Ecological Modelling 222 (2011) 1793–1799

Contents lists available at ScienceDirect

Ecological Modelling

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

Parameter estimation in a simple stochastic differential equation for phytoplankton modelling

Jan Kloppenborg Møller[a,b,*], Henrik Madsen[a], Jacob Carstensen[b]

[a] DTU Informatics, Richard Pedersens Plads, Technical University of Denmark, Building 321, DK-2800 Lyngby, Denmark
[b] National Environmental Research Institute, Fredriksborgvej 399, DK-4000 Roskilde, Denmark

ARTICLE INFO

Article history:
Received 11 November 2010
Received in revised form 14 March 2011
Accepted 18 March 2011
Available online 12 April 2011

Keywords:
Phytoplankton modelling
Stochastic differential equations
Parameter estimation
Extended Kalman filter
Maximum likelihood estimation

ABSTRACT

The use of stochastic differential equations (SDEs) for the simulation of aquatic ecosystems has attracted increasing attention in recent years. The SDE setting also provides the opportunity for statistical estimation of ecosystem parameters. We present an estimation procedure, based on Kalman filtering and likelihood estimation, which has proven useful in other fields of application. The estimation procedure is presented and the development from ordinary differential equations (ODEs) to SDEs is discussed with emphasis on autocorrelated residuals, commonly encountered with ODEs. The estimation procedure is applied to a simple nitrogen-phytoplankton model, with data from a Danish estuary (1988–2006). The resulting SDE is simple enough to have a well-known stationary distribution and this distribution is presented and compared with observed phytoplankton data.

1. Introduction

Ecosystems, and marine ecosystems in particular, are complex mosaics of interconnected processes, with many known and unknown drivers affecting the systems. Marine ecosystems have traditionally been modelled by means of Ordinary Differential Equations (ODEs) that have gradually evolved from simple descriptions of the nutrient–phytoplankton interaction (NP models) to include an increasing number of components, e.g. zooplankton (NPZ models), detritus (NPZD models), benthic fauna and vegetation, as well as fish. Moreover, nutrients and organisms have also gradually been partitioned into different constituents and species groups, often in response to inadequate description of observed dynamics. The consequence of employing such a detailed mechanistic approach is increasing model complexity with an escalating number of unknown parameters that are calibrated using values obtained from the literature or tuned to mimic observations. In addition these observations are often aggregates of several states, according to the modeller's subjective assessment. For example, Fasham et al. (1990) considered a relatively simple NPZD model with 7 states (3 different nitrogen pools, phytoplankton, zooplankton, detritus and bacteria) and 25 parameters, whereas Bartell et al. (1999) introduced a flexible, though complex, modelling framework and applied it to Canadian lakes using 44 states for the biological components. Ecological models seem to have grown in size and complexity as computational constraints have been alleviated and understanding of sub-processes have grown over time. Matear (1995) employs a more objective stochastic optimisation (simulated annealing), but the underlying system is still deterministic.

Scientists have come to realise that even the most complex ecosystem model will not be able to capture all mechanisms and drivers of the real ecosystem, Dowd (2006, 2007) presents NPD models with differential random forcing, the formulation of the noise does however not allow a stochastic differential equation formulation (Øksendal, 2003).

Drivers that are unobservable or not accounted for in a model will lead to systematic deviations from the model in the form of autocorrelated residuals between observations and short term predictions. This kind of autocorrelation is actually also evident in the results presented by Fasham et al. (1990, Figures 4 and 5 in the reference). In fact, these residuals can be modelled as stochastic perturbations working within the model. In this case, ODE models with stochastic input of internal randomness are referred to as stochastic differential equations (SDEs) (e.g. Øksendal, 2003). This internal stochastic perturbation will also, to some extent, be able to indirectly capture drivers of the system not implicitly contained in the model formulation.

SDEs are an emerging field and the use of SDEs in mathematical finance and option pricing (e.g. Black and Scholes models (see

* Corresponding author at: DTU Informatics, Richard Pedersens Plads, Technical University of Denmark, Building 321, DK-2800 Lyngby, Denmark.
Tel.: +45 4525 3375.
E-mail address: jkm@imm.dtu.dk (J.K. Møller).

0304-3800/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2011.03.025

e.g. Øksendal, 2003) is the standard example in many text books (e.g. Øksendal, 2003). Early applications of SDEs in other fields can be found in Madsen et al. (1987) (climatology), Madsen and Holst (1995) (engineering), and Jacobsen and Madsen (1996) (oxygen level in a stream). More recently SDEs have been applied to pharmaceutical problems, e.g. Tornøe et al. (2004). Modelling of motion patterns for larger animals has also been the subject of SDE-modelling in recent years, e.g. Brillinger et al. (2002) and Pedersen et al. (2008).

The use of stochastic differential equations (SDEs) to introduce stochastic forcing in NP-like ecosystem models has attracted increasing attention over recent years. Carpenter and Brock (2006) and Guttal and Jayaprakash (2008) are examples, where SDEs are used in analysis of non-linear stochastic systems with emphasis on regime shifts, i.e. these studies analyse known deterministic regime shift models with stochastic forcing. These studies are, however, pure simulation studies of how random behaviour affects the dynamics of regime shift models. Stollenwerk et al. (2001) present estimation, of phytoplankton in an SDE-based model, based on stationary distribution and consequently on data from the growth season only. The present study use data and SDEs for NP modelling and parameter estimation based on 19 years of data including the winter period. The aim is to provide a simple example to illustrate the usefulness of SDEs for modelling of phytoplankton.

The paper is organised in the following way: Section 2 gives a short introduction to SDEs, focusing on the development from ODEs to SDEs, and introduces the statistical method used for parameter and state estimation in SDEs, Section 3 provides a small simulation example to illustrate the presented theory, and Section 4 presents an example of a simple NP-SDE model with data from an estuary in the northern Denmark. The SDE model includes phytoplankton as the state with water column nitrogen and global radiation (total (i.e. direct and diffuse) incoming solar radiation) as drivers.

2. Stochastic differential equations and ordinary differential equations

As a general rule it is only possible to observe continuous time processes in discrete time. Let $\boldsymbol{x}_t \in \mathcal{X} \subset \mathbb{R}^n$ be the continuous time state variable which is observed through an observation equation in discrete time, and let $\boldsymbol{y}_k \in \mathcal{Y} \subset \mathbb{R}^l$ denotes the observation at time t_k ($k \in \{0, \ldots, N\}$), let the observation equation be given by

$$\boldsymbol{y}_k = \boldsymbol{h}(\boldsymbol{x}_k, \boldsymbol{u}_k, t_k, \boldsymbol{\theta}, \boldsymbol{e}_k), \tag{1}$$

where $\boldsymbol{x}_k$ and $\boldsymbol{u}_k \in \mathcal{U} \subset \mathbb{R}^r$ is the state variable and the inputs (forcing or control variables) at time $t = t_k$, $\boldsymbol{e}_k \in \mathbb{R}^l$ is the random observation error, $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$ is a set of parameters to be estimated and $\boldsymbol{h}(\cdot) \in \mathbb{R}^l$ is the function that links the states to the observations. Simple forms of $\boldsymbol{h}(\cdot)$ include the identity link ($\boldsymbol{h}(\cdot) = \boldsymbol{x}_{t_k} + \boldsymbol{e}_k$) and the loglink ($\boldsymbol{h}(\cdot) = \log(\boldsymbol{x}_{t_k}) + \boldsymbol{e}_k$ (if $\boldsymbol{x}_t > \boldsymbol{0}$)), with $\boldsymbol{e}_k \sim N(0, \boldsymbol{S}_k)$ and $N(\cdot)$ is the normal distribution.

2.1. Ordinary differential equation representation

In the ordinary differential equation setting the evolution in time of the state variable is given by the deterministic system equation

$$d\boldsymbol{x}_t = \boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{u}_t, t, \boldsymbol{\theta})dt, \tag{2}$$

where $t \in \mathbb{R}$ is time (the structure of) $\boldsymbol{f}(\cdot) \in \mathbb{R}^n$ is deduced from physical (or biological) knowledge of the system, and $\boldsymbol{u}_t$ and $\boldsymbol{\theta}$ are similar to the input and parameters presented in the observation Eq. (1).

If $\boldsymbol{e}_k$ takes a simple form (i.e. additive and Gaussian) and $\boldsymbol{x}_t$ follows the deterministic formulation (2), then the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is equivalent to minimising the weighted sum of squared observation errors ($\boldsymbol{e}_k^T \boldsymbol{S}_k^{-1} \boldsymbol{e}_k$), where $\boldsymbol{S}_k$ is the observation covariance matrix for the kth observation.

2.2. Stochastic differential equation representation

Natural systems are subject to random perturbation, such as random variation of the input (specified by $\boldsymbol{u}_t$) or non-specified random forcing, e.g. processes not specified in the model description, working within the system. Such perturbations create autocorrelated noise in the observations ($\boldsymbol{y}_k$), which cannot be captured by Eqs. (1) and (2), since observation noise is present only. Further, when the parameters, $\boldsymbol{\theta}$ in Eq. (2), have been estimated then the uncertainty of a forecast will be independent of the forecast horizon, which is somewhat counterintuitive.

SDEs can be formulated by introducing a noise term, perturbing the differential of $\boldsymbol{x}_t$ (Øksendal, 2003)

$$\frac{d\boldsymbol{x}_t}{dt} = \boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{u}_t, t, \boldsymbol{\theta}) + \boldsymbol{\sigma}(\boldsymbol{x}_t, \boldsymbol{u}_t, t, \boldsymbol{\theta})\boldsymbol{w}_t, \tag{3}$$

where $\boldsymbol{w}_t \in \mathbb{R}^m$ is an m-dimensional standard Wiener process and $\boldsymbol{\sigma}(\cdot) \in \mathbb{R}^{n \times m}$ is a matrix function (Øksendal, 2003). Multiplying with dt gives the standard SDE formulation

$$d\boldsymbol{x}_t = \boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{u}_t, t, \boldsymbol{\theta})dt + \boldsymbol{\sigma}(\boldsymbol{x}_t, \boldsymbol{u}_t, t, \boldsymbol{\theta})d\boldsymbol{w}_t, \tag{4}$$

$\boldsymbol{\sigma}(\boldsymbol{x}_t, \boldsymbol{u}_t, t, \boldsymbol{\theta})$ is referred to as the diffusion term, and $\boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{u}_t, t, \boldsymbol{\theta})$ is referred to as the drift term. The solution to (4) is a stochastic process with transition probabilities given by the Fokker–Planck equation (e.g. Klebaner, 2005). Furthermore, one path of the solution is an autocorrelated stochastic process, which can be realised by considering Eq. (3) where the increments of $\boldsymbol{x}_t$ are subject to random perturbations.

2.3. Parameter estimation in SDEs

Estimation of parameters in SDEs is a difficult task because evaluation of the likelihood of observation requires knowledge about the transition densities between discrete time observations. Transitions densities are generally unknown except for very simple SDEs and approximate methods has to applied. To enable general estimation of SDEs simulation based methods has to be applied (e.g. Nicolau, 2002), including sampling techniques (e.g. Pastorello and Rossi, 2010) and particle filters (e.g. Givon et al., 2009). While simulation based method has the advantage of dealing effectively with general differential noise terms like Poisson noise (e.g. Givon et al., 2009), the draw- back is the computational effort needed in the simulation part of the algorithms.

Closed form likelihood expansions are also available (e.g. Aït-Sahalia, 2008), while tractable from a computational point of view, these are complicated to apply involving Hermite series expansion of the local log-likelihood function. Further the assumption is that all states are observed, which will not be the case in general. In the present study we will base the analysis on the Extended Kalman filter (EKF), where the prediction of the first and second moments of the process is based on a set of ODEs, which are solved numerically. While the methodology is well-known (Kristensen et al., 2004) it has, to our knowledge, not been applied to marine ecosystems previously. A key advantage of the methodology is that the method is implemented in the easily accessible open source software CTSM[1] (Kristensen and Madsen, 2003; Kristensen et al., 2004).

[1] The software is available at www2.imm.dtu.dk/~ctsm.

2.3.1. Likelihood estimation by EKF

Consider the continuous-discrete time stochastic state-space model formulation in Eqs. (1) and (4)

$$d\boldsymbol{x}_t = \boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{u}_t, t, \boldsymbol{\theta})dt + \boldsymbol{\sigma}(\boldsymbol{x}_t, \boldsymbol{u}_t, t, \boldsymbol{\theta})d\boldsymbol{w}_t, \tag{5}$$

$$\boldsymbol{y}_k = \boldsymbol{h}(\boldsymbol{x}_{t_k}, \boldsymbol{u}_{t_k}, t_k, \boldsymbol{\theta}, \boldsymbol{e}_k), \tag{6}$$

where parameters and variables are as described in Section 2.

For the continuous-discrete time stochastic state-space model (5) and (6), the problem that needs to be solved is: find the set of parameters ($\hat{\boldsymbol{\theta}}$) such that some objective function is maximised given a set of observations $\mathcal{Y}_N = \{\boldsymbol{y}_0, \ldots, \boldsymbol{y}_N\}$. A natural choice of such an objective function is the joint probability density of the observations, considered as a function of the unknown parameters ($\boldsymbol{\theta}$) (the likelihood function), i.e.

$$L(\boldsymbol{\theta}; \mathcal{Y}_N) = p(\mathcal{Y}_N|\boldsymbol{\theta}) = \left(\prod_{k=1}^{N} p(\boldsymbol{y}_k|\mathcal{Y}_{k-1}, \boldsymbol{\theta})\right) p(\boldsymbol{y}_0|\boldsymbol{\theta}), \tag{7}$$

where Bayes rule has been applied recursively to form the product of conditional densities (e.g. Madsen, 2008). In principle the solution of this problem would be an application of the Fokker–Planck equation for predictions and Bayes rule for updating given a new observation. Such a strategy is, however, infeasible, except for systems with very simple structures of the system equation (5), because it involves the solution of a very complex partial differential equation.

The estimation procedure, which will be introduced in the following, relies on an implementation of EKF techniques (Jazwinski, 1970). This implementation requires the system and observation equations to have the form

$$d\boldsymbol{x}_t = \boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{u}_t, t, \boldsymbol{\theta})dt + \boldsymbol{\sigma}(\boldsymbol{u}_t, t, \boldsymbol{\theta})d\boldsymbol{w}_t, \tag{8}$$

$$\boldsymbol{y}_k = \boldsymbol{h}(\boldsymbol{x}_{t_k}, \boldsymbol{u}_{t_k}, t_k, \boldsymbol{\theta}) + \boldsymbol{e}_k, \tag{9}$$

where 1) the diffusion matrix is quadratic, i.e. $\boldsymbol{\sigma}(\cdot) \in \mathbb{R}^{n\times n}$ and $\boldsymbol{w}_t \in \mathbb{R}^n$, 2) the diffusion term is not allowed to depend on the state, and 3) the observation noise is additive Gaussian white noise ($\boldsymbol{e}_k \in N(\boldsymbol{0}, \boldsymbol{S}_k(\boldsymbol{\theta}, \boldsymbol{u}_k))$). In a weak solution sense (equality in distribution) (see Øksendal (2003) for a discussion of weak and strong solutions), 1) is not a restriction since $\boldsymbol{\sigma}(\cdot)$ can only be identified up to the "square root" of $\boldsymbol{\sigma}\boldsymbol{\sigma}^T(\cdot)$, 2) is clearly a restriction in the multivariate case, but to some extent this can be dealt with by transformations, and a class of diffusion processes can be dealt with by transformations (e.g. all processes with diffusion given by $\mathrm{diag}(\boldsymbol{x}_t)\boldsymbol{\sigma}(t, \boldsymbol{\theta})$; Luschgy and Pagés, 2006). Finally, the restriction (3) should be dealt with by transformation of the observation if possible.

Since the systems (8) and (9) are driven by Wiener noise, and the observation noise is additive Gaussian, a reasonable local approximation of the conditional densities in (7) is the Gaussian distribution, which is completely characterised by its mean and covariance.

The one-step prediction, covariance and the innovation are defined as

$$\hat{\boldsymbol{y}}_{k|k-1} = E\{\boldsymbol{y}_k|\mathcal{Y}_{k-1}, \boldsymbol{\theta}\}, \tag{10}$$

$$\boldsymbol{R}_{k|k-1} = V\{\boldsymbol{y}_k|\mathcal{Y}_{k-1}, \boldsymbol{\theta}\}, \tag{11}$$

$$\boldsymbol{\epsilon}_k = \boldsymbol{y}_k - \hat{\boldsymbol{y}}_{k|k-1}, \tag{12}$$

where $E\{\cdot\}$ is the expectation and $V\{\cdot\}$ is the variance. Using this notation the likelihood can be written as

$$L(\boldsymbol{\theta}; \mathcal{Y}_N) = \left(\prod_{k=1}^{N} \frac{\exp((1/2)\boldsymbol{\epsilon}_k^T \boldsymbol{R}_{k|k-1}^{-1}\boldsymbol{\epsilon}_k)}{\sqrt{\det(\boldsymbol{R}_{k|k-1})(2\pi)^l}}\right) p(\boldsymbol{y}_0|\boldsymbol{\theta}), \tag{13}$$

where l is the dimension of the sample space (see Eq. (1)) and $(\cdot)^T$ is the vector transpose. The actual optimisation is done in the log-domain and the (approximate) maximum likelihood estimate of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\arg\max}\{\log(L(\boldsymbol{\theta}; \mathcal{Y}_N))\}. \tag{14}$$

The Kalman gain is essential for the state updating procedure. The Kalman gain governs how much the one-step prediction ($\hat{\boldsymbol{x}}_{k|k-1}$) should be adjusted to form the reconstruction ($\hat{\boldsymbol{x}}_{k|k}$) of the state based on the observation, and is given by

$$\boldsymbol{K}_k = \boldsymbol{P}_{k|k-1}\boldsymbol{C}^T\boldsymbol{R}_{k|k-1}^{-1}, \tag{15}$$

where $\boldsymbol{C}$ is the first order Taylor expansion (the Jacobian) of $\boldsymbol{h}$ and $\boldsymbol{P}_{k|k-1}$ is the covariance of the one-step prediction. Note that the Kalman gain is proportional to the information ($\boldsymbol{R}_{k|k-1}^{-1}$) provided by the kth observation. The state reconstruction is given by

$$\hat{\boldsymbol{x}}_{k|k} = \hat{\boldsymbol{x}}_{k|k-1} + \boldsymbol{K}_k\boldsymbol{\epsilon}_k, \tag{16}$$

$$\boldsymbol{P}_{k|k} = \boldsymbol{P}_{k|k-1} - \boldsymbol{K}_k\boldsymbol{R}_{k|k-1}^{-1}\boldsymbol{K}_k^T, \tag{17}$$

i.e. a combination of the predicted state and the information obtained by the kth observation ($\boldsymbol{y}_k$). The state predictions are governed by a set of ordinary differential equations (Kristensen et al., 2004; Kristensen and Madsen, 2003).

In addition to the state reconstruction and parameter estimates discussed above, the optimisation of the likelihood function provides an estimate of the parameter covariance given by the negative inverse Hessian of the log-likelihood evaluated at the optimal parameter values. As the estimation is based on the maximum likelihood, the procedure allows for likelihood ratio tests of nested models and t-tests for all estimated parameters.

3. A simulation example

This section presents a simple simulation example, with synthetically generated observation. The example resembles some features of the case study presented in Section 4, and illustrates the points discussed above. Consider the SDE

$$dx_t = \left[\sin\left(\frac{2\pi}{12}t\right) + 1 - ax_t\right]dt + \sigma x_t dw_t, \tag{18}$$

x_t can be considered as an ecosystem component (e.g. phytoplankton) with a periodic growth process ($\sin((2\pi/12)t)+1$), which is independent of the state, a death-rate (a), and a diffusion term that is proportional to the state of the system. The solution to $dx_t = -ax_t dt + \sigma x_t dw_t$, for each fixed time horizon T, is a log-normal distributed random variable if the initial state (x_0) is larger than zero (e.g. Øksendal, 2003, Example 5.1.1). Therefore adding a positive forcing, will still guarantee that $x_t > 0\,\forall\, t$, if $x_0 > 0$. Assume that the observation equation is given by

$$\log(y_k) = \log(x_{t_k}) + e_k, \tag{19}$$

this implies that the standard deviation of the observation noise is proportional to the state of the system. The synthetic data are a realisation of the stochastic process defined by (18) and (19) with $a = 0.5$, $\sigma = 0.2$ and $e_k \sim N(0, 0.1^2)$. The simulation of (18) is performed by the Euler approximation scheme (Kloden and Platen, 1999) with $\Delta t = 10^{-4}$, and the sampling frequency of the synthetic observations is $3/\pi$.

As discussed above, the ODE solution (Fig. 1) is a deterministic function, with autocorrelated residuals. The SDE solution is a stochastic process represented by the expectation (Fig. 1) and covariance (not shown) of the state given all observations. Clearly this expectation captures the autocorrelation of the underlying process quite well (Fig. 1).

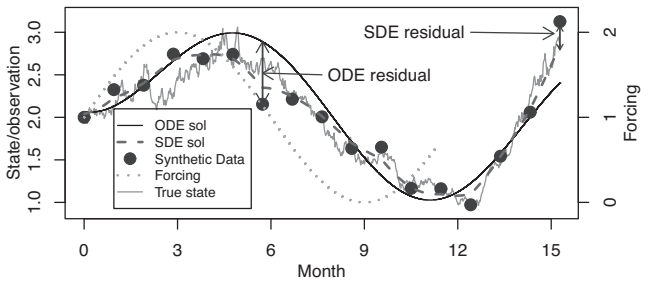

Fig. 1. Simulation results from the system described in Eqs. (18) and (19). "ODE sol" refers to the least square solution of (18) and (19) when σ is assumed to be zero, "SDE sol" refers to the smoothen state, "Synthetic Data" refers to one realisation of the stochastic process (18) and (19), "Forcing" is the input function in (18) ($\sin(2\pi t/12)+1$) and "True state" refers to one realisation of (18).

The example demonstrates that the SDE solution captures the dynamics of the underlying process better than the ODE solution (the residual sum of square is 0.18 and 1.60 for the SDE solution and the ODE solution, respectively). When SDEs are used for long-term forecasts autocorrelation between the residuals will be observed, since x_t is an autocorrelated process, see e.g. Madsen (2008) for this result in linear time series analysis. However, the distributional properties of long-term forecasts will be captured better with SDEs than with ODEs.

The estimated death-rate was 0.508 ($\pm$0.14) and 0.536 ($\pm$0.04) for the SDE model and the ODE model, respectively. Thus, the ODE solution deviates substantially from the true value (0.5) and the 95% confidence interval does not even contain the true value, indicating that the estimate of the dynamics of the process is also better when taking the correlation structure into account.

4. Skive Fjord case study

This section presents a simple model to describe the phytoplankton nitrogen dynamics in an estuary located in the northern Denmark. The model aims at describing total phytoplankton nitrogen ($X_{p,t}$) as a function of total nitrogen in the water column ($U_{w,t}$) and incoming global radiation ($U_{gr,t}$). The period with overlapping time series of input data is September 18th 1987 through December 18th 2006, which will be the modelling period.

4.1. Data

Skive Fjord has been extensively monitored during the Danish National Aquatic Monitoring and Assessment Program (DNAMAP), where various ecosystem components and water-chemistry variables have been recorded since the 1980s.

The data set includes chlorophyll in μg chla/l, which is converted to nitrogen units using the standard chlorophyll to carbon ratio of 1:50 (weight) (e.g. Pedersen et al. (2010), which report a ratio of 1:47), the Redfield ratio (C:N = 106:16 (M)), and assuming that the monitoring station is representative of the entire estuary (the observation is denoted $Y_{p,t}$). Because nitrogen in the water column acts as an input to the system, missing observations are not allowed and are filled in by linear interpolation between data points.

Global radiation data (provided by the Danish Meteorological Institute) are available from two sites around Skive Fjord and reported on an hourly basis. The global radiation data contain both missing observations and what we will refer to as "false zeros". A false zero is when global radiation equal to zero is reported during daytime. To identify such points, a general yet simple periodic function for global radiation is fitted to the non-missing data

$$f_{gr}(t;\boldsymbol{\theta}) = \left(a_0 + a_1 \sin\left(\frac{2\pi}{P_{\text{year}}}t + \phi_1\right) + a_2 \sin\left(\frac{2\pi}{P_{\text{day}}}t + \phi_2\right)\right)_+, \tag{20}$$

where $(x)_+ = \max(x, 0)$, $P_{\text{year}} = 24 \times 365.25$ is the average number of hours in one year and $P_{\text{day}} = 24$ is the number of hours in one day. If the observation at time t_{i_0} is zero and $f_{gr}(t_i) > 0$ for $i \in \{i_0 - 1, i_0, i_0 + 1\}$ then the observation is considered a false zero and marked as missing. The number of observations removed in this way is 144 out of a total of about 270×10^3 observations.

The hourly global radiation is found as a simple average (over stations) of the non-missing observations at each time point. As global radiation acts as an input, missing observations are not allowed (the number of missing observations is about 1% of the total number of observations). If the sequence of missing data is shorter than three, or the same hour of the day before and after is available, then the missing observations are filled in by linear interpolation. Remaining gaps in data are filled in by equating with $f_{gr}(t)$, and $U_{gr,t}$ is created by average daily global radiation.

The seasonal variation in both input variables and phytoplankton nitrogen is evident (Fig. 2), but strong fluctuations overlaying this signal are apparent.

4.2. A simple SDE-model

The simple model set up is a constant mortality rate and a growth process which is a function of available nitrogen and global radiation

$$\frac{dX_{p,t}}{dt} = b(U_{w,t}, U_{gr,t}) - aX_{p,t} + \text{noise}, \tag{21}$$

where $X_{p,t}$ is total N in phytoplankton, and the growth process $b(U_{w,t}, U_{gr,t})$ and the mortality rate (a) are both strictly positive. The growth process will be assumed to be proportional to the available nitrogen and the inflow of solar energy, i.e.

$$b(U_{w,t}, U_{gr,t}) = b_0 U_{w,t} U_{gr,t}, \tag{22}$$

where $b_0 > 0$ is a constant. This formulation governs a stochastic process and a natural requirement for the process is that $P(X_{p,t} < 0) = 0 \; \forall \, t$. A formulation that meets this constraint is

$$dX_{p,t} = (b_0 U_{w,t} U_{gr,t} - aX_{p,t})dt + \sigma_x X_{p,t} dw_t, \tag{23}$$

where w_t is the standard Wiener process. In this formulation the diffusion for the process is proportional to the level of the process, i.e. the higher the abundance of phytoplankton the higher the variance (in absolute terms). This choice of noise process is the simplest in terms of estimation, because the state space of the Lamperti transformed process (see below) is the entire real axis, but it is also in good agreement with the generality of the log-normal distribution (Limpert et al., 2001) and coincide with the choice in Dowd (2006).

Assuming $b(U_{w,t}, U_{gr,t})$ is constant, then Eq. (23) is a special case of the Pearson diffusion, which is defined as (Iacus, 2008)

$$dX_t = -\theta(X_t - \mu)dt + \sqrt{2\theta(\sigma_1 X_t^2 + \sigma_2 X_t + \sigma_3)}dw_t, \tag{24}$$

which does not have a closed form solution for the transient. For $\sigma_2 = \sigma_3 = 0, \theta > 0, \mu > 0$, and $\sigma_1 > 0$, the stationary distribution for this process is, however, known to be an inverse Gamma distribution with shape parameter $1 + \sigma_1^{-1}$ and scale parameter μ/σ_1 (see Iacus 2008, observe the misprint on p. 54 in the reference though). With $b(\cdot)$ a function of t (the inputs vary in time), the process will always be in the transient, nonetheless the process should be close to the stationary distribution, if the time constant for the system ($1/a$) is

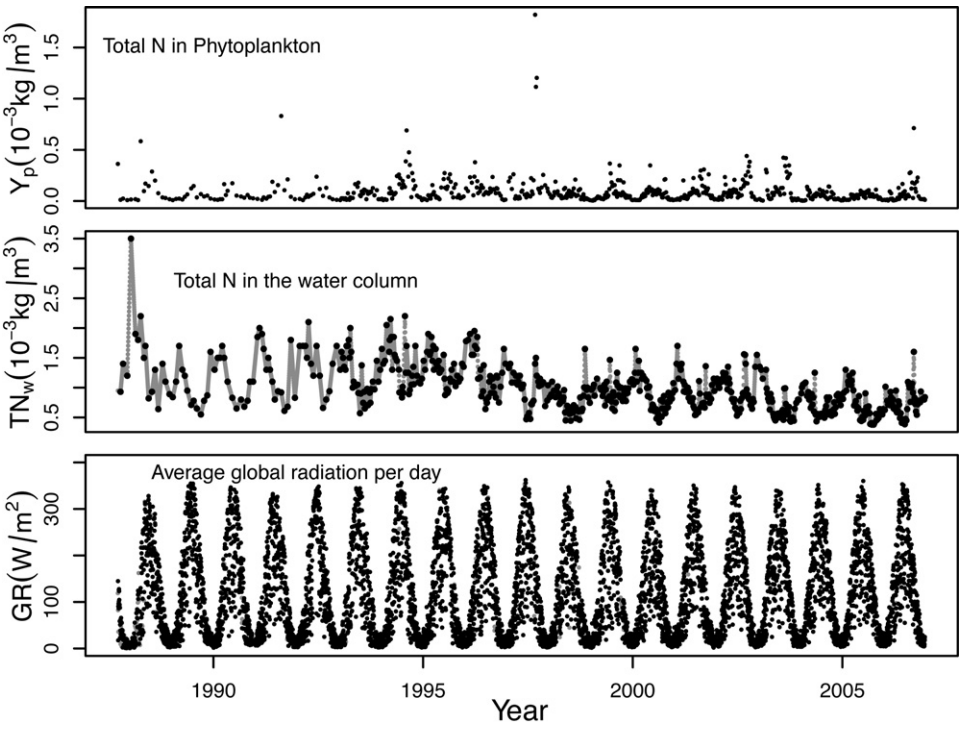

Fig. 2. Observations of total nitrogen in phytoplankton, total nitrogen in the water column and global radiation. Black dots are observations, while grey dots/lines are interpolated data points.

fast compared to the variation in $b(\cdot)$. Suppressing the time index and comparing (23) and (24) gives

$$\theta = a, \quad \mu = \frac{b_0 U_w U_{gr}}{a}, \quad \sigma_1 = \frac{\sigma_x^2}{2a}, \tag{25}$$

which implies that for each fixed t the stationary distribution for $X_{p,t}$ follows an inverse gamma distribution with shape parameter $1 + (a/\sigma_x^2)$ and scale parameter $2b_0 U_{w,t} U_{gr,t}/\sigma_x^2$.

As discussed in Section 2.3 the diffusion term is required to be independent of the state of the system, which is not the case for the presented system. Fortunately it is always possible to transform a one-dimensional system with only one continuously differentiable diffusion term into a system with constant diffusion, by applying the Lamperti transform (e.g. Iacus, 2008; Baadsgaard et al., 1997)

$$\psi(X_t) = \int \frac{1}{\sigma(\xi)} d\xi|_{\xi=X_t}, \tag{26}$$

by choosing $Z_t = \psi(X_{p,t}) = \int 1/(\sigma_x \xi) d\xi|_{\xi=X_t} = (1/\sigma_x)\log(X_{p,t})$, and applying Ito's lemma (e.g. Øksendal, 2003) we get

$$dZ_{p,t} = \psi_t(X_{p,t})dt + \psi_x(X_{p,t})dX_{p,t} + \psi_{xx}(X_{p,t})(dX_{p,t})^2, \tag{27}$$

$$= \frac{(b_0 U_{w,t} U_{gr,t} - aX_{p,t})dt + \sigma_x X_{p,t} dw_t}{\sigma_x X_{p,t}} - \frac{1}{2}\frac{(\sigma_x X_{p,t})^2}{\sigma_x (X_{p,t})^2} dt, \tag{28}$$

$$= \left(\frac{b_0 U_{w,t} U_{gr,t}}{\sigma_x X_{p,t}} - \frac{a}{\sigma_x} - \frac{1}{2}\sigma_x\right) dt + dw_t, \tag{29}$$

$$= \left(\frac{b_0}{\sigma_x} e^{-\sigma_x Z_{p,t}} U_{w,t} U_{gr,t} - \frac{a}{\sigma_x} - \frac{1}{2}\sigma_x\right) dt + dw_t, \tag{30}$$

which is now a non-linear SDE with unit diffusion. In addition to the system equation described above, there has to be a description of the observation equation. Under the assumption that observations are log-normal distributed around the true state, the observation equation is

$$\log(Y_{p,k}) = \sigma_x Z_{p,t_k} + e_k, \tag{31}$$

where $Y_{p,k}$ is the observed nitrogen content in phytoplankton and $e_k \sim N(0, \sigma_y^2)$.

4.3. Results

The parameters of the model Eqs. (30) and (31) are now estimated using the estimation procedure presented in Section 2.3. All parameters score well in t-tests (Table 1). The time constant $(1/a)$ of the deterministic skeleton (remove the noise term) is about 59 days. As discussed above, the stationary distribution is known when the forcing $(b_0 U_{w,t} U_{gr,t})$ is constant, which is clearly (Fig. 2) not the case for the system analysed here. Even though we do not analyse the dynamics of the forcing compared to the time constant of the system, it is reasonable to assume that the state (total phytoplankton nitrogen) is close to the stationary distribution in some sense. To explore this we define a moving average growth process

$$\tilde{b}_t = \frac{b_0}{14} \sum_{i=t-13}^{t} U_{w,i} U_{gr,i}. \tag{32}$$

This moving average growth process is now used (in Eq. (25)) to calculate confidence intervals around the mode of the stationary distribution of each t (Figs. 3 and 4).

In addition to parameter estimates, the implementation of the EKF allows us to calculate the smoothed state (the conditional

Table 1
Estimation results.

	$[\theta_{\min}, \theta_{\max}]$	$\hat{\theta}$	Std. dev.	t-Score	$P(x > \lvert t \rvert)$
$Z_{p,0}$	[−20, 20]	−9.4e+00	3.1e+00	−3.0	0.003
b_0	[0, 1]	1.9e−03	2.2e−04	8.4	0.000
a	[0, 1]	1.7e−02	3.4e−03	5.0	0.000
σ_x	[0, 1]	1.6e−01	1.3e−02	12.0	0.000
σ_y	[0, 1]	1.9e−01	2.6e−02	7.6	0.000

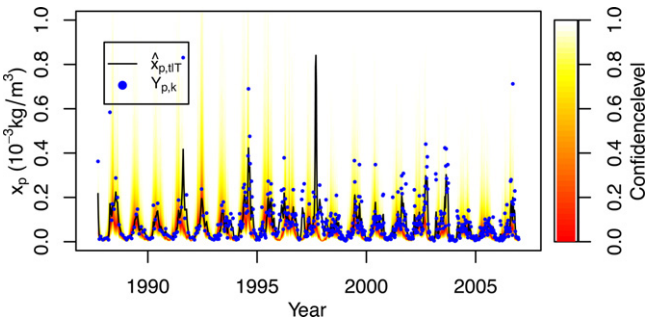

Fig. 3. Time series of stationary distributions, smoothed state (black line) and observations (blue dots). The color key refers to confidence intervals around the mode of the stationary distributions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

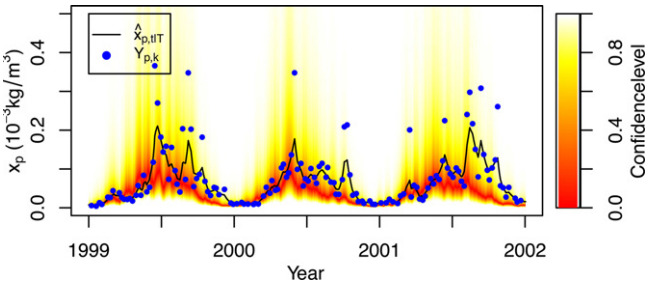

Fig. 4. Time series of stationary distributions, smoothed state (black line) and observations (blue dots). The color key refers to confidence intervals around the mode of the stationary distributions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

expectation of the state given all observations), i.e.

$$\hat{\boldsymbol{x}}_{t|T} = E[\boldsymbol{x}_t|\mathcal{Y}_N]. \tag{33}$$

When considering that the stationary distribution does not depend on the local information given by the observations the smoothen state is in general quite close to the mode of the stationary distribution (Figs. 3 and 4). Both the stationary distribution and the smoothen state reproduce the periodic dynamics of data quite well, while the extreme values are not well captured. In particular, the stationary distribution fails to reproduce extreme values (Figs. 3 and 4), because it does not use local information given by the observations.

Very large observations influence the smoothed state of the system strongly (Fig. 3), implying that observations in the tail of the stationary distribution draw the smoothed state away from the centre of the stationary distribution. This indicates that the model does not capture the extremes in the dynamics very well, which is not surprising since the model does not include any mechanism to capture extreme events. It might possible to capture such behaviour by more effects (e.g. $b(t) = b(U_{w,t}, U_{gr,t}, X_{p,t})$) in the system equations.

5. Conclusion

We have demonstrated that the presented approach based on embedded stochastic differential equations provides an alternative tool for phytoplankton modelling. In particular the procedure (as illustrated in Section 3) accounts for the autocorrelated residuals often seen when ODEs are used for modelling. Furthermore, as the model is formulated in continuous time, the states can be updated and parameters estimated from data that are not sampled at equidistant points in time, which often happens to be the case with ecosystem monitoring. This is exemplified with the Skive Fjord case study presented in Section 4. The higher flexibility of the estimating procedure, compared to discrete-time models, is a trade-off with the computational effort. The ODEs given by the filter equations are computationally expensive when the system equations are complex. Further, the optimisation requires many iterations when the number of parameters to be estimated is high. In practice, this limits the complexity of ecosystem models formulated as SDEs that can be estimated, acknowledging that the more complex and less significant mechanisms will be contained in the stochastic processes of the covariance model, that will be regularly updated through the filter equations. Consequently, SDEs are per se data-driven and less appropriate for long-term predictions or interpolation over larger gaps in the time series compared to ODEs. However, an important feature of SDEs is the uncertainty quantification of the model outputs, such uncertainty quantification cannot be readily and reliably provided by ODEs.

The skive Fjord case study provide two qualitatively different results 1) the stationary distribution, which represents long-term predictions under given loading conditions and 2) the smoothen state which represents the conditional mean of the phytoplankton state given all observations (both past and future), the model structure and the parameters. The stationary distribution reproduces the long-term dynamics of the data quite well, while local information from the observation does not influence the predictions and extreme observations are quite far from the mode of the distribution. The smoothen state clearly describes data better than the stationary distribution as it is adapted to the local information provided by the observed phytoplankton. The smoothen state can, however, not describe extreme observations.

The problem of reproducing extreme observations could potentially be solved by including more effects (e.g. the phytoplankton state and local weather conditions) in the growth process. Moreover, for the model to constitute a realistic representation with desired mathematical properties (stationarity of the solution), a two-state system, where phytoplankton remove nitrogen from the water column, is more appropriate. However, the aim of this study was to introduce SDEs and the estimation procedure, and not a modelling exercise.

References

Aït-Sahalia, Y., 2008. Closed-form likelihood expansions for multivariate diffusions. The Annals of Statistics 32 (2), 906–937.
Baadsgaard, M., Nielsen, J.N., Spliid, H., Madsen, H., Preisel, M., 1997. Estimation in stochastic differential equations with state dependent diffusion term. In: SYSID '97 – 11th IFAC Symposium on System Identification, IFAC.
Bartell, S.M., Lefebvre, G., Kaminski, G., Carreau, M., Campbell, K.R., 1999. An ecosystem model for assessing ecological risks in Québec river, lakes and reservoirs. Ecological Modelling 124, 43–67.
Brillinger, D.R., Preisler, H.K., Ager, A.A., Kie, J.G., Stewart, B.S., 2002. Employing stochastic differential equation to model wildlife motion. Bulletin Brazilian Mathematical Society, New Series 33 (3), 385–408.
Carpenter, S.R., Brock, W.A., 2006. Rising variance: a leading indicator of ecological transition. Ecology Letters 9, 311–318.
Dowd, M., 2006. A sequential Monte Carlo approach for marine ecological prediction. Environmetrics 17, 435–455.
Dowd, M., 2007. Bayesian statistical data assimilation for ecosystem models using Markov Chain Monte Carlo. Journal of Marine Systems 68, 439–456.
Fasham, M.J.R., Ducklow, H.W., McKelvie, S.M., 1990. A nitrogen-based model of plankton dynamics in the oceanic mixed layer. Journal of Marine Research 58, 591–639.
Givon, D., Stinis, P., Weare, J., 2009. Variance reduction for particle filters of systems with time scale separation. IEEE Transactions on Signal Processing 57 (2), 424–435.
Guttal, V., Jayaprakash, C., 2008. Changing skewness: an early warning signal of regime shift in ecosystems. Ecology Letters 201 (3), 420–428.
Iacus, S.M., 2008. Simulation and Inference for Stochastic Differential Equations – with R Examples. Springer Series in Statistics.
Jazwinski, A.H., 1970. Stochastic Processes and Filtering Theory. Academic Press, New York, USA.
Jacobsen, J.L., Madsen, H., 1996. Grey box modelling of oxygen levels in a small stream. Environmetrics 7, 109–121.
Klebaner, F.C., 2005. Introduction to Stochastic Calculus with Applications. Imperial College Press.
Kloden, P., Platen, E., 1999. Numerical Solutions of Stochastic Differential Equations. Springer-Verlag, Berlin.

Kristensen, N.R., Madsen, H., Jørgensen, S.B., 2004. Parameter estimation in stochastic grey-box models. Automatica 40, 225–237.
Kristensen, N.R., Madsen, H., 2003. Continuous Time Stochastic Modeling – CTSM 2.3 – Mathematics Guide. Technical University of Denmark.
Limpert, E., Stahel, W.A., Abbt, M., 2001. Log-normal distributions across the sciences: keys and clues. BioScience 51 (5), 341–352.
Luschgy, H., Pagés, G., 2006. Functional quantization of a class of Brownian diffusions: a constructive approach. Stochastic Processes and their Applications 116, 310–336.
Madsen, H., Holst, J., Thyregod, P., 1987. A continuous time model for the variations of air temperature. In: 10th Conference on Probability and Statistics in Atmospheric Science, American Meteorological Society, Edmonton, pp. 52–58.
Madsen, H., Holst, J., 1995. Estimation of continuous-time models for the heat dynamics of a building. Energy and Building 22, 67–79.
Madsen, H., 2008. Time Series Analysis. Chapman & Hall/CRC.
Matear, R., 1995. Parameter optimization and analysis of ecosystem models using simulated annealing: a case study at Station P. Journal of Marine Research 53, 571–607.
Nicolau, J., 2002. A new technique for simulating the likelihood of stochastic differential equations. Econometrics 5, 91–103.
Pastorello, S., Rossi, E., 2010. Efficient importance sampling maximum likelihood estimation of stochastic differential equation. Computational Statistics and Data Analysis 54, 2753–2762.
Pedersen, M.W., Righton, D., Thygesen, U.H., Andersen, K., Madsen, H., 2008. Geolocation of North Sea cod using Hidden Markov Models and behavioural switching. Canadian Journal of Fisheries and Aquatic Sciences 65, 2367–2377.
Pedersen, T.M., Almeda, R., Fotel, F.L., Jacobsen, H.H., Mariani, P., Hansen, B.W., 2010. Larval growth in the dominant polychaete Polydora ciliata is food-limited in a eutrophic Danish estuary (Isefjord). Marine Ecology Progress Series 407, 99–110.
Stollenwerk, N., Friedhelm, D.R., Siegel, H., 2001. Testing nonlinear stochastic models on phytoplankton biomass time series. Ecological Modelling 144, 1261–1277.
Tornøe, C.W., Jacobsen, J., Pedersen, O., Hansen, T., Madsen, H., 2004. Grey-box modelling of pharmacokinetic/pharmacodynamic systems. Journal of Pharmacokinetics and Pharmacodynamics 31, 401–417.
Øksendal, B., 2003. Stochastic Differential Equations – An Introduction with Applications, sixth edition. Springer.