**Abstract.** Diffusion processes are natural modelling frameworks for many stochastic phenomena. However since inference for nonlinear diffusion models is difficult, the diffusion models used within the ecological literature are predominantly linear thus ignoring the more complex forces that many populations and environments are subject to. We suggest a parameter estimation procedure that is computationally efficient and valid for a wide class of diffusion models. This includes diffusions that are nonlinear and/or multivariate. The procedure is used to fit a diffusion model to a time series with three variables: the plankton abundances for both *Pseudo-nitzschia australis* and *Prorocentrum micans* in addition to the surface water temperature. *Pseudo-nitzschia australis* are shown to achieve maximal growth with a water temperature in the region of 7.8–16.0°C. The corresponding optimum temperature region for *Prorocentrum micans* is 12.5–24.5°C with the dinoflagellate appearing to be less sensitive to sub-optimal temperatures. Both species are governed by r-type strategies and are subject to regulated growth. Though *Prorocentrum micans* are known to feed on diatoms, we are unable to find evidence of feeding on *Pseudo-nitzschia australis*. We thus demonstrate that by fitting a diffusion model to an ecological dataset, one may learn much about the factors that affect the ecosystem. A good understanding of these factors is vital in order to manage an ecosystem appropriately.

**Key words:** diffusion process; Fokker-Planck equation; Markov Chain Monte Carlo; parameter estimation; plankton; population model.

**INTRODUCTION**

Diffusion processes are powerful instruments for modelling continuous-time phenomena. Not only can these models account for a wide range of stochastic behavior, but they also enable one to infer the dynamics that govern the evolution of a process from discretely-sampled data. Such inference remains straightforward even when the data is irregularly spaced. Consequently diffusion models have found widespread application in many different fields of scientific endeavor.

In ecology, diffusion processes are used to model the movements of organisms and hence to explain the distribution and dispersal of a species (Tufto et al. 2012). They are also widely used for analyzing a population’s viability (Takimoto 2009). Furthermore, diffusion models can be used to study the impact of established ecological concepts—such as the Allee effect—on a population (Takimoto 2009). Generalizations of tradi-
Diffusion processes to include time delayed input have also found application in population modelling (Mao et al. 2005).

Diffusion processes may also serve as natural extensions to the ordinary differential equations often used in ecology. Mao et al. (2002) showed how the addition of a noise component to an ordinary differential equation, thus reformulating the model in terms of a stochastic differential equation (SDE), can suppress potential population explosions. Indeed diffusion processes can approximate many well-known population models. This includes branching processes (Allen and Allen 2003), age-structured population models (Vindenes et al. 2011) as well as birth and death processes (Varughese 2009). Diffusion processes have also been used to model environmental variables such as phosphorous levels in lake water and sediment (Carpenter and Brock 2006). The usage of diffusion processes in both population and environmental modelling suggests that higher dimensional diffusion processes can provide a unified framework for inferring relationships between populations and the environment (Varughese 2011).

Compared to the many ecological papers studying the population behavior implied by a prescribed diffusion model, relatively few perform statistical inference using a diffusion model. Within an ecological context, parameter estimation has been primarily performed for simple diffusion models. For example, Gaston and Nicholls (1995) fitted a Wiener process with drift to bird populations in order to estimate their times to extinction. Some extensions have been made to more complex diffusion models. Knape and de Valpine (2012) fitted a more complex quadratic diffusion model to kangaroo data by combining Markov Chain Monte Carlo (MCMC) with particle filters. The authors point out that this is a brute-force method that is computationally intensive. Humbert et al. (2009) showed how one may estimate the parameters of a simple diffusion process with observation error. Ovaskainen (2004) developed a diffusion model of species movement across a heterogeneous landscape and used a finite element scheme to estimate the model parameters. Ananthasubramanian et al. (2011) focused on the estimation of the volatility of a diffusion process subsequently enabling their study of the behavior of a size-structured population.

This article demonstrates a recently developed, computationally efficient parameter estimation procedure that is applicable for a wide class of diffusion models (Varughese 2013). The methodology allows quite arbitrary model specification thus widening the scope of inference substantially: whether the aim is to measure specific phenomena or to encapsulate multiple points of interest in a model, the method remains the same. This allows for the fitting of more realistic diffusion models of an ecosystem in a computationally efficient manner. The method is used to fit a diffusion model to a plankton time series. The flexibility of the parameter estimation procedure enables us to fit a diffusion model that simultaneously allows for species self-regulation, interspecific interactions and environmental drivers. The fitted model can subsequently be used to better understand the dynamics of plankton fluctuations. As such, the purpose of this paper is threefold: firstly, introduce the diffusion process and its properties; secondly, introduce an efficient parameter estimation method that is valid for a wide class of diffusion processes and finally, fit a model to an ecological dataset and report on the subsequent findings.

**Diffusion Processes**

A multivariate diffusion process can be represented by the following stochastic differential equation (SDE):

\[
d\Phi(t) = \mu(\Phi(t), t; \theta)dt + \Sigma^{1/2}(\Phi(t), t; \theta)dB(t),
\]

where \(\Phi(t) = (\Phi_i(t))_{i=1, \ldots, m}\) is the process vector at time \(t\) and \(d\Phi(t) = \Phi(t + dt) - \Phi(t)\). \(\Phi(t)\) can for example represent a collection of species abundances and environmental factors. The parameter vector \(\theta = (\theta_i)_{i=1, \ldots, p}\) of the diffusion system may represent ecological parameters such as a crowding coefficient or environmental sensitivity.

\(\mu(\Phi(t), t; \theta) = (\mu_i(\Phi(t), t; \theta))_{i=1, \ldots, m}\) is the drift vector; \(\Sigma(\Phi(t), t; \theta) = (\sigma_{ij}(\Phi(t), t; \theta))_{i,j=1, \ldots, m}\) is the diffusion matrix with \(\Sigma(\Phi(t), t; \theta) = \Sigma^{1/2}(\Phi(t), t; \theta)\Sigma^{1/2}(\Phi(t), t; \theta)^T\). Finally, \(B(t) = (B_i(t))_{i=1, \ldots, m}\) is a vector of independent Brownian motions with \(dB(t) = B(t + dt) - B(t)\). For a general introduction to stochastic differential equations consult Stirzaker (2005).

Table 1 gives a summary of all notation used.
The stochastic component of the diffusion process $\Phi(t)$ is driven by the $m$ independent Brownian motions. Brownian motion $B(t)$ is a continuous-time stochastic process characterized by the following properties: firstly, $B(0) = 0$; secondly, the function mapping $t \rightarrow B(t)$ is continuous and lastly, $B(t)$ has independent increments with $B(t) - B(s) \sim \text{Normal}(0, t - s)$.

From Eq. 1, it is easy to see that the diffusion system is characterized by $\mu(\Phi(t), t; \theta)$ and $\Sigma(\Phi(t), t; \theta)$ as they govern the evolution of $\Phi(t)$ through time. Their individual elements are defined as:

$$
\mu_i(\Phi(t), t; \theta) = \lim_{\Delta t \rightarrow 0} \frac{E[\phi_i(t + \Delta t) - \phi_i(t)] | \Phi(t), t; \theta]}{\Delta t},
$$

$$
\sigma_{ij}(\Phi(t), t; \theta) = \lim_{\Delta t \rightarrow 0} \text{Cov}([\phi_i(t + \Delta t) - \phi_i(t), \phi_j(t + \Delta t) - \phi_j(t)] | \Phi(t), t; \theta) / \Delta t.
$$

Over an infinitesimally small interval, the diffusion system is distributed (Stirzaker 2005):

$$
\Phi(t + \Delta t) - \Phi(t) \sim \text{Multivariate Normal} \left( \mu(\Phi(t), t; \theta) \Delta t, \Sigma(\Phi(t), t, \theta) \Delta t \right).
$$

Since the drift vector and diffusion matrix only depend on the current process value $\Phi(t)$, the process is Markov. Furthermore, as the time interval $(t; t + \Delta t)$ lengthens, the mean and variance of $\Phi(t + \Delta t) - \Phi(t)$ only remain proportional to $\Delta t$ for a special case (a Wiener process with drift).

In general, the evolution of the probability density function of $\Phi(t) = (\phi_i(t))_{i=1,...,m}$ through time—which we denote as $p(\Phi(t))$—is governed by the Kolmogorov forward equation (Cox and Miller 1977):

$$
\frac{\partial p(\Phi(t))}{\partial t} = -\sum_{i=1}^{m} \frac{\partial}{\partial \phi_i} \left[ \mu_i(\Phi(t), t; \theta) p(\Phi(t)) \right] + \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \frac{\partial^2}{\partial \phi_i \partial \phi_j} \left[ \sigma_{ij}(\Phi(t), t; \theta) p(\Phi(t)) \right].
$$

This is also known as the Fokker-Planck equation. There is a corresponding partial differential equation for the moment generating function (MGF) (for more on MGFs consult Mood et al.)
Let $\Lambda = (\nu_1, \ldots, \nu_m)$ be the parameters for the MGF $M(\Lambda, t) = E[\exp[\nu_1 \phi_1(t) + \cdots + \nu_m \phi_m(t)]]$ of the system. In addition, let $\Lambda^* = (\partial/\partial \nu_1, \ldots, \partial/\partial \nu_m)$. The MGF is governed by the following partial differential equation (Varughese 2013):

$$\frac{\partial M(\Lambda, t)}{\partial t} = \left[ \sum_{i=1}^{m} \nu_i \mu_i(\Lambda^*, t; \theta) \right] M(\Lambda, t) + \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \nu_i \nu_j \sigma_{ij}(\Lambda^*, t; \theta) M(\Lambda, t).$$

(4)

Unless required, we may suppress any dependence on the parameter vector $\theta$ within our notation. Though the terms $\mu_i(\Lambda, t)$ and $\sigma_{ij}(\Lambda, t)$ denote functions of $\Lambda$ and $t$, the terms $\mu_i(\Lambda^*, t)$ and $\sigma_{ij}(\Lambda^*, t)$ represent differential operators that act on the MGF $M = M(\Lambda, t)$. So, for example, if we have a 2-dimensional diffusion system with drift term $\mu_i(\phi_1, \phi_2) = a \phi_1 - b \phi_1^2 - c \phi_1 \phi_2$, then

$$\mu_i \left( \frac{\partial}{\partial \phi_1}, \frac{\partial}{\partial \phi_2} \right) = a \frac{\partial}{\partial \phi_1} - b \frac{\partial^2}{\partial \phi_1^2} + c \frac{\partial^2}{\partial \phi_1 \phi_2}$$

is a differential operator and

$$\mu_i \left( \frac{\partial}{\partial \phi_1}, \frac{\partial}{\partial \phi_1}, \frac{\partial}{\partial \phi_2} \right) M = a \frac{\partial M}{\partial \phi_1} - b \frac{\partial^2 M}{\partial \phi_1^2} + c \frac{\partial^2 M}{\partial \phi_1 \phi_2}$$

represents the action of the operator on the MGF $M(\Lambda, t)$. In general, neither Eq. 3 nor Eq. 4 are analytically tractable.

Since a diffusion process is Markovian, the likelihood of a diffusion system sampled at discrete time points $(t_1, t_2, \ldots, t_N)$ can be decomposed multiplicatively (Ait-Sahalia 2002). That is,

$$L(\theta) = p(\Phi(t_1)) \prod_{i=2}^{N} p(\Phi(t_i) | \Phi(t_{i-1})).$$

(5)

Asymptotically, for large $N$, the term $p(\Phi(t_1))$ may be ignored whilst the transitional distribution $p(\Phi(t_i) | \Phi(t_{i-1}))$ is the solution to Eq. 3 at time $t_i$ with the boundary condition that $p(\Phi(t))$ is given by the Dirac delta function centered around $\Phi(t_{i-1})$ at time $t_{i-1}$. The Dirac delta function is zero everywhere except at $\Phi(t_{i-1})$, where it must be infinite in order to ensure the probability density integrates to one. Using this probability density at time $t_{i-1}$ reflects the fact that we are conditioning on $\Phi(t_{i-1})$ in Eq. 5 so there is no uncertainty about the process value at time $t_{i-1}$. Unfortunately, except for a few special cases, Eq. 3 (and hence also the likelihood) is analytically intractable.

A number of approaches have been suggested to approximate the likelihood when Eq. 5 is intractable. One may use Bayesian imputation to augment the data so that the distribution of the increments is approximated by Eq. 2 (Roberts and Stramer 2001). Alternatively, Monte Carlo methods may be used to approximate the likelihood (Kleinhans and Friedrich 2007). One can also solve Eq. 3 numerically. Wojtkiewicz and Bergman (2000) discretized the spatial domain and solved the partial differential equation numerically at each point on the lattice. Another possibility is to seek a closed form approximation to the transitional probability. Ait-Sahalia (2002) approximated the solution to Eq. 3 by a Hermite polynomial expansion.

Instead of approximating the transitional distribution directly, this paper first seeks to approximate the conditional moments of the diffusion system. These may be subsequently used to approximate the likelihood. Huang (2012) approximated the first two conditional moments using a Wagner-Platen approximation. Varughese (2013) approximated the conditional moments to an arbitrary order using the cumulant truncation procedure. This method has a number of features which makes it suitable for ecological datasets: the method is applicable for multivariate diffusion systems and it can be naturally incorporated within a MCMC procedure. Furthermore, in comparison to small-time expansions like the Hermite and the Wagner-Platen approximations, the method is more robust to the large lapses between observations that are prevalent with ecological datasets. Consequently, we apply the method to fit a diffusion model to an ecological dataset.

The Dataset

We introduce the methodology of the present paper by analyzing part of the seminal dataset compiled by W. E. Allen during 1917–1939 on the abundance of phytoplankton and the state of relevant environmental variables. Thomas et al. (2001) give a very detailed analysis of how the data was collected and managed. This time series...
contains dense information over long periods of time, warranting the analysis of a diffusion representation of the underlying population dynamics and environmental factors. We give a brief description of the data collection scheme and then discuss the attributes of the species to be modelled.

Observations on the number of each species present in water samples, taken from sites along the Southern California coastline, were collected daily and at the end of each week compiled into weekly averages by mixing the samples for each site and re-analyzing the content for observations (Thomas et al. 2001). Our focus will be on weekly observations taken from the Scripps Institute of Oceanography station for the dinoflagellate *Prorocentrum micans* and the diatom *Pseudo-nitzschia australis* as well as the accompanying surface water temperatures for the time period 1930 to 1937. We focus on the post 1930 data as the data collection method changed in this period. Fig. 1 displays log-population numbers together with the surface water temperature. The dataset consists of observations made at non-equidistant intervals, up to 17 weeks apart.

Mixotrophic behavior among dinoflagellates is now a well-established phenomenon in marine biology (Jeong et al. 2010). Burkholder et al. (2008) give evidence of the ingestion of prey by *Prorocentrum micans* together with the morphological adaptations required to support this trophic mode. These phenomena suggest the possible existence of a predator-prey relationship between *Prorocentrum micans* and *Pseudo-nitzschia australis* which lead to interesting possibilities from a population dynamics perspective, given that both species are known to exhibit blooming dynamics caused by changes in ambient media (Cochlan et al. 2008), as well as the use of photosynthesis as a primary method of energy production. Thus the rules that govern the underlying dynamics are potentially more complex than those of specialist predators and their prey, where predation components dominate the underlying dynamics. Smayda (2002) explore contrasts in the blooming behavior of dinoflagellates and diatoms in general, citing specific adaptations and strategies (in this instance transience in trophic modes and nutrient uptake) under ambient changes as significant population driving components. The preference of species to water temperature is difficult to establish, with *Pseudo-nitzschia australis* and other harmful *Pseudo-nitzschia* species being cosmopolites (Lundholm and Moestrup 2006), showing up in temperate coastal waters (Guiry 2012). *Prorocentrum micans* also tolerates a wide range of temperatures (Guiry 2012) and shows vast geographical distribution. Temperature does still however play a crucial role in growth rates and contribute to conditions that result in blooms (Rose and Caron 2007).

**ANALYSIS**

**A multivariate model of the data**

Our attention now turns towards the development of an appropriate model for this dataset. An ecological model should be capable of mimicking the behavior of the primary forces that affect the system. By fitting such a model to the plankton data, one can infer the system’s dynamics. This however is constrained by the need for model parsimony. Parsimonious models are usually easier to interpret. By definition, the parameter space of a parsimonious model will be of a lower dimension than that for a complex model. This is desirable as not only is it easier to find the global maximum of the resulting model’s posterior distribution but it is also less likely that the model will suffer from model degeneracies (that is, having large regions in the parameter space for which the posterior distribution is very close to its maximum). Difficulties arising from high dimensionality get compounded in multivariate diffusion models making parsimony all the more important.

In this paper, the focus is not only on the nature of the interactions between the two plankton species, but also on the impact of the fluctuating water temperature on their abundances. The surface water temperature \( T(t) \) consists of a deterministic and a random component. Let:

\[
T(t) = w(t) + \Gamma(t),
\]

where \( w(t) \) represents the deterministic fluctuation in temperatures and \( T(t) \) is the residual component. \( w(t) \) is given by the sum of the annual seasonal variation in temperatures (assumed to be sinusoidal) and a longer term cyclical component (assumed to be a polynomial
of order $p$) that is driven by the Southern Oscillation Index. That is:

$$w(t) = \sum_{i=0}^{p} a_i t^i + \text{Asin} \left( \frac{2\pi t}{365} + \varphi \right).$$  \hspace{1cm} (7)

The cyclical component is estimated by fitting a degree seven polynomial to the rolling mean of the water temperature. By subsequently subtracting the resulting polynomial from the data, the seasonal component can be fitted using the method of least squares. Fig. 2 shows the deterministic component $w(t)$ superimposed onto the temperatures.

The residual temperature $\Gamma(t)$ is modelled as a diffusion process.

$$d\Gamma(t) = h\Gamma(t)dt + \sqrt{q}dB(t),$$  \hspace{1cm} (8)

where $q > 0$ is the instantaneous volatility of the temperature and $h < 0$. Keeping $h < 0$ ensures that the temperature tends to drift towards the deterministic component $w(t)$. We may derive the corresponding stochastic differential equation (SDE) for $T(t)$ through Itô’s lemma.

Given that a diffusion process $X(t)$ behaves according to the SDE $dX(t) = \mu(t, X(t))dt + \sigma(t, X(t))dB(t)$, Itô’s lemma gives the corresponding SDE for a transformation $g(t; X(t))$:

$$dg(t; X(t)) = \left( \frac{\partial g}{\partial t} + \mu \frac{\partial g}{\partial x} + \frac{\sigma^2}{2} \frac{\partial^2 g}{\partial x^2} \right) dt + \sigma \frac{\partial g}{\partial x} dB(t).$$  \hspace{1cm} (9)

By applying Eq. 9 to Eqs. 6 and 8, we obtain:

$$dT(t) = \left( h(T(t) - w(t)) + w'(t) \right)dt + \sqrt{q}dB(t).$$  \hspace{1cm} (10)

The SDE is nonlinear in time since $w(t)$ is a nonlinear function. In addition to the SDE for
water temperature, the SDEs for the two species must also be specified.

Due to the very large range in the species abundances, we take their log transforms. Let $A(t)$ and $C(t)$ be the logs of species abundances for *Pseudo-nitzschia australis* and *Prorocentrum micans*, respectively. Note that:

$$dA(t) = \log\left( \frac{P. australis \text{ abundance at time } t}{P. australis \text{ abundance at time } t + dt} \right).$$

That is both $dA(t)$ and $dC(t)$ may be interpreted as the logs of the instantaneous growth rates for the two species. We model these as follows:

$$dA(t) = [a_1 + b_{1,1}A(t) + b_{1,2}T(t) + c_1T(t)^2 + b_{1,3}C(t)] \ dt + \sqrt{d_1}dB_1(t)$$

$$dC(t) = [a_2 + b_{2,1}C(t) + b_{2,2}T(t) + c_2T(t)^2 + b_{2,3}A(t)] \ dt + \sqrt{d_2}dB_2(t).$$

The terms $b_{1,1}$ and $b_{2,1}$ represent intraspecies competition coefficients. As such, both $b_{1,1}$ and $b_{2,1}$ will be negative hence ensuring that the growth rates of the two species are regulated. By making the diffusion’s drift terms quadratic in $T(t)$, we allow their relationship to water temperature to be concave with the degree of concavity controlled by the parameters $c_1$ and $c_2$. Consequently, the optimal temperatures for maximal growth of *P. australis* and *P. micans* are $-b_{1,2}/2c_1$ and $-b_{2,2}/2c_2$ respectively with $c_1$ and $c_2$ representing the respective species’ sensitivity to departures from the optimal temperature. Interactions between the two species are allowed.
through the parameters $b_{1,3}$ and $b_{2,3}$. This enables us to look for possible predation upon *P. australis* by *P. mican*. Finally, the parameters $d_1$ and $d_2$ represent the respective volatilities in the growth rates of the two species.

In total, there are 14 parameters that need to be estimated: $a_1, a_2, b_{1,1}, b_{1,2}, b_{1,3}, b_{2,1}, b_{2,2}, b_{2,3}, c_1, c_2, d_1, d_2, h$ and $q$. These are subsumed within the parameter vector $\theta$. The value of $\theta$ determines the transitional trivariate distribution $p(\Phi(t)|\Phi(t_{i-1}))$ for $\Phi(t) = (A(t), C(t), T(t))$ and hence also the value of the likelihood within the notation.

**Parameter estimation**

We fit the plankton model to the W. E. Allen dataset. As the model is multivariate and nonlinear, neither Eq. 3 nor the corresponding likelihood are analytically tractable consequently making inference difficult. However the parameter estimation method suggested by Varughese (2013) can still be used to analyze the data. The resulting analysis within this section is didactic, showcasing the application of this estimation procedure. Except where it may be unclear, any functional dependence on $\Lambda$ and $t$ is suppressed within the notation.

The first step in the procedure is to derive a differential equation governing the evolution of the model’s MGF $M(\Lambda, t) = E[\exp\{v_1A(t) + v_2C(t) + v_3T(t)\}]$ through time. By substituting Eqs. 10 and 11 within Eq. 4, we obtain:

$$ \frac{\partial M(\Lambda, t)}{\partial t} = \left[ v_1 a_1 + v_2 a_2 - v_3 h w(t) + v_3 w'(t) \right] M 
+ v_1^2 b_{1,1} \frac{\partial M}{\partial v_1} 
+ v_2^2 b_{1,1} \frac{\partial M}{\partial v_1} 
+ v_1 b_{1,2} + v_2 b_{2,1} \frac{\partial M}{\partial v_3} 
+ v_1^2 c_1 + v_2^2 c_2 \frac{\partial^2 M}{\partial v_3^2}. $$

It is subsequently possible to derive the corresponding partial differential equation for the cumulant generating function (CGF) $K(\Lambda, t) = \ln M(\Lambda, t)$ (for more on the CGF, consult Mood et al. (1974)). In order to do so, the following relation is needed:

$$ \frac{1}{M} \frac{\partial^2 M}{\partial \Lambda \partial t} = \frac{\partial K}{\partial \Lambda} - \left( \frac{1}{M} \frac{\partial M}{\partial \Lambda} \right) \frac{1}{M} \frac{\partial K}{\partial \Lambda} \frac{\partial M}{\partial \Lambda} + \frac{\partial^2 K}{\partial \Lambda \partial t}, $$

where $x$ is a variable that $M$ and $K$ depend on. One can for example show:

$$ \frac{1}{M} \frac{\partial^2 M}{\partial \Lambda \partial t} = \frac{1}{M} \frac{\partial M}{\partial \Lambda} \frac{1}{M} \frac{\partial M}{\partial \Lambda} = \frac{1}{M} \frac{\partial K}{\partial \Lambda} \frac{\partial M}{\partial \Lambda} + \frac{1}{M} \frac{\partial^2 K}{\partial \Lambda \partial t}.$$ 

From Eq. 12, we thus have:

$$ \frac{\partial K(\Lambda, t)}{\partial t} = v_1 a_1 + v_2 a_2 - v_3 h w(t) + v_3 w'(t) + \frac{v_1^2}{2} a_1 
+ \frac{v_2^2}{2} d_2 + \frac{v_3^2}{2} q 
+ [v_1 b_{1,1} + v_2 b_{2,1}] \frac{\partial K}{\partial v_1} 
+ [v_1 b_{1,2} + v_2 b_{2,1}] \frac{\partial K}{\partial v_2} 
+ [v_1^2 c_1 + v_2^2 c_2] \left[ \frac{\partial^2 K}{\partial v_3^2} + \frac{\partial^2 K}{\partial \Lambda \partial v_3} \right]. $$

Neither Eq. 12 nor Eq. 13 are analytically tractable. Indeed, even the numerical solution of these equations is difficult as they are both nonlinear partial differential equations. Instead, we apply the cumulant truncation procedure (Daniels 1954) as it provides approximate, computationally efficient solutions to Eq. 13.

Note that the CGF $K(\Lambda, t) = \ln E[\exp\{v_1 \phi_1(t) + \ldots + v_m \phi_m(t)\}]$ may be written as a series expansion in terms of the model cumulants.

$$ K = \sum_{r_1=0}^\infty \sum_{r_2=0}^\infty \ldots \sum_{r_m=0}^\infty \left( \prod_{i=1}^m v_i^{r_i} \right) \frac{\kappa_{r_1,r_2,\ldots,r_m}(t)}{\prod_{i=1}^m r_i!}, $$

where:

$$ \kappa_{r_1,r_2,\ldots,r_m}(t) $$


The cumulant truncation procedure assumes that all cumulants above a certain order are zero. For this paper we shall assume all the cumulants above the second order are zero, which is equivalent to assuming that the data is normally distributed. That is $K(\Lambda, t) = \ln E[\exp\{v_1 A(t) + v_2 C(t) + v_3 T(t)\}]$ is approximated by:

$$K(\Lambda, t) \approx 1 + v_1 K_{100}(t) + v_2 K_{010}(t) + v_3 K_{001}(t) + \frac{v_1^2}{2} K_{200}(t) + \frac{v_2^2}{2} K_{020}(t) + \frac{v_2 v_3}{2} K_{002}(t) + v_1 v_2 K_{110}(t) + v_1 v_3 K_{101}(t) + v_2 v_3 K_{011}(t).$$

(15)

By substituting Eq. 15 within Eq. 13 and subsequently equating the coefficients of $(\prod_{j=1}^m v_j^r)$, it is possible to derive a system of ordinary differential equations:

$$\dot{K}_{100}(t) = a_1 + b_1 A_{100}(t) + c_1 K_{100}(t)^2 + K_{002}(t)$$
$$+ b_{1,1} K_{010}(t) + b_{1,2} K_{100}(t) + b_{1,3} K_{001}(t)$$

$$\dot{K}_{010}(t) = a_2 + b_2 A_{010}(t) + c_2 K_{010}(t)^2 + K_{002}(t)$$
$$+ b_{2,1} K_{100}(t) + b_{2,3} K_{001}(t)$$

$$\dot{K}_{001}(t) = h[K_{001}(t) - w(t)] + w'(t)$$

$$\dot{K}_{200}(t) = d_1 + 2b_1 A_{200}(t) + 4c_1 K_{100}(t) K_{100}(t)$$
$$+ 2b_{1,2} K_{100}(t) + 2b_{1,3} K_{001}(t)$$

$$\dot{K}_{020}(t) = d_2 + 2b_2 A_{020}(t) + 4c_2 K_{010}(t) K_{100}(t)$$
$$+ 2b_{2,1} K_{010}(t) + 2b_{2,3} K_{100}(t)$$

$$\dot{K}_{002}(t) = 2h K_{002}(t) + q$$

$$\dot{K}_{110}(t) = b_1 A_{110}(t) + 2c_1 K_{100}(t) K_{100}(t)$$
$$+ b_{1,2} K_{100}(t) + b_{2,1} K_{110}(t)$$
$$+ 2c_2 K_{010}(t) K_{100}(t) + b_{1,3} K_{100}(t) + b_{1,4} K_{110}(t) + b_{2,3} K_{001}(t)$$

$$\dot{K}_{101}(t) = b_1 A_{101}(t) + 2c_1 K_{100}(t) K_{100}(t)$$
$$+ b_{1,2} K_{100}(t) + b_{1,3} K_{100}(t) + b_{1,4} K_{100}(t) + h K_{101}(t)$$

$$\dot{K}_{011}(t) = b_2 A_{011}(t) + 2c_2 K_{010}(t) K_{010}(t)$$
$$+ b_{2,1} K_{010}(t) + b_{2,3} K_{001}(t) + h K_{011}(t) + b_{2,4} K_{010}(t).$$

(16)

This system of ordinary, nonlinear differential equations are much easier to solve numerically than the partial differential equation in 13: in order to predict the cumulants at a future time $t_{i+1}$, we solve the system of equations in 16 using a Runge-Kutta scheme (Butcher 2008) with the boundary conditions dictated by the process values at the present time $t_i$. These are: $K_{100}(t_i) = A(t_i)$, $K_{010}(t_i) = A(t_i)$, $K_{001}(t_i) = A(t_i)$, $K_{200}(t_i) = 0$, $K_{020}(t_i) = 0$, $K_{002}(t_i) = 0$, $K_{110}(t_i) = 0$, $K_{101}(t_i) = 0$, $K_{011}(t_i) = 0$. The resulting conditional cumulants may be plugged into a multivariate normal distribution and subsequently used as an approximation to $p(\Phi(t_{i+1})|\Phi(t_i))$. This approximation is denoted $f(\Phi(t_{i+1})|\Phi(t_i))$. By substituting $f(\Phi(t_{i+1})|\Phi(t_i))$ within Eq. 5, an approximation to the likelihood can be derived.

This approximation to the likelihood may be used within an MCMC algorithm (see Robert and Casella 2004) for a detailed description of the MCMC algorithm) in order to sample the ecological parameters from their posterior distribution and consequently derive parameter estimates. We define such an MCMC algorithm below:

1. Apply the cumulant truncation procedure to the diffusion system thus deriving a system of ordinary differential equations (such as in 16) that describe the cumulants’ evolution. Note these differential equations will depend on the model parameters $\theta$.
2. Choose an arbitrary, starting set of parameter values $\theta_0$. Set $\theta_{old} = \theta_0$.
3. Propose a jump from the old set of parameter values $\theta_{old}$ to a new set of parameter values, $\theta_{new} = \theta_{old} + \Delta \theta$ using a suitably chosen proposal distribution $q(\theta_{new}|\theta_{old})$.
4. Calculate the likelihoods $L(\theta_{old})$ and $L(\theta_{new})$. $L(\theta)$ is calculated as follows:
   i. Set the likelihood, $L(\theta) = 1$ and set the index $i$ to 1.
   ii. Use the system of ordinary differential equations from Step 1 together with the data values observed at time $t_i$, $\Phi(t_i) = \left(\phi_1(t_i), \phi_2(t_i), \ldots, \phi_m(t_i)\right)$ to predict the values of the cumulants at time $t_{i+1}$.
   iii. Given the data $\Phi(t_i)$ at time $t_i$, the transitional probability distribution at time $t_{i+1}$ can be approximated by a multivariate normal distribution $f(\Phi(t_{i+1})|\Phi(t_i))$. After substituting for the cumulants derived from step ii.
   iv. Set the likelihood $L(\theta) = L(\theta) \times$
This is an implementation of Eq. 5.

v. Set \( i = i + 1 \) and if \( i < N \) (where \( N \) is the number of data points) go to step ii.

5. Calculate the acceptance ratio \( R \). This is given by:

\[
R = \frac{L(\theta_{\text{new}}) \pi(\theta_{\text{new}}) q(\theta_{\text{old}} \mid \theta_{\text{new}})}{L(\theta_{\text{old}}) \pi(\theta_{\text{old}}) q(\theta_{\text{new}} \mid \theta_{\text{old}})},
\]

where \( \pi(\theta) \) is the prior density. We then accept the proposed parameter values \( \theta_{\text{new}} \) with probability \( \min(1, R) \). That is, set \( \theta_{\text{new}} = \theta_{\text{old}} \) with probability \( 1 - \min(1, R) \). Otherwise leave \( \theta_{\text{new}} \) unchanged.

6. Go back to step 3.

The proposal distributions used within the MCMC are normally distributed and hence symmetric. In addition, we specify priors for the model parameters \( \theta \). These priors are shown in Table 2. The priors reflect initial expectations on the behavior of the modelled system, including the optimal temperatures for the two species as well as species competition.

MCMC chains of length 300,000 are run and the first 10,000 steps are trimmed as part of the burn-in period. It took 2 hours, 23 minutes to run such a chain, implemented in R, on an Intel i5, 2.53GHz processor. However, if instead of using a Runge-Kutta scheme to solve the system of equations in 16 we rather used an Euler scheme, the run time drops to 43 minutes. Despite its slower speed, we recommend using the Runge-Kutta algorithm as it is more robust against severe non-linearities. Most scientific software will have built-in procedures to solve a system of ordinary differential equations. We provide our R code as a Supplement.

In comparison to the cumulant truncation method, using a well-known computationally efficient data augmentation method (Kalogeropoulos et al. 2011) to evaluate the likelihood resulted in the MCMC chain taking 1 hour, 41 minutes to run. The two methods are comparable in speed. In general the speed of the imputation algorithm is tied to the resolution at which the imputation is performed. The appropriate resolution to use for the imputation is determined by the dataset so it is difficult to draw general conclusions about the relative speeds of the two methods. It should however be pointed out that, unlike the imputation algorithm, the cumulant truncation algorithm introduced allows for non-zero off-diagonal terms in the model’s diffusion matrix \( \Sigma(\Phi(t), t; \theta) \). Such off-diagonal terms may be required to model an ecosystem of interest.

### RESULTS

By applying the MCMC algorithm to the dataset, parameter estimates were obtained. These estimates together with the corresponding 90% credibility intervals are shown in Table 3. A quick comparison of \( d_1 \) and \( d_2 \) reveals that *Pseudo-nitzschia australis* is the more volatile species, which is perhaps not surprising as it tends to be the more numerous species. In addition the parameters \( b_{1,2} \) and \( b_{2,1} \) are negative. This implies that the lower the current species numbers are, the greater the instantaneous growth rate will be—all else being equal. This serves to regulate the growth of both species.

We also find no evidence of a standard predator-prey interaction between the two species as neither of the parameters \( b_{1,3} \) and \( b_{2,3} \) are significantly non-zero. Their estimated posterior distributions are shown in Fig. 3. Any possible predator-prey link between *Prorocentrum micans* and *Pseudo-nitzschia australis* may be obscured by competition for nutrients and/or by the species having different ecological niches. This is a topic for future research.

Perhaps more interestingly, the model enables us to find the optimal temperatures for species growth. The optimal temperatures \( T^* \) may be found by maximizing the instantaneous growth rates in Eq. 11 with respect to the temperature. We thus have:

\[
T_{\text{australis}}^* = \frac{-b_{1,2}}{2c_1}; \quad T_{\text{micans}}^* = \frac{-b_{2,2}}{2c_2}.
\]

We can find the optimal temperatures sampled by the MCMC chain. Our estimate for the optimum temperature for *Pseudo-nitzschia aus-

Table 2. Priors for the model parameters. All the priors are taken to be normally distributed. The prior means and standard deviations are shown.

| Parameter       | Mean | SD  | Parameter       | Mean | SD  |
|-----------------|------|-----|-----------------|------|-----|
| *Pseudo-nitzschia australis* |      |     | *Prorocentrum micans* |      |     |
| \( a_1 \)       | 100  | 35  | \( a_2 \)       | 10   | 10  |
| \( b_{1,2} \)   | 3.5  | 7.5 | \( b_{2,2} \)   | 3.5  | 7.5 |
| \( c_1 \)       | 0    | 8   | \( c_2 \)       | 0    | 8   |
tralis is 12.3°C with a 90% credibility interval of (7.8°C; 16.0°C). For *Pseudo-nitzschia australis*, the estimated optimum temperature is 17.5°C with a 90% credibility interval of (12.5°C; 24.5°C). Fig. 4 shows the estimated posterior distributions for the two species. *Prorocentrum micans* seems to prefer warmer temperatures than *Pseudo-nitzschia australis*. Furthermore note that $c_1$ and $c_2$ controls the concavity of the species relationships to the temperature. As such $c_1$ and $c_2$ represent their respective sensitivities to temperature. *Pseudo-nitzschia australis* appears to be more sensitive to unfavorable temperatures. This is further evidenced by the narrower credibility interval for its optimal temperature.

### Table 3. Parameter estimates together with the corresponding 90% credibility intervals.

| Parameter | Pseudo-nitzschia australis | Prorocentrum micans | Temperature |
|-----------|----------------------------|---------------------|-------------|
| $a_1$     | 139.5                      | 8.0                 | 1.0         |
| $b_{1,1}$ | -36.9                      | 10.7                | -4.9.       |
| $b_{2,1}$ | 14.4                       | 6.6                 | 3.6         |
| $c_1$     | -0.58                      | 9.2                 | -1.53       |
| $d_1$     | 293.5                      | 75.2                | 2.92        |
| $a_2$     | 10.7                       | 26.4                | 14.5        |
| $b_{2,2}$ | 10.7                       | 3.6                 | -0.52       |
| $c_2$     | -0.58                      | 9.2                 | 2.92        |
| $d_2$     | 93.8                       | 113.6               | 14.5        |
| $h$       | -39.0                      | -51.6               | 130.5       |
| $q$       | 98.6                       | 171.4               | -10.0       |

### Discussion

Diffusion models are capable of capturing the behavior of a vast array of ecological phenomena (Drake and Griffen 2009). It is demonstrated how a nonlinear, multivariate diffusion model may be fitted to an ecological time series through the use of the cumulant truncation procedure. The method remains applicable even with large, non-equal time lapses between observations within a time series (Varughese 2013). This makes it suitable for many ecological datasets, including the plankton dataset investigated in this paper.

The bloom dynamics exhibited by both species investigated are typical for r-selected species (Smayda and Reynolds 2001). In addition, there
is evidence that the growth of the two species is regulated as the parameters $b_{1,1}$ and $b_{2,1}$ are negative. We also find no evidence of predation of the diatom *Pseudo-nitzschia australis* by the dinoflagellate *Prorocentrum micans*. Any possible predation could possibly be obscured by competition or a more complex interaction between the two species. The results however suggest that the dinoflagellate prefers warmer temperatures than the diatom. *Pseudo-nitzschia australis* is the more volatile species and appears to be more vulnerable to unfavorable temperatures. This suggests that the two species occupy different niches within the marine environment (Smayda and Reynolds 2001).

By performing statistical inference upon the diffusion model, we were able to learn much about the species relationship to surface water temperature as well as the nature of inter-and intra-species interactions. Diffusion models may also be used to assess boundary crossing events and hence are useful in population viability analysis (Gaston and Nicholls 1995). They are also useful for forecasting population trajectories under varying conditions. This makes diffusion models highly valuable tools in the management and understanding of ecosystems.

**ACKNOWLEDGMENTS**

Melvin Varughese was hosted by the Department of Mathematics and Statistics at the University of Melbourne whilst working on this project. Etienne Pienaar and Melvin Varughese were both funded by the National Research Foundation of South Africa. We thank the editor and the reviewers for their constructive comments that have greatly improved this paper.

**LITERATURE CITED**

Ait-Sahalia, Y. 2002. Maximum likelihood estimation of discretely sampled diffusions: a closed-form approximation approach. Econometrica 70:223–262.

Allen, L., and E. Allen. 2003. A comparison of three different stochastic population models with regard to persistence time. Theoretical Population Biology 64:439–449.

Ananthasubramaniam, B., R. Nisbet, W. Nelson, E. McCauley, and W. Gurney. 2011. Stochastic growth reduces population fluctuations in daphnia-algal systems. Ecology 92:362–372.

Burkholder, J., P. Glibert, and H. Skelton. 2008. Mixotrophy, a major mode of nutrition for harmful algal species in eutrophic waters. Harmful Algae 8:77–93.

Butcher, J. 2008. Numerical methods for ordinary differential equations. John Wiley and sons, Chichester, UK.

Carpenter, S., and W. Brock. 2006. Rising variance: a
leading indicator of ecological transition. Ecology Letters 9:311–318.

Cochlan, W., J. Herndon, and R. Kudela. 2008. Inorganic and organic nitrogen uptake by the toxigenic diatom *Pseudo-nitzschia australis* (Bacillariphycceae). Harmful Algae 8:111–118.

Cox, D., and H. Miller. 1977. The theory of stochastic processes. CRC Press, Boca Raton, Florida, USA.

Daniels, H. 1954. Saddlepoint approximations in statistics. Annals of Mathematical Statistics 25:631–650.

Drake, J., and B. Griffen. 2009. Speed of expansion and extinction in experimental populations. Ecology Letters 12:772–778.

Gaston, K., and A. Nicholls. 1995. Probable times to extinction of some rare breeding bird species in the united kingdom. Proceedings of the Royal Society B 259:119–123.

Guiry, M. 2012. How many species of algae are there? Journal of Phycology 48:1057–1063.

Huang, X. 2012. Quasi-maximum likelihood estimation of multivariate diffusions. Studies in Nonlinear Dynamics & Econometrics 5:390–455.

Humbert, J., L. Mills, J. Horne, and B. Dennis. 2009. A better way to estimate population trends. Oikos 118:1940–1946.

Jeong, H., Y. D. Yoo, J. S. Kim, K. A. Seong, N. S. Kang, and T. H. Kim. 2010. Growth, feeding and ecological roles of the mixotrophic and heterotrophic dinoflagellates in marine planktonic food webs. Ocean Science Journal 45:65–91.

Kalogeropoulos, K., P. Dellaportas, and G. Roberts. 2011. Likelihood-based inference for correlated diffusions. Canadian Journal of Statistics 39:52–72.

Kleinhans, D., and R. Friedrich. 2007. Maximum likelihood estimation of drift and diffusion functions. Physics Letters A 368:194–198.

Knape, J., and P. de Valpine. 2012. Fitting complex population models by combining particle filters with Markov chain Monte Carlo. Ecology 93:256–263.

Lundholm, N. and . Moestrup. 2006. The biogeography of harmful algae. Ecology of Harmful Algae 189:23–35.

Mao, X., G. Marion, and E. Renshaw. 2002. Environmental brownian noise suppresses explosions in population dynamics. Stochastic Processes and Their Applications 97:95–110.

Mao, X., C. Yuan, and J. Zou. 2005. Stochastic differential delay equations of population dynamics. Journal of Mathematical Analysis and Applications 304:296–320.

Mood, A., F. Graybill, and D. Boes. 1974. Introduction to the theory of statistics. McGraw-Hill, New York, New York, USA.

Ovaskainen, O. 2004. Habitat-specific movement parameters estimated using mark-recapture data and a diffusion model. Ecology 85:242–257.

Robert, C., and G. Casella. 2004. Monte Carlo statistical methods. Springer, New York, New York, USA.

Roberts, G., and O. Stramer. 2001. On inference for partially observed nonlinear diffusion models using the metropolis-hastings algorithm. Biometrika 88:603–621.

Rose, J., and D. Caron. 2007. Does low temperature constrain the growth rates of heterotrophic protists? Evidence and implications for algal blooms in cold waters. Limnology & Oceanography 52:886–895.

Smyda, T. 2002. Adaptive ecology, growth strategies and the global bloom expansion of dinoflagellates. Journal of Oceanography 58:281–294.

Smyda, T., and C. Reynolds. 2001. Community assembly in marine phytoplankton: Application of recent models to harmful dinoflagellate blooms. Journal of Plankton Research 23:447–461.

Stirzaker, D. 2005. Stochastic processes and models. Oxford University Press, New York, New York, USA.

Takimoto, G. 2009. Early warning signals of demographic regime shifts in invading populations. Population Ecology 51:419–426.

Thomas, W., T. Zavoico, and C. Hewes. 2001. Historical phytoplankton species time-series data (1917–1939) from the north american pacific coast. Technical report. Scripps Institute of Oceanography. http://escholarship.org/uc/item/39v2v8w5

Tufto, J., R. Lande, T. Ringsby, S. Engen, B. Saether, T. Walla, and P. DeVries. 2012. Estimating brownian motion dispersal rate, longevity and population density from spatially explicit mark-recapture data on tropical butterflies. Journal of Animal Ecology 81:756–769.

Varughese, M. M. 2009. On the accuracy of a diffusion approximation to a discrete state-space markovian model of a population. Theoretical Population Biology 76:241–247.

Varughese, M. M. 2011. A framework for modelling ecological communities and their interactions with the environment. Ecological Complexity 8:105–112.

Varughese, M. M. 2013. Parameter estimation for multivariate diffusion systems. Computational Statistics and Data Analysis 57:417–428.

Vindenes, Y., S. Engen, and B. Saether. 2011. Integral projection models for finite populations in a stochastic environment. Ecology 92:1146–1156.

Wojtkiewicz, S., and L. Bergman. 2000. Numerical solution of high dimensional fokkerplanck equations. In Proceedings of the 8th Specialty Conference on Probabilistic Mechanics and Structural Reliability. Paper PMC2000-167.
Supplemental Material

Supplement

An R script and data file for the analysis conducted in the main paper (Ecological Archives C004-011-S1).