99)$	1	0.2898 (100%)	0.3748 (100%)	0.2312 (100%)
		2	0.3365 (100%)	0.3343 (100%)	0.2337 (100%)
		3	0.2064 (100%)	0.2086 (100%)	0.1922 (100%)
Migration (2 locations, Ricker)	$m_1 \in (0.25,0.8)$	1	−0.0252 (31.4%)	0.0422 (82.9%)	0.0293 (71.4%)
		2	0.0687 (80.0%)	0.0371 (74.3%)	0.0227 (74.3%)
		3	0.1223 (91.4%)	0.1332 (97.1%)	0.0351 (80.0%)
Maternal Effects	$r \in (1.5,5.2)$	1	0.4471 (100%)	0.3256 (100%)	0.1228 (97.1%)
		2	0.4283 (100%)	0.3205 (100%)	0.1271 (100%)
		3	0.4381 (100%)	0.3204 (100%)	0.1080 (88.6%)

Please cite this article in press as: S.B. Munch, et al., Circumventing structural uncertainty: A Bayesian perspective on nonlinear forecasting for ecology, Ecol. Complex. (2016), http://dx.doi.org/10.1016/j.ecocom.2016.08.006

We simulate 75 years of data from each of the models, use the first 50 years to determine the model parameters and compute one-step predictions for the last 25 years conditional on the MAP estimates. We compare the forecast error of the dynamic embedding model against that of time-constant model as in Section 2.

Although the attractor changes shape as the driving variable changes, the standard GP does a fair job of forecasting, but does so by inflating the estimated process noise (Fig. 4). In contrast, the dynamic embedding model produces more accurate one-step forecasts with smaller confidence bands.

Overall, the dynamic embedding model reduces forecast error for most simulations (Table 3). The magnitude of the improvement is 20–30% for most of the small-noise scenarios ($s^2 = 0.05$ and 0.1). For $s^2 = 0.2$, the error reductions are typically 10–20%. The main exception to this is for parameter set 1 of the migration model. For this model the time-constant GP seems to do quite well despite the time-varying parameter. Thus, adding drift actually reduces model performance.

3. Discussion

Here we have developed Bayesian approaches to time delay embedding for use in ecological forecasting. The Bayesian paradigm offers a number of advantages over algorithmic approaches including automatic quantification of uncertainty, automatic detection of relevant lags, and incorporation of prior information. Another advantage is the ease with which the models are generalized. Hierarchical and dynamic embedding models were developed to cope with the short, noisy, and nonstationary time series encountered in ecology. As shown in Appendix C, combining these models is trivial, though we have not yet tested the combined model with simulations.

As with any Bayesian model, we must be mindful of the influence of the prior. For more general statistical modeling, Zexun and Wang (2016) have shown that the GP is relatively insensitive specification of the priors on the hyperparameters and more sensitive to assumptions about the covariance function. Here, we have attempted to incorporate some minimal, biological information into the prior, asserting that we expect to see functions that are smooth, continuously differentiable, and have on average one local maximum. These assertions are captured in the squared exponential covariance with the ARD prior on the length scale parameters. For the small-sample applications we have in mind it is particularly useful in eliminating spurious structure (see e.g. the illustration in Appendix A) and is analogous to the use of a 'wiggliness' penalty in frequentist approaches.

For some applications, however, this prior specification may be overly restrictive. Two cases seem particularly relevant. First, the squared exponential prior implicitly asserts that the curvature of the function to be estimated is roughly constant throughout its domain. When this is not the case, the squared exponential

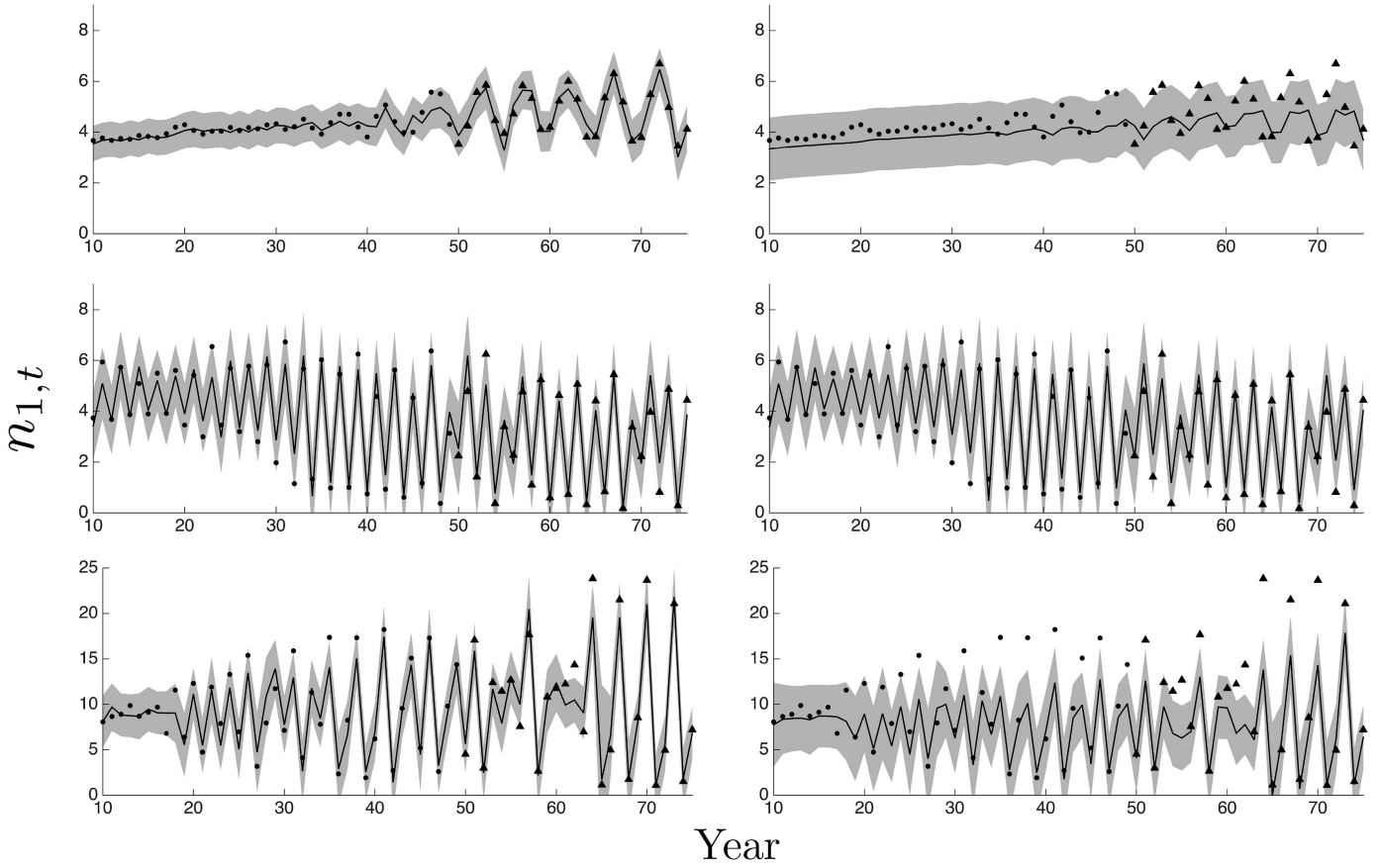

Fig. 4. Performance of the dynamic embedding model (left) relative to one without drift (right). The black line (mean) and grey patch (±2 sd) illustrate one-step-ahead predictions based on data (black dots) from three different ecological scenarios (Top row: Density dependent maturation, center row: Two location Ricker, bottom row: Maternal effects). In each case one parameter of the simulation model is changing through time. The stationary GP compensates for this by making the noise variance high and the length scale short, which tends to produce imprecise forecasts. In contrast, the dynamic embedding model is robust to slow changes in the underlying dynamics.

Please cite this article in press as: S.B. Munch, et al., Circumventing structural uncertainty: A Bayesian perspective on nonlinear forecasting for ecology, Ecol. Complex. (2016), http://dx.doi.org/10.1016/j.ecocom.2016.08.006

covariance will not work well. E.g. if the function has a high, narrow peak near 0 and is otherwise fairly constant, the inferred length scale will be some compromise that results in underestimating f near the peak and too much uncertainty in the constant region. Second, we may wish to infer the multi-step ahead map directly from the data {e.g. to forecast k steps ahead, fit $y_{t+k} = f(y_{t-1}, \ldots y_{t-L})$}. While it is biologically reasonable to assert that the one-step map is unimodal, the corresponding k-step map may have up to 2^k local maxima. When we naively use the ARD prior in this case, the resulting GP is too smooth and assigns most of the variation to process noise, rather than dynamics. In both cases, we could potentially improve inference by changing the prior specification (e.g. using nonstationary covariance functions) but more work is needed to evaluate this in ecological applications.

Combining short time series through hierarchical GP models increases forecast precision across a wide range of ecological scenarios. In cases where more than two or three series are available, it may be worthwhile to generalize the hierarchical structure to allow different degrees of dependence across pairs. Rather than have the pointwise correlation be constant for all pairs, ρ, we could instead model pairwise correlations separately, e.g., $Corr\left(f_i(x_t), f_j(x_t)\right) = \rho_{ij}$ for the i^{th} and j^{th} populations. This approach offers a means of identifying clusters of series with similar dynamics. If the dynamics are expected to vary along spatial gradients, a natural choice would be to use a spatial covariance function to constrain the correlations across embeddings.

We have focused primarily on the situation where multiple series are available for a single species. For any given location it is often the case that data are available for more than one species. In this case, the most direct thing to do is to fit GP models for each species using the data for all species as predictors, i.e. we could write $x_t = f(x_{t-1}, y_{t-1}, \ldots) + \varepsilon_t$. This could be done easily using the methods we have described with little modification. However, it is worth noting that the multivariate embedding theorem (Deyle and Sugihara, 2011) tells us that we are justified in using any combination of L lags from all interacting species. This suggests that we might improve forecast precision using model averaging to combine multiple embeddings.

The dynamic embedding approach we propose for handling nonstationary series substantially improves our ability to forecast when the underlying dynamics are changing. In addition to allowing us to forecast in the presence of nonstationarity, we suspect that the dynamic embedding approach could be of use in anticipating ecological regime shifts. Specifically, we might compare the fit of models with and without temporal drift to test for the presence of nonstationarity.

Our approach to nonstationary dynamics assumes that we know little beyond the fact that the system is changing. If we had some information on *how* the system was changing there are several obvious alternatives. For instance, if the driving variable is actually known, we could simply include it as another state in the GP model. Alternatively, if the driver is unknown, we could write the dynamics in terms of a single extra variable that drifts through time, rather than allowing the shape of the inferred map to change. That is, we could write $x_t = f(x_{t-1}, \ldots, x_{t-L}, u_t) + \varepsilon_t$ with $u_t = u_{t-1} + \omega_t, \omega_t \sim N(0,1)$. It may be that a model with a low-dimensional drift term (ω_t) is more efficient than the infinite-dimensional model we have proposed.

We note that although the simulations results are promising, they are certainly not exhaustive. Our primary intent was to demonstrate proof of concept. Having done so, it would be worthwhile to determine the limits under which the models fail. For instance, although the hierarchical model reduces error when the simulation parameters differ by up to 25%, there must surely be a limiting difference beyond which the model is not useful. Similarly, we expect the dynamic embedding model to be useless when the driving variable changes sufficiently fast.

We have thus far ignored observation uncertainty. This makes it possible to analytically marginalize over the unknown f, simplifying the inferential problem down to a handful of hyperparameters. With observation errors, this is no longer the case and we must resort to either MCMC (Hastings, 1970; Ming-Hui et al., 2000) or Laplace approximation (Tierney and Kadane, 1986). Accounting for observation errors in GP-based dynamical models is certainly possible (Thorson et al., 2014), though at the cost of dramatically increased computation. It is also worth noting that in simulations using Sugihara's S-map for forecasting, the results are robust to modest levels of measurement error (Perretti et al., 2013; Deyle et al., 2013). Extending the models shown here to state-space settings is an important area for future development.

Nonlinear dynamics and chaos have been clearly demonstrated in multiple experimental systems (Ellner and Turchin, 1993; Desharnais et al., 2001; Becks and Arndt, 2008), and are likely to be the norm in ecology. Structural uncertainty is a pervasive problem in ecological modeling and ecosystem management. The models we propose do not offer a solution to the problems of structural uncertainty. Rather, they give us a way around them – particularly when the goal is to generate short term predictions. We anticipate that these methods will be of value in conservation and management whenever the 'true' dynamics of the system are unknown.

Acknowledgements

This work was supported by funding from the Lenfest Ocean Program and NOAA's IAM program. It has greatly benefitted from input by George Sugihara, Ethan Deyle, Hao Ye, Athansios Kottas, Alec MacCall, and Marc Mangel. S Munch is grateful to MBI, Alan Hastings, Jennifer Dunne, and Andrew Morozov for organizing the symposium on Uncertainty, Sensitivity, and Predictability in Ecology.

Appendix A. Prior specification

Prior specification

The univariate GP forecasting model has L + 2 hyperparameters: $V_\varepsilon, \tau^2, \phi$ corresponding to the process variance, the pointwise-prior variance in f, and L lag-specific length scale parameters $\phi_{1,\ldots}\phi_L$. We used weakly informative priors for V_ε and τ^2. Specifically, we set $P(\tau^2/\mathrm{Var(y)}) = P\left(\mathrm{V}_\varepsilon/\mathrm{Var(y)}\right) = \mathrm{Beta}(1.1, 1.1)$ such that the prior mean for each parameter is $\mathrm{Var}(y)/2$. This prior constrains the total variance in the predicted population size, e.g, y_{T+1}, to be less than twice the observed variance in $[y_1, \ldots, y_T]$.

Our prior for ϕ is more informative. We set $p(\phi)$ such that the expected number of local extrema is 1, in keeping with our intuition that the function we are attempting to estimate should not be too 'wiggly.' Our argument is as follows: Using the squared exponential correlation function $R = exp\left[-(\phi d)^2\right]$ with $d = |y_t - y_s|/r$, the distances,d, have been scaled between 0 and 1. This allows us to make use of the fact that the expected number of zero crossings of a stationary GP on the unit interval is given by E (# of crossings in [0,1]) = $\mathrm{Z} = \pi^{-1}[-\mathrm{R}''(0)]^{1/2}$ (Sacks and Ylvisaker, 1966). Combining this with the fact that the derivative of a GP is also a GP with covariance function given by $\partial^2\Sigma/\partial y_i \partial y_j$ (See Rasmussen and Williams, 2006), the expected number of local extrema on the unit interval is given by $\mathrm{Z} = \pi^{-1}[\mathrm{R}^{(4)}(0)]^{1/2}$.

Please cite this article in press as: S.B. Munch, et al., Circumventing structural uncertainty: A Bayesian perspective on nonlinear forecasting for ecology, Ecol. Complex. (2016), http://dx.doi.org/10.1016/j.ecocom.2016.08.006

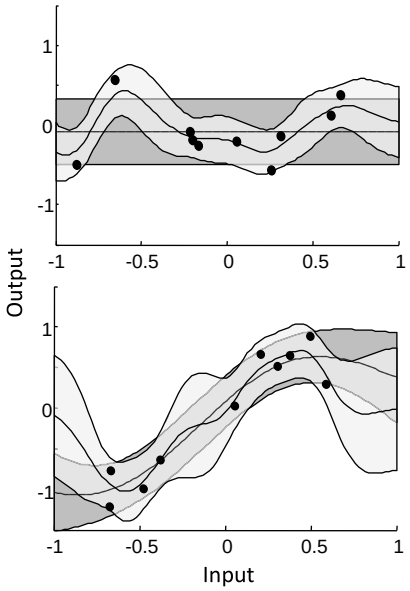

Fig. A1. Effects of the ARD prior on the inferred function. In both panels the black dots are simulated data where x_i is drawn from a uniform distribution on [-1,1]. In the top panel, the y's are independent of x, with $y_i = 0.3z_i$ where $z \sim N(0,1)$. In the bottom panel, $y_i = \sin(\square x_i) + 0.3z_i$. In each panel, the solid black line and light gray band indicate the posterior mean ±2 s.d. for f(x) using a uniform prior on Φ, while the dashed line and dark gray band indicate the results using the ARD prior.

Evaluating this for the squared exponential, we have $R^{(4)}(0) = 12\phi^4$ so that $Z(\phi) = \pi^{-1}\sqrt{12}\phi^2$. Since we want $E(Z) = \pi^{-1}\sqrt{12}E(\phi^2) \approx 1$, we set our prior for ϕ so that this is true. Here we use a half-normal prior, $(\phi) = 2\exp[-\phi^2/2v]/\sqrt{2\pi v}$. Since $E(\phi^2) = v$, setting $v = \pi/\sqrt{12}$ means $E(\text{turns in}[0,1]) = 1$

The value of this prior specification is illustrated in Fig. A1. In the top panel we simulate data for which there is no functional dependence of y on x while in the bottom panel we simulate a sine function with noise. Under the ARD prior, when there is no functional relationship between x and y, the inferred function is flat. When there is a functional dependence, the result under the ARD prior is quite smooth. In contrast, using a uniform prior on Φ leads to spurious shapes for f(x).

Appendix B. Updating

(A2.1) Updating

For all of the cases described in the main text, updating uses a two-stage approach. In the first stage, the parameters and hyperparameters $\{V_\varepsilon, \tau^2, \phi,$ etc.$\}$ are estimated from the marginal posterior obtained by integrating out f. The data consist of $\mathbf{X}$, the collection of delay vectors for the observed series $\{x_L, \ldots, x_T\}$ and y, the vector of observed next states $\{y_L, \ldots, y_T\}$. The marginal likelihood for $\mathbf{y}$ is multivariate normal and the log of the marginal posterior is given by

$$\ln p(V_\varepsilon, \theta | data) = -\frac{1}{2}\ln|\boldsymbol{\Sigma} + V_\varepsilon I| - \frac{1}{2}y^T[\boldsymbol{\Sigma} + V_\varepsilon I]^{-1}y - \ln p(V_\varepsilon, \theta) + c \quad \text{(A2.1)}$$

where the log of the prior (Appendix A) is indicated by $\ln p(V_\varepsilon, \theta)$ and θ is the vector of hyper parameters for the covariance function (i.e.τ^2, ϕ). The covariance matrix $\boldsymbol{\Sigma} = \Sigma(X, X^T)$ is obtained by evaluating the covariance function for all pairs of inputs, and $\boldsymbol{I}$ is the identity matrix. The constant is a normalizing factor that does not depend on the parameters. We use the R-prop algorithm (Blum and Riedmiller, 2013) to find the parameters that maximize the marginal log posterior.

In the second stage, we use the fact that the posterior distribution for f (given the MAP estimates of the parameters) is also a GP

$$p[f|y, \tau^2, \phi] \sim GP(m_c, \Sigma_c)$$

where the conditional mean and covariance functions, evaluated at new states x_{new} are given by

$$m_c(x_{new}) = \Sigma(x_{new}, X^T)[\boldsymbol{\Sigma} + V_\varepsilon I]^{-1}y$$
$$\Sigma_c(x_{new}, x^T{}_{new}) = \Sigma(x_{new}, x^T{}_{new}) - \Sigma(x_{new}, X^T)[\boldsymbol{\Sigma} + V_\varepsilon I]^{-1}\Sigma(X, x^T{}_{new}) \quad \text{(A2.2)}$$

Using this updating scheme, we can produce 1-step ahead forecasts directly by setting $x_{new} = x_{T+1}$.

Updating in the hierarchical and non-stationary models is quite similar. For example, in the hierarchical case, we stack the population-specific arrays into $x_{new} = x_{T+1}$ and $X_{stack} = \{X_1^T, \ldots, X_p^T\}^T$ where p is the number of populations. As in the single population case, the covariance matrix $\boldsymbol{\Sigma}_{stack}$ is assembled from the covariance functions, C and Σ, evaluated at all pairs of inputs in X_{stack}. The marginal likelihood is then given by replacing $\mathbf{y}$ and $\boldsymbol{\Sigma}$ with y_{stack} and $\boldsymbol{\Sigma}_{stack}$ in (A2.1). The conditional mean and covariance functions are evaluated using (A2.2) modified analogously. To make this concrete, imagine we have two populations. The covariance matrix is then given by

$$\boldsymbol{\Sigma}_{stack} = \begin{bmatrix} \Sigma(X_1, X_1^T) + C(X_1, X_1^T) & C(X_1, X_2^T) \\ C(X_2, X_1^T) & \Sigma(X_2, X_2^T) + C(X_2, X_2^T) \end{bmatrix} \quad \text{(A2.3)}$$

and the row vector $\Sigma(x_{new}, X)$ that appears in the conditional mean and covariance (A2.2) becomes $C(x_{i,new}, [X_1^T X_2^T]) + [\delta_{i,1}\Sigma(x_{i,new}, X_1^T)\delta_{i,2}\Sigma(x_{i,new}, X_2^T)]$ obtained by evaluating the covariance functions C and Σ at $x_{i,new}$ and each element of $\boldsymbol{X}_1$ *and* $\boldsymbol{X}_2$. The $\delta_{i,j}$ are Kronecker deltas, taking the value 1 if i = j and 0 otherwise.

Multi-step forecasts

If we are interested in k −step ahead forecasts for $k > 1$, two approaches are possible. The first approach is to estimate the future states by simply iterating the estimated f over several steps. A second approach, is to estimate a new function, say f_k that produces the k −step forecast directly by replacing the vector $y = \{y_L, \ldots, y_T\}$ with $y_k = \{y_{L+k}, \ldots, y_{T+k}\}$. Here we have adopted the former approach as it is more intuitive and k is not limited by the length of the time series.

Appendix C. Hierarchical models with time-varying dynamics

(6) Hierarchical models with time-varying dynamics

The hierarchical and nonstationary models can be readily combined, allowing the population-specific f' s and the across-population mean μ to drift through time. As in the previous two cases there is an underlying additive representation for f, which is now $f_{it} = \mu_t + z_{it}$ and $\mu_{t+1} = \mu_t + \eta_t$ with the addition of a random walk component for the deviations from the mean, i.e. $z_{it+1} = z_{it} + \zeta_{it}$. The intial conditions are represented by GP's: $\mu_0 \sim GP(0, C)$ and $z_{i0} \sim GP(0, \Sigma)$ and scalars are used to parameterize the temporal variations: $\eta_t \sim GP(0, \delta C)$ and $\zeta_{it} \sim GP(0, \psi\Sigma)$. The discount factor δ controls the rate at which μ changes through time, while the new discount factor ψ controls the rate at which populations drift independently. As in the hierarchical case, we set

Please cite this article in press as: S.B. Munch, et al., Circumventing structural uncertainty: A Bayesian perspective on nonlinear forecasting for ecology, Ecol. Complex. (2016), http://dx.doi.org/10.1016/j.ecocom.2016.08.006

the point-wise variance for μ_0 to $\rho\varpi^2$ and the pointwise variance in z_{i0} to $(1-\rho)\varpi^2$. The two previous models are obtained as special cases of this model by setting $\psi = \delta = 0$ for the hierarchical case or setting $\rho = 0$ to obtain independent time varying models for each population.

For building intuition, it is again useful to think about the correlation between two f' s at a single state x and time t. Across populations, we have

$$Corr(f_{jt}(x), f_{kt}(x)) = \frac{\rho(1+\delta t)}{\rho(1+\delta t) + (1-\rho)(1+\psi t)}$$

If the drift rates are equal, $\psi = \delta$, then the correlation remains ρ for all t. If the series do not drift independently ($\psi = 0$), then the correlation goes to 1 and if $\delta = 0$ then the correlation goes to 0.

The correlation through time for a single population is given by

$$Corr(f_{j0}(x), f_{jt}(x)) = \frac{1}{\sqrt{\rho(1+\delta t) + (1-\rho)(1+\psi t)}}$$

The introduction of z as an additional GP was just a notational convenience. Since $z_{it} = f_{it} - \mu_t$, we can eliminate it from the model. Doing so, the hierarchical model with time varying dynamics is given by

$$\begin{gathered} p[y_{it}|f_{it}, x_{it}, V_{\varepsilon i}] \sim N\left(f_{it}(x_{it}), V_{\varepsilon i}\right) \\ p[f_{i,t}|\mu_t, f_{i,t-1}, \mu_{t-1}, \boldsymbol{\theta}] \sim GP(\mu_t + f_{i,t-1} - \mu_{t-1}, \psi\Sigma) \\ p[\mu_{t+1}|\mu_t, \boldsymbol{\theta}] \sim GP(\mu_t, \delta C) \\ p[f_{i,0}|\mu_0, \boldsymbol{\theta}] \sim GP(\mu_0, \Sigma) \\ p[\mu_0|\boldsymbol{\theta}] \sim GP(0, C) \\ p[V_\varepsilon, \boldsymbol{\theta}] \end{gathered} \tag{6}$$

References

Bakker, R., Schouten, J.C., Giles, C.L., Takens, F., Van Den Bleek, C.M., 2000. Learning chaotic attractors by neural networks. Neural Comput. 12, 2355–2383.

Becks, L., Arndt, H., 2008. Transitions from stable equilibria to chaos, and back, in an experimental food web. Ecology 89, 3222–3226.

Beddington, J.R., Free, C.A., Lawton, J.H., 1975. Dynamic complexity in predator-prey models framed in difference equations. Nature 225, 58–60.

Bjornstad, O.N., Grenfell, B.T., 2001. Noisy clockwork: time series analysis of population fluctuations in animals. Science 293, 638–643.

Blum, M., Riedmiller, M.A., 2013. Optimization of Gaussian process hyperparameters using Rprop. European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Belgium, pp. 24–26.

Buzug, T., Pfister, G., 1992. Optimal delay time and embedding dimension for delay-time coordinates by analysis of the global static and local behavior of strange attractors. Phys. Rev. A 45 (10), 7073–7084.

Carpenter, S.R., 2003. Regime Shifts in Lake Ecosystems: Pattern and Variation, Excellence in Ecology Series, Vol. 15. Ecology Institute, Oldendorf/Luhe, Germany.

Chen, B., Tong, H., 1992. On consistent nonparametric order determination and chaos. J. R. Stat. Soc. Ser. B (Methodolo.) 54 (2), 427–449.

Cressie, N., 1993. Statistics for Spatial Data (revised). Wiley.

Desharnais, R.A., Constantino, R.F., Cushing, J.M., Henson, S.M., Dennis, B., 2001. Chaos and population control of insect outbreaks. Ecol. Lett. 4, 229–235.

Deyle, E.R., Sugihara, G., 2011. Generalized theorems for nonlinear state space reconstruction. PLoS One 6 (3), e18295.

Deyle, E.R., Fogarty, M., Hsieh, C.-H., Kaufmann, L., MacCall, A.D., Munch, S.B., Perretti, C., Ye, H., Sugihara, G., 2013. Predicting climate effects on Pacific sardine. PNAS 110, 6430–6435.

Doebeli, M., de Jong, G., 1999. Genetic variability in sensitivity to population density affects the dynamics of simple ecological models. TPB 55, 37–52.

Ellner, S., Turchin, P., 1993. Chaos in a "noisy" world: new methods and evidence from time series analysis. Am. Nat. 145, 343–375.

Gerber, L.R., Karieva, P.M., Bascompte, J., 2002. The influence of life history attributes and fishing pressure on the efficacy of marine reserves. J. Biol. Conserv. 106, 11–18.

Ginzurg, L.R., Taneyhill, D.E., 1994. Population cycles of forest Lepidoptera: a maternal effect hypothesis. J. Anim. Ecol. 63, 79–92.

Grenfell, B.T., Brornstad, O.N., Kappey, J., 2001. Travelling waves and spatial hierarchies in measles epidemics. Nature 414, 716–723.

Halstead, B.J., Wylie, G.D., Coates, P.S., Valcarcel, P., Casazza, M.L., 2012. Bayesian shared frailty models for regional inference about wildlife survival. Anim. Conser. 15, 117–124.

Hare, S.R., Mantua, N.J., 2000. Empirical evidence for North Pacific regime shifts in 1977 and 1989. Prog. Oceanogr. 47, 103–145.

Hastings, A., Hom, C.L., Ellner, S., Turchin, P., Godfray, H.C.J., 1993. Chaos in ecology: is mother nature a strange attractor? Annu. Rev. Ecol. Syst. 24, 1–33.

Hastings, W.K., 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrics 57, 97–109.

Hsieh, C.-H., Anderson, C., Sugihara, G., 2008. Extending nonlinear analysis to short ecological time series. Am. Nat. 171, 71–80.

Ives, A.R., Dakos, V., 2012. Detecting dynamical changes in nonlinear time series using locally linear state-space models. Ecosphere 3 (6), 58.

Kannathal, N., Choo, M.L., Acharya, U.R., Sadasivan, P.K., 2005. Entropies for detection of epilepsy is EEG. Comput. Methods Programs Biomed. 80, 187–194.

Lee, T.H., Shiba, S., Wood, R.C., 1999. Integrated Management Systems: A Practical Approach to Transforming Organizations. Jon Wiley and Sons Inc..

Mayfield, E.S., Mizrach, B., 1992. On determining the dimension of real-time stock-price data. J. Bus. Econ. Stat. 10 (3), 367–374.

Ming-Hui, C., Shao, Q.-M., Ibrahim, J.G., 2000. Monte Carlo Methods in Bayesian Computation. Springer-Verlag, New York.

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

Munch, S.B., Kottas, A., Mangel, M., 2005. Bayesian nonparametric analysis of stock–recruitment relationships. Can. J. Fish Aquat. Sci. 62, 1808–1821.

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

Neubert, M.G., Caswell, H., 2000. Density-dependent vital rates and their population dynamic consequences. J. Math. Biol. 41, 103–121.

Nocedal, J., Wright, S., 1999. Numerical Optimization. Springer-Verlag, New York.

O'Hagan, A., 1978. Curve fitting and optimal design for prediction. J. R. Stat. Soc. Ser. B 40, 1–42 (with discussion).

Paciorek, C., Schervish, M., 2004. Nonstationary covariance functions for Gaussian process regression. Adv. Neural Inf. Process. Syst. 16, 273–280.

Patterson, T.A., Thomas, L., Wilcox, C., Ovaskainen, O., Matthiopoulos, J., 2008. State-space models of individual animal movement. Trends Ecol. Evol. 32 (2), 87–94.

Perretti, C.T., Munch, S.B., Sugihara, G., 2013. Model-free forecasting outperforms the correct mechanistic model for simulated and experimental data. PNAS 110, 5253–5257.

Rasmussen, C.E., Williams, C., 2006. Gaussian Processes for Machine Learning. The MIT Press.

Riedmeiller, M., Braun, H., 1993. A direct adaptive method for faster backpropagation: the RPROP algorithm. Proceedings of the IEEE International Conference on Neural Networks, San Francisco, CA.

Royle, J.A., Dorazio, R.M., 2008. Hierarchical Modeling and Inference in Ecology: the Analysis of Data from Populations, Metapopulations and Communities. Elsevier/Academic Press, London, UK.

Sacks, J., Ylvisaker, D., 1966. Designs for regression problems with correlated errors. Ann. Math. Stat. 37 (1), 66–89.

Schaffer, W.M., 1985. Order and chaos in ecological systems. Ecology 66, 93–106.

Schoombie, S.W., Getz, W.M., 1998. Evolutionary stable strategies and trade-offs in generalized Beverton and Holt growth models. Theor. Popul. Biol. 53, 216–235.

Shi, J.Q., Murray-Smith, R., Titterington, D.M., 2005. Hierarchical Gaussian process mixtures for regression. J. Stat. Comput. 15, 31–41.

Stark, J., Broomhead, D.S., Davies, M.E., Huke, J., 2003. Delay embeddings for forced systems. II. Stochastic forcing. J. Nonlinear Sci. 13, 519–577.

Stark, J., 1999. Delay embeddings for forced systems. I. Deterministic forcing. J. Nonlinear Sci. 9, 255–332.

Sugeno, M., Munch, S.B., 2013. Semiparametric Bayesian method for detecting Allee effects. Ecology 94, 1196–1204.

Sugihara, G., Grenfell, B., May, R.M., Chesson, P., Platt, H.M., Williamson, M., 1990. Distinguishing error from chaos in ecological time series. Philos. Trans.: Biol. Sci. 330, 235–251.

Sugihara, G., 1994. Nonlinear forecasting for the classification of natural time series. Philos. Trans.: Phys. Sci. Eng. 348, 477–495.

Summers, D., Cranford, J.G., Healey, B.P., 2000. Chaos in periodically forced discrete-time ecosystem models. Chaos, Solitons Fractals 11, 2331–2342.

Takens, F., 1981. Detecting strange attractors in turbulence. In: Rand, D.A., Young, L.-S. (Eds.), Dynamical Systems and Turbulence, Lecture Notes in Mathematics. Springer-Verlag.

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

Tierney, L., Kadane, J.B., 1986. Accurate approximations for posterior moments and marginal densities. J. Am. Stat. Assoc. 81, 82–86.

Turchin, P., 1990. Rarity of density dependence or population regulation with lags? Nature 344, 660–663.

Walters, C., Christensen, V., Fulton, B., Smith, A.D., Hilborn, R., 2016. Predictions from simple predator-prey theory about impacts of harvesting forage fishes. Ecol. Modell. 337, 272–280.

West, M., Harrison, J., 1997. Bayesian Forecasting and Dynamic Models. Springer-Verlag, New York.

Wikle, C.K., 2003. Hierarchical Bayesian models for predicting the spread of ecological processes. Ecology 84, 1382–1394.

Wood, S.N., Thomas, M.B., 1999. Super−sensitivity to structure in biological models. Proceedings of the Royal Society of London B: Biological Sciences 565–570 (266).

Zexun, C., Wang, B., 2016. How priors of initial hyperparameters affect Gaussian process regression models. arXiv:1605.07906 [stat.ML], http://arxiv.org/abs/1605.07906.

Please cite this article in press as: S.B. Munch, et al., Circumventing structural uncertainty: A Bayesian perspective on nonlinear forecasting for ecology, Ecol. Complex. (2016), http://dx.doi.org/10.1016/j.ecocom.2016.08.006