Available online at www.sciencedirect.com

ScienceDirect

Journal of Marine Systems 68 (2007) 439–456

JOURNAL OF MARINE SYSTEMS

www.elsevier.com/locate/jmarsys

Bayesian statistical data assimilation for ecosystem models using Markov Chain Monte Carlo

Michael Dowd [*]

Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia, Canada B3H 3J5

Received 17 August 2006; received in revised form 13 January 2007; accepted 19 January 2007
Available online 8 February 2007

Abstract

This study considers advanced statistical approaches for sequential data assimilation. These are explored in the context of nowcasting and forecasting using nonlinear differential equation based marine ecosystem models assimilating sparse and noisy non-Gaussian multivariate observations. The statistical framework uses a state space model with the goal of estimating the time evolving probability distribution of the ecosystem state. Assimilation of observations relies on stochastic dynamic prediction and Bayesian principles. In this study, a new sequential data assimilation approach is introduced based on Markov Chain Monte Carlo (MCMC). The ecosystem state is represented by an ensemble, or sample, from which distributional properties, or summary statistical measures, can be derived. The Metropolis-Hastings based MCMC approach is compared and contrasted with two other sequential data assimilation approaches: sequential importance resampling, and the (approximate) ensemble Kalman filter (including computational comparisons). A simple illustrative application is provided based on a 0-D nonlinear plankton ecosystem model with multivariate non-Gaussian observations of the ecosystem state from a coastal ocean observatory. The MCMC approach is shown to be straightforward to implement and to effectively characterize the non-Gaussian ecosystem state in both nowcast and forecast experiments. Results are reported which illustrate how non-Gaussian information originates, and how it can be used to characterize ecosystem properties.

Keywords: Marine ecosystem models; Data assimilation; State space models; Monte Carlo; Prediction; MCMC; Ensemble Kalman filter; Sequential importance resampling; Ecological statistics; Stochastic models; Particle filters; Non-Gaussian; Differential equations

1. Introduction

Data assimilation is fundamentally a problem in statistical estimation, *i.e.* combining dynamical models and data to provide state or parameter estimates. Marine ecosystem models for lower trophic levels (biogeochemical and plankton models) typically take the form of time-dependent nonlinear differential equations (Fennell and Neumann, 2004), and are known to exhibit a wealth of complex dynamical behaviour (Huisman and Weissing, 2001; Edwards, 2001). These ecosystem models are generally treated as deterministic, and frequently coupled, as interacting tracers, to physical oceanographic models to allow for transport and mixing (Oschlies and Schartau, 2005). Stochastic elements enter ecosystem models through environmental forcing such as mixed layer dynamics or rapid fluctuations in the light environment (Marion et al., 2000; Edwards et al., 2004; Dowd, 2006). Observations of marine ecological state variables are complex data types coming from a variety of sources and sensors (*e.g.* satellites, water samples, moored and profiling instruments), and characterized by being sparse, noisy and non-Gaussian in their distributions (Dickey,

[*] Tel.: +1 902 494 1048; fax: +1 902 494 5130.
E-mail address: Michael.Dowd@Dal.Ca.

0924-7963/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jmarsys.2007.01.007

2003). A major challenge for marine prediction is the identification and development of appropriate data assimilation methods to integrate this variety of data sources with marine ecosystem models.

Oceanographic data assimilation has generally been divided into two approaches: (i) variational methods for estimation of parameters (and initial conditions), and (ii) sequential methods for state estimation. Parameter estimation is concerned with model calibration: models are considered as deterministic functions of the parameters, a cost function is then posed which measures the discrepancy between the model and data, and optimization procedures are used to minimize it (Lawson et al., 1995; Vallino, 2000; Evans, 2003; Oschlies and Schartau, 2005). Sequential methods, on the other hand, are recursive algorithms concerned with estimation of the ecological state as the system evolves through time, in other words nowcasting and forecasting (Bertino et al., 2003). After model calibration, these sequential approaches provide the basis for biological forecasting systems that are emerging as part of ocean observing systems (Allen et al., 2003; Pinardi et al., 2003). This study is concerned with the identification and application of advanced statistical methods for sequential data assimilation for nonlinear and non-Gaussian interdisciplinary oceanographic studies.

The integration of modern statistical approaches into oceanographic data assimilation is in its infancy. Methods such as Markov Chain Monte Carlo (MCMC) have revolutionalized Bayesian statistical computation (Gelman et al., 2003). Indeed, the data assimilation problem has long been formulated from a probabilistic perspective using Bayesian principles (Jazwinski, 1970; van Leeuwen and Evensen, 1996). However, for sequential data assimilation, the main approach has been to formulate approximate methods based on extensions of the Kalman filter to treat nonlinear systems. For example, Pham et al. (1997) introduced the singular evolutive extended Kalman (SEEK) filter based on an EOF approximation of the updating equations. A popular Monte Carlo approach is the ensemble Kalman filter, or EnKF (Tsuji and Nakamura, 1973; Evensen, 1994). This uses stochastic dynamic (Monte Carlo) prediction, but approximates the Bayesian assimilation of observations with the Kalman filter updating equations. It has been widely used in oceanography (Evensen, 2003), including applications to marine ecological data assimilation (Eknes and Evensen, 2002; Allen et al., 2003; Natvik and Evensen, 2003). An exact statistical approach for sequential data assimilation in nonlinear and non-Gaussian systems is sequential importance resampling, or SIR (Gordon et al., 1993; Kitagawa, 1996). SIR has become well established in the statistical and signal processing literature (see texts by Doucet et al. (2001) and Ristic et al. (2004)), with a few pilot applications in data assimilation in physical oceanography (van Leeuwen, 2003) and biogeochemical modelling (Losa et al., 2003).

In this study, an alternative statistical approach for sequential data assimilation is introduced based on an MCMC approach. Like SIR, but unlike the EnKF, the proposed MCMC method provides an exact solution for general nonlinear and non-Gaussian data assimilation. The well-known drawback of the SIR algorithm is that the ensemble that represents the system state can degenerate due to the repeated resampling (or bootstrapping) steps that are fundamental to its operation (*e.g.* Arulampalam et al., 2002). In recent work (Dowd, 2006), a modification to the SIR algorithm was proposed to alleviate this problem of sample degeneracy by appending an MCMC post-processing step to the SIR algorithm. However, it was subsequently realized that the MCMC ideas developed in Dowd (2006) could be adapted to stand-alone as a useful and novel approach for sequential data assimilation. This study explores this idea and develops and tests the methodology, with an emphasis on estimation of the non-Gaussian features of the ecological state.

The paper is structured as follows. Section 2 introduces the statistical framework for treating the nowcasting and forecasting problems of sequential data assimilation. It also outlines the MCMC approach, and contrasts it to the SIR and EnKF algorithms. Section 3 provides an illustration of the method with a simple application to a 0-D nonlinear plankton model using non-Gaussian multivariate observations from a coastal ocean observing system. It is shown how the non-normal distributional information for nowcasts and forecasts can be used to diagnose and describe properties of the ecosystem state variables. The computational performance of the candidate data assimilation algorithms are also compared and contrasted. A discussion and conclusions are given in Section 4.

2. Methods

2.1. Problem statement

The statistical framework for ecological data assimilation is provided by the nonlinear and non-Gaussian state space model (*e.g.* Dowd and Meyer, 2003),

$$x_t = f_t(x_{t-1}, \theta_t, n_t) \tag{1}$$

$$y_t = h_t(x_t, \phi_t, v_t) \tag{2}$$

defined for t=1,... T. The first equation (Eq. (1)) is a stochastic difference equation representing a Markovian transition over a unit time interval (from time $t-1$ to time t). It is identified with the numerical model for the ecological dynamics. The ecosystem state variables at time t are contained in the vector x_t, which also includes any spatial dimension (*e.g.* gridded fields being mapped into vectors). The state evolution, or ecosystem dynamics, is embodied in the operator f_t, which may be time dependent. The model depends on a set of parameters θ_t which includes static rate constants, as well as time-dependent parameters and forcing (including boundary fluxes). The system noise n_t includes stochastic elements due to environmental forcing; in the estimation context it can also be considered to incorporate model errors due to structural uncertainty in the governing equations.

The second equation (Eq. (2)) is the measurement equation which incorporates observations on all or part of the ecological state. The observation vector at time t is given by y_t, but may not be defined for all times (*i.e.* missing values). Observations, y_t, are related to the ecosystem state, x_t, through the measurement operator h_t. This depends on a parameter set, θ_t, and the measurement errors v_t. This allows for indirect and nonlinear relations between the observations and the state (*e.g.* measuring optical properties and modelling phytoplankton). Note that the special case of direct measurements of the complete ecosystem state implies that h_t is the identity matrix.

Designate the available observation set from times 1 through T inclusive as $y_{1:T}=(y_1', y_2',\ldots, y_T')'$ (that is, we stack the observation vectors). Suppose we are interested in state estimates at some analysis time τ (*i.e.* for x_τ). Three classes of time-dependent estimation problems can be defined corresponding to hindcasting ($\tau<T$), nowcasting ($\tau=T$) and forecasting ($\tau>T$). Statistically, these correspond to the problems of smoothing, filtering and prediction. In this study we focus on the filtering (nowcasting) and prediction (forecasting) problems using online recursive approaches for sequential state estimation.

A complete description of the ecosystem state at any time is given by the joint probability density function (pdf) of x_t defined for $t=1,\ldots, T$. These pdfs embody all information on the ecological state and are commonly summarized in terms of measures of the central tendency (the mean or mode), uncertainty (variance), and higher order moments such as skewness and kurtosis. Relationships between variables can also be summarized by measures of dependence, such as covariance. The goal of this study is to estimate the time evolving probability distribution of the ecological state using both ecosystem models and measurement information.

The target quantity to be computed by the sequential data assimilation procedure is $p(x_t|y_{1:t})$, which is the conditional pdf for the ecological state given all the information (observations) up to and including time t (where $t=1,\ldots, T$). This is referred to as the filter, or nowcast, density at time t. Note that while the conditioning in the pdf only explicitly considers the observations, it is implicit that the following also be specified: the model equations, f_t, along with the parameters (forcing), θ_t; the measurement operator, h_t, with its parameters, ϕ_t; and the statistics of both the system noise, n_t, and observation error, v_t. Initial conditions, x_0, must also be provided.

Sequential estimation of the ecosystem state can be conceptualized in the data assimilation cycle in Fig. 1. It starts with an estimate of the ecological state at the current time (*i.e.* a nowcast at time $t-1$). This is designated by $p(x_{t-1}|y_{1:t-1})$ which is the pdf of x_{t-1} using measurements up to and including time $t-1$. Two steps are then taken. First, a prediction is made for time t using $p(x_{t-1}|y_{1:t-1})$ as the initial conditions for the (stochastic) ecological model dynamics in Eq. (1), including any external forcing or boundary conditions. This yields the predictive density $p(x_t|y_{1:t-1})$. Second, suppose an observation y_t becomes available at time t. This measurement information is blended with the accumulated information on x_t contained in the forecast. This yields the nowcast at time t as $p(x_t|y_{1:t})$. For sequential data assimilation, a recursive sequence of these single stage transitions of the system is applied to the entire analysis period of interest.

2.2. Analytic solution

The rules for manipulating the pdfs in order to carry out the prediction and measurement steps of the data

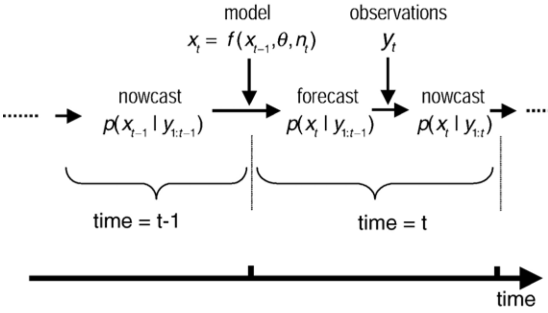

Fig. 1. The data assimilation cycle showing a single stage transition of the ecological system. It starts with a nowcast distribution of the state at time $t-1$, or $p(x_{t-1}|y_{1:t-1})$. This is used as the initial condition for a forecast to time t using model dynamics to yield the $p(x_t|y_{1:t-1})$. Finally, this forecast is combined with the observation y_t to produce the desired nowcast at time t, or $p(x_t|y_{1:t})$. See the text for further details.

assimilation cycle are given by the following (c.f. Jazwinski, 1970). The prediction step is

$$p(x_t|y_{1:t-1}) = \int p(x_t|x_{t-1})p(x_{t-1}|y_{1:t-1})\mathrm{d}x_{t-1}. \tag{3}$$

The predictive distribution $p(x_t|y_{1:t-1})$ is computed as a product of the nowcast distribution, $p(x_{t-1}|y_{1:t-1})$, and a transition density $p(x_t|x_{t-1})$. This latter quantity moves the state forward one time step and is identified with the model dynamics in Eq. (1). The measurement update step occurs at time t when observational information y_t becomes available. The updating relies on straightforward application of Bayes' formula, *i.e.*

$$p(x_t|y_{1:t}) = \frac{p(y_t|x_t)p(x_t|y_{1:t-1})}{p(y_{1:t})} \tag{4}$$

and yields the desired nowcast density at time t. Note that the joint pdf of the observations $p(y_{1:t})$ in the denominator acts simply as a normalizing constant and modern computational Monte Carlo techniques (as below) do not require its evaluation. The nowcast density at time t, $p(x_t|y_{1:t})$, (the posterior) is expressed as the product of the likelihood $p(y_t|x_t)$ and the forecast density at time t, $p(x_t|y_{1:t-1})$, (the prior). The likelihood can be evaluated based on the measurement Eq. (2).

These two steps can be combined into a single expression for the single stage transition of the data assimilation cycle by substituting Eq. (3) into Eq. (4) to yield

$$p(x_t|y_{1:t}) \propto p(y_t|x_t) \int p(x_t|x_{t-1})p(x_{t-1}|y_{1:t-1})\mathrm{d}x_{t-1}. \tag{5}$$

where the proportionality means that the denominator can be dropped. This equation provides the basis for the Markov Chain Monte Carlo based sequential data assimilation procedure introduced below.

2.3. Numerical solutions

2.3.1. Sampling based solutions

Numerical solutions for sequential data assimilation for the general nonlinear and non-Gaussian case relies on Monte Carlo methods (Kitagawa, 1987). These are based on algorithms that generate samples or ensembles from the desired nowcast (and forecast) distributions. These ensembles can then be used to reconstruct approximations to the distributions of interest, or any summary quantities desired (*e.g.* the mean and variance). Consider a set of n realizations of the ecosystem state vector, each of which is denoted by $x^{(i)}$ where $i=1,\ldots, n$. Suppose this sample is drawn from the (multivariate) probability distribution $p(x)$, as designated by the notation

$$\{x^{(i)}\} \sim p(x), \qquad i = 1, \ldots, n.$$

Here, superscript i is an index which refers to the ith member of the ensemble, and the curly braces refer to the entire set, or the ensemble itself. Hence, $\{x^{(i)}\}$ denotes the sample from $p(x)$ which has n elements. As the ensemble size $n \rightarrow \infty$ it provides for an exact representation of the pdf of the random variable x.

Fig. 2 shows a illustration of such an ensemble based representation of a pdf. Here, random samples of size $n=25$ and $n=250$ have been drawn from the standard normal distribution (the $N(0,1)$) using a random number generator. The positions of these ensemble members (or particles) are shown on the graph. Given this ensemble, estimates for the underlying distributional parameters (the sample mean and sample variance) can be determined and are shown on the plots. Clearly the larger

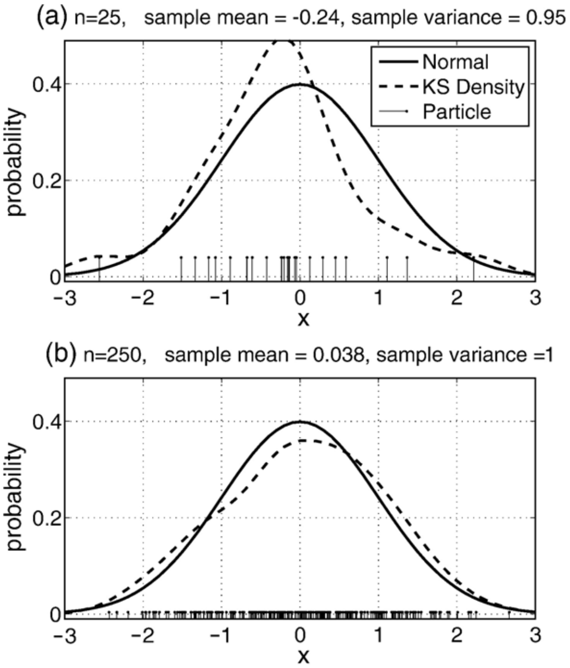

Fig. 2. Ensemble based representation of a probability distribution. The x-axis is the value of the state variable and the y-axis denotes probability (probability density for the distributions, and probability mass for the particles). The standard normal with mean zero and variance one is shown (solid line). The ensemble members or particles (short vertical lines) are shown in terms of their values (x-axis position) and their probability (height above $x=0$). The estimated kernel smoothed density determined from the ensemble is also given (dashed line). Panel (a) shows the case of $n=25$ ensemble size, while panel (b) shows the $n=250$ case.

ensemble gives answers that are closer to the true mean and variance. The distribution itself can also be reconstructed from the smoothed and scaled histogram, or through kernel density estimation (Silverman, 1986). It is seen in Fig. 2 that the larger ensemble yields an estimated pdf much closer to the standard normal. This simple idea of using samples to characterize distributions is the basis for the Monte Carlo based data assimilation methods outlined below.

This same sample based characterization of the ecosystem state can be applied to the single stage transition of the data assimilation cycle as follows. The starting point is an ensemble of n ecological states given as

$$\{x_{t-1|t-1}^{(i)}\} \sim p(x_{t-1}|y_{1:t-1}), \qquad i = 1, \ldots, n \tag{6}$$

that represents the nowcast pdf at time $t-1$ (it is shown how this is generated recursively below; for now assume that it is given). The subscripting on x designates a sample at time $t-1$ using information (data) up to and including time $t-1$. The target distribution of the sequential estimation procedure is the nowcast ensemble at time t represented as

$$\{x_{t|t}^{(i)}\} \sim p(x_t|y_{1:t}), \quad i = 1, \ldots, n. \tag{7}$$

Monte Carlo methods can be used to carry out the transition from Eq. (6) to Eq. (7) for sequential data assimilation. Sequential importance resampling (Gordon et al., 1993; Kitagawa, 1996) offers an exact solution, while the ensemble Kalman filter (Tsuji and Nakamura, 1973; Evensen, 1994) provides an approximate one. Markov Chain Monte Carlo methods are another possible approach for exact solutions to the problem of sequential data assimilation. However, these have not been widely applied except as part of more comprehensive SIR algorithms (Gilks and Berzuini, 2001; Lee and Chia, 2002; Dowd, 2006). MCMC methods have not, to the author's knowledge, been applied for sequential data assimilation for marine systems. In the next section, a flexible and general Metropolis-Hasting based MCMC algorithm is proposed for sequential data assimilation.

2.3.2. MCMC for sequential data assimilation

Markov Chain Monte Carlo (MCMC) methods are numerical methods which evaluate Bayes' formula to generate ensembles drawn from a desired posterior (or target) distribution (*e.g.* Gamerman, 1997; Gelman et al., 2003). For Bayesian sequential data assimilation, Eq. (5) indicates that the desired target distribution is $p(x_t|y_{1:t})$, or the nowcast at time t. To obtain the posterior *via* generation of an ensemble $\{x_{t|t}^{(i)}\}$, we evaluate Eq. (5) using a Metropolis-Hasting MCMC technique. The Metropolis-Hastings algorithm is the basis for a general and flexible class of MCMC methods (Metropolis et al., 1953; Hastings, 1970), and below it is tailored to treat the problem of sequential data assimilation.

The iterative procedure used to generate the target ensemble $\{x_{t|t}^{(i)}\}$ is as follows. Suppose we are at the $i-1$th iteration of the Metropolis-Hastings algorithm and so the associated ensemble member, $x_{t|t}^{(i-1)}$, is available. To generate the next member in the target ensemble, $x_{t|t}^{(i)}$, the following steps are taken:

1. Generate a candidate $x_{t|t-1}^{*}$ as a sample from the forecast density $p(x_t|y_{1:t-1})$. To do this, first randomly choose one member of the nowcast ensemble at the previous time $\{x_{t-1|t-1}^{(i)}\}$, which is designated as $x_{t-1|t-1}^{*}$. This is then used as an initial condition for a forecast of the stochastic dynamic model of Eq. (1), *i.e.*

$$x_{t|t-1}^{*} = f(x_{t-1|t-1}^{*}, \theta_t, n_t^{*}), \tag{8}$$

where n_t^{*} represents an independent realization of the system noise.

2. Calculate the probability of accepting the candidate $x_{t|t-1}^{*}$ as the ith ensemble member of the target distribution, *i.e.* as $x_{t|t}^{(i)}$. This probability is computed as

$$\alpha = \min\left(1, \frac{p(y_t|x_{t|t-1}^{*})}{p(y_t|x_{t|t}^{(i-1)})}\right). \tag{9}$$

3. Accept the candidate as the ith ensemble member with probability α. This is carried out according to the following rule. Draw z from Uniform(0,1) distribution. Then

$$x_{t|t}^{(i)} = \begin{cases} x_{t|t-1}^{*} & \text{if} \quad z \le \alpha \\ x_{t|t}^{(i-1)} & \text{if} \quad z > \alpha \end{cases}$$

Therefore given a starting value $x_{t|t}^{(1)}$, the algorithm can be run n times (or indeed any number of times) to generate the required ensemble $\{x_{t|t}^{(i)}\}$, $i=1,\ldots, n$. This provides a draw from the target nowcast distribution at time t, $p(x_t|y_{1:t})$. Pseudo-code for this algorithm is given in Appendix A.

The MCMC algorithm offers significant advantages for sequential data assimilation. First, it is easy to implement. Appendix A shows that to sequentially generate candidates, the ecosystem model is called as a subroutine (called *ModelForecast*) for one step ahead stochastic dynamic prediction of a single ensemble member. The candidates are then included in the target ensemble using a simple accept/reject rule. The acceptance probability (9) takes the form of a simple ratio of likelihoods since here the candidates are drawn from the prior, $p(x_t|y_{1:t-1})$ (Chib and Greenberg, 1995). The second advantage is that the sequence generated is an independence chain (Tierney, 1994). Candidates are drawn independently from the predictive density and so have no dependence on the current state of the algorithm. Dependence only arises due to the fact that the chain does not necessarily move (or accept the new candidate) every iteration so that ensemble members can be repeated. This makes the acceptance probability α is an important diagnostic for algorithm performance. As a consequence of this independence, some well known issues with MCMC are alleviated: the sequence effectively mixes to rapidly explore the state space; and the "burn-in" time is negligible since once the first candidate is accepted, the chain is generating draws from the posterior.

2.3.3. Other candidate approaches

Here, two other statistical approaches for sequential data assimilation are briefly reviewed and contrasted to the MCMC approach (computational comparisons are given in Section 3). The first method is sequential importance resampling (SIR), which, like MCMC, produces exact solutions for sequential data assimilation. The interested reader is referred to Ristic et al. (2004) for details of the algorithm; summaries in a marine ecological context are given in Losa et al. (2003) and Dowd (2006). The second method considered, the ensemble Kalman filter (EnKF), provides an approximate solution for sequential data assimilation; implementation details are given in Evensen (2003).

2.3.3.1. Sequential importance resampling. SIR involves separate evaluations of both the prediction step (3) and the measurement step (4). The starting point is again the ensemble $\{x^{(i)}_{t-1|t-1}\}$ drawn from the nowcast density at time $t-1$, $p(x_{t-1}|y_{1:t-1})$. Prediction can be based on an ensemble forecast, *i.e.* moving each of the n ensemble members forward one time step using the stochastic dynamical model (1), *i.e.*

$$x^{(i)}_{t|t-1} = f(x^{(i)}_{t-1|t-1}, \theta_t, n^{(i)}_t). \qquad i = 1, \ldots, n. \tag{10}$$

However, note that other proposal densities (other than the prior or predictive density) are possible. In the above, $n^{(i)}_t$ represents an independent realization of the system noise for the ith ensemble member. This yields a new ensemble $\{x^{(i)}_{t|t-1}\}$ which is a draw from the predictive density $p(x_t|y_{1:t-1})$. The measurement step (4) then starts with the ensemble $\{x^{(i)}_{t|t-1}\}$. Each of the n ensemble members is assigned a weight $w^{(i)}$ according to the likelihood

$$w^{(i)} = p(y_t|x^{(i)}_{t-1|t}), \qquad i = 1, \ldots n. \tag{11}$$

That is, ensemble members that are close to observations will be given a higher weight, while those that are more distant will be given a smaller weight. The probability model used to evaluate the likelihood follows the measurement distribution in Eq. (2). To generate the required (target) ensemble $\{x^{(i)}_{t|t}\}$, the ensemble $\{x^{(i)}_{t|t-1}\}$ is resampled (with replacement) wherein members are chosen with a probability proportional to their weight.

This weighted resampling procedure is at the core of the SIR methods. Its main drawback is sample impoverishment wherein ensemble members with high weights are chosen more frequently and the resultant target sample $\{x^{(i)}_{t|t}\}$ may contain many repeats. Much research effort has been aimed at developing modified SIR schemes which alleviate these problems (Gilks and Berzuini, 2001; Arulampalam et al., 2002; Dowd, 2006).

2.3.3.2. Ensemble Kalman filter. The ensemble Kalman filter also evaluates the prediction and measurement steps separately. As with SIR, prediction relies on ensemble forecasts *via* Eq. (10). However, the measurement step is simplified by using the Kalman filter updating equations which follow from the linear, Gaussian version of the state space model (1)–(2).

To illustrate the procedure, consider a version of Eq. (2) describing a linear observation equation defined by a matrix H and having additive errors v_t. Suppose the forecast ensemble $\{x^{(i)}_{t|t-1}\}$ is available. The ith ensemble member of the target nowcast at time t is determined as

$$\tilde{x}^{(i)}_{t|t} = x^{(i)}_{t|t-1} + K(y^{(i)}_t - Hx^{(i)}_{t|t-1}), \qquad i = 1, \ldots, n \tag{12}$$

and the resultant ensemble $\{\tilde{x}^{(i)}_{t|t}\}$ represents the nowcast at time t. The tilde notation is used to emphasize that the ensemble will not in general be a draw from $p(x_t|y_{1:t})$, but only an approximation. As part of the updating procedure (12), a measurement ensemble $\{y^{(i)}_t\}$ has been introduced and is generated as $y^{(i)}_t = y_t + v^{(i)}_t$ for $i = 1, \ldots, n$ where $v^{(i)}_t$ is an independent realization of the observation error. The Kalman gain matrix K has also been used and is defined as

$$K = PH'(HPH' + R)^{-1} \tag{13}$$

with P and R being the sample covariance matrices of the forecast ensemble $\{x^{(i)}_{t|t-1}\}$ and the observation ensemble $\{y^{(i)}_t\}$, respectively.

The ensemble Kalman filter therefore does not evaluate directly Bayes formula in Eq. (4). Rather it approximates the result by using the Kalman gain to update each forecast ensemble member by taking its old value and adding to it an increment based on the discrepancy between the observation and the forecast. For some cases, the EnKF can be transformed to an exact procedure (Bertino et al., 2003; Evensen, 2003). However, general observation error processes as in (2) are not supported directly.

3. Application

3.1. Observations

Measurements of the ecosystem state are taken from an observing system in Lunenburg Bay, Canada (44.36°N, 64.26°W). Lunenburg Bay is a small (8 km long), shallow (max depth 20 m) tidal embayment. Observations for both phytoplankton and nutrients are available. Phytoplankton observations were derived from optical time series at three locations in the bay using the algorithm of Huot et al. (submitted for publication). Observations on inorganic nitrogen included both nitrates and ammonia and were based on biweekly water sampling at 5 stations in the bay. Units for these ecological state variables were expressed in μmol nitrogen l^{-1}.

Observations for phytoplankton, P, and nutrients, N, from Lunenburg Bay are shown in Fig. 3. Since little coherent spatial variation was evident (in either the vertical or horizontal) in these variables, their daily values have been spatially pooled and are reported as time series. Phytoplankton observations have some gaps but are generally available on a regular daily basis (since they are derived from ocean optical data). The median value cycles near 1 μmol N l^{-1} until after day 250 when it rises to near two. Significant high frequency variability is seen. In contrast, the *in situ* nutrient observations are much more sparse and exhibit significant sampling variability with few clear trends evident.

The multiple observations on P and N for any given day were treated as replicates and used to identify the appropriate probability distributions and estimate their (time-varying) parameters. Only distributions which support non-negative values for the concentrations were considered. For the P observations, a gamma(ν, β) distribution was found to well characterize the observations. The scale parameter was determined to be β=0.025, and the shape parameter ν varied daily and depends on the mean level, μ, of the process (*i.e.* $\nu=\mu/\beta$). For the N observations, a lognormal distribution was chosen. The mean of this distribution varied daily and its standard deviation, σ, was found the be related to the mean, μ, according the following regression equation: σ=0.67–

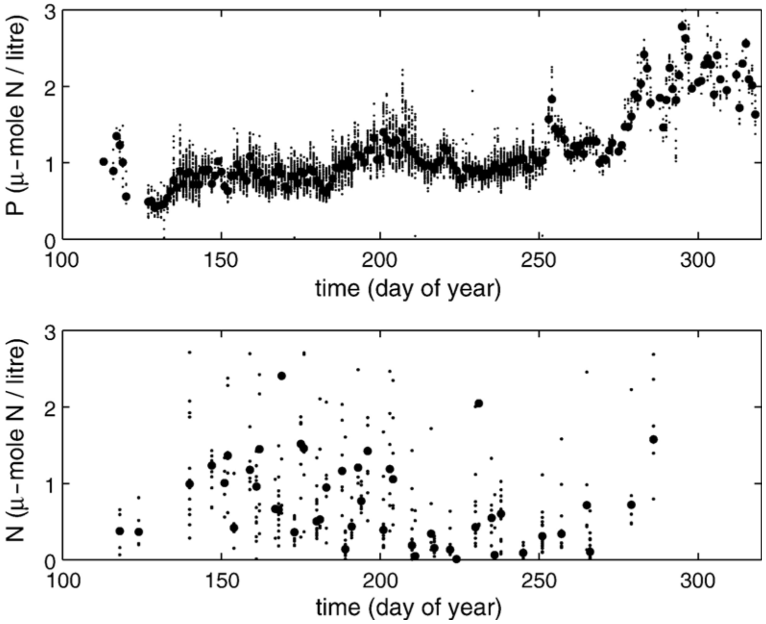

Fig. 3. Observations on phytoplankton and nutrients in Lunenburg Bay for 2004. Available measurements are given by small black dots. The median value on any given day is also shown (large circles).

0.25μ. For days with too few replicates for distribution fitting, these relationships were assumed to be valid. The gamma and lognormal distributions for P and N serve to specify the observation Eq. (2), and hence the likelihood used in the measurement step of Eq. (4) and elsewhere.

3.2. Ecological model

The prototype ecological model for Lunenburg Bay is a simple 0-D biogeochemical model with ecosystem components: phytoplankton (P), nutrients (N) and detritus (D). These prognostic variables are defined within a finite volume and co-evolve according to the following equations:

$$\frac{\mathrm{d}P}{\mathrm{d}t}=\frac{N}{k_N+N}\gamma P-\lambda P^2+\varepsilon_P \tag{14}$$

$$\frac{\mathrm{d}N}{\mathrm{d}t}=\phi D-\frac{N}{k_N+N}\gamma P+\varepsilon_N \tag{15}$$

$$\frac{\mathrm{d}D}{\mathrm{d}t}=-\phi D+\lambda P^2+\varepsilon_D. \tag{16}$$

The state variables are in nitrogen concentration units. All quantities used in this model are summarized in Table 1. The deterministic part of the model is a simplified version of Dowd (2005) and further details can be found there. Additive dynamical noise (ε_P, ε_N, ε_D) is appended to each of the equations as non-conservative source and sink terms (Bailey et al., 2004). These accounts for the exchange (advection and mixing) of ecosystem components with the far field, so that Eqs. (14)–(16) acts as an open system.

Table 1
State variables and parameters in the ecosystem model

Quantity	Units	Value	Definition
State variables (x_t)			
P	μmol nitrogen l^{-1}	–	Phytoplankton biomass
N	μmol nitrogen l^{-1}	–	Inorganic nutrients
D	μmol nitrogen l^{-1}	–	Organic detritus
γ	d^{-1}	–	Phytoplankton growth rate
Parameters (θ_t)			
k_N	μmol nitrogen l^{-1}	2.5	Half-saturation for N uptake by P
λ	μmol nitrogen $\mathrm{l}^{-1}\,\mathrm{d}^{-1}$	0.05	Grazing loss of P
$\phi(t)$	d^{-1}	0.02–0.1	Remaining rate of D to N
$g_{\mathrm{seas}}(t)$	μmol nitrogen $\mathrm{l}^{-1}\,\mathrm{d}^{-1}$	0–1	Seasonally varying growth rate
a	d^{-1}	0.1	Decay/memory for γ

Explicit functional dependence on time (t) is indicated for parameters, along with the range of their values. Here, l is litres and d is days.

The daily averaged light-limited growth parameter is also considered stochastic and time dependent. It evolves according to the (Langevin) equation

$$\frac{\mathrm{d}\gamma}{\mathrm{d}t}=g_{\mathrm{seas}}(t)-a\gamma+\varepsilon_\gamma \tag{17}$$

where $g_{\mathrm{seas}}(t)$ represents deterministic forcing corresponding to the maximum light limited growth rate computed using surface irradiance, attenuation, and a photosynthesis–irradiance curves (c.f. Dowd, 2005). The remaining terms are a decay term and a random forcing term ε_γ. This stochastic dynamic parameter is treated as an additional state variable (Kitagawa, 1998; Dowd, 2006). The decorrelation scale is set at $1/a=10$ days, matching the meterological band and accounting for the effects of fluctuations in light levels and mixing on P growth.

This system of coupled, nonlinear stochastic differential Eqs. (14)–(17) was discretized to yield a stochastic difference equation corresponding to the system model (1). A 4×1 vector thus describes ecosystem state at any time t, *i.e.* $x_t=(P_t, N_t, D_t, \gamma_t)'$. Initial conditions in the form of an initial ensemble must also be specified. These have simply been defined using a normal distribution (truncated to be non-negative) over a reasonable set of values. Note that once observations are assimilated, initial condition have almost no effect on the subsequent analysis for this 0-D ecosystem model.

The system noise process for the ecosystem state variables (ε_P, ε_N, ε_D) and the stochastic parameter (ε_γ) are assumed normally distributed, zero-mean, and independent through time. The variance of ε_γ was set such that it had a level of 20% of the mean of the seasonal growth curve. To specify the variance for the ecosystem state variables consider its interpretation as source and sink terms due to mixing. Denote $\varepsilon_{i,t}$ as the system noise for the ith element of the state vector x_t (similarly for $x_{i,t}$). The quantity $\varepsilon_{i,t}/\Delta t$ can then be identified with the concentration flux in a time increment Δt. If we assume Fickian diffusion wherein this flux into the finite volume scales as $K\times(\Delta x_{i,t})$, where K is an exchange coefficient and $\Delta x_{i,t}$ is a concentration difference. Thus, $\mathrm{var}(\varepsilon_{i,t})=\Delta t^2\times K^2\times\Delta x_{i,t}^2$. For this study, we assume $K=0.5$ day^{-1} (flushing time scale of 2 days), and $\Delta x_{i,t}=0.2x_{i,t}$. Note that these system noise terms have been added in such a way as to ensure non-negative concentrations.

3.3. State estimation

Implementation of the MCMC data assimilation method followed the pseudo-code of Appendix A. From this algorithm it can been seen that there are two main subroutines: (i) *ModelForecast* which corresponds to the ecological model and moves the ecosystem state forward one time increment (note that this is done for each member of the ensemble); and (ii) *Assimilate* which incorporates available measurements according to Bayes theorem by applying the accept/reject rule. The solutions obtained by the MCMC algorithm were verified by comparison to analytic solutions provided by the Kalman filter for simple linear, Gaussian cases. The consistency of the MCMC and SIR solutions for the application here was also verified for very large ensemble sizes.

As a baseline run, a very large ensemble size of $n=250{,}000$ was used to ensure a close match with the true target (posterior) distribution. (This is clearly an unrealistic ensemble size for practical application, but was here used to provide an "exact" solution which facilitates assessment of convergence and computational properties of the algorithms in the next section). The acceptance probabilities of the Metropolis-Hasting MCMC algorithm over time were examined and had a median of 0.65 with an inter-quartile range of 0.14. Occasionally low acceptance probabilities were associated with abrupt shifts in the values of the measurements. This is consistent with Bayes' formula being a measure of the overlap of the likelihood and the predictive density. Also note that the algorithm as given can be easily altered to run longer chains which may be themselves sub-sampled to yield the desired target ensemble.

Fig. 4 shows nowcast results from the sequential MCMC estimation procedure for the ecosystem state variables and the stochastic dynamic growth parameter. Two summary quantities are shown: the median and the 90% confidence region. The observed P are clustered tightly around the median state, and the confidence region contains most of the observations. The ecosystem state variable N is also observable but with fewer and much noisier measurements, and hence wider confidence regions. Its blocky appearance is a result of the sequential nature of the data assimilation cycle wherein after analysis, a (smooth) prediction of the state forward in time is made (with variance growth) followed by an abrupt correction upon encountering the next observations (with variance collapse). The influence of the P measurements on the N state is evident since these variance adjustments are occurring at times with no N observations (*e.g.* at day 130 and 270). The unobserved state variable D is similarly indirectly influenced by the P and N observations, but is smoothed and has a wider confidence interval. The estimates for the dynamic growth parameter show the imposed deterministic seasonal cycle but also the fluctuations which are a consequence of the need to alter the phytoplankton growth rate in a manner consistent with the observations.

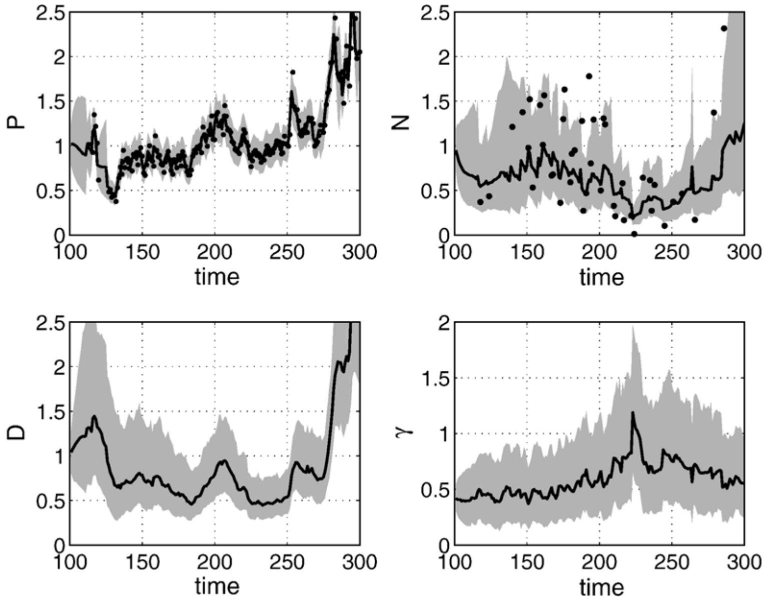

Fig. 4. State estimates for the ecosystem variables and the stochastic dynamic growth parameter. Each panel shows the median (solid line) and the 90% confidence regions (gray shaded area). Median values for observations on P and N are also shown (black dots).

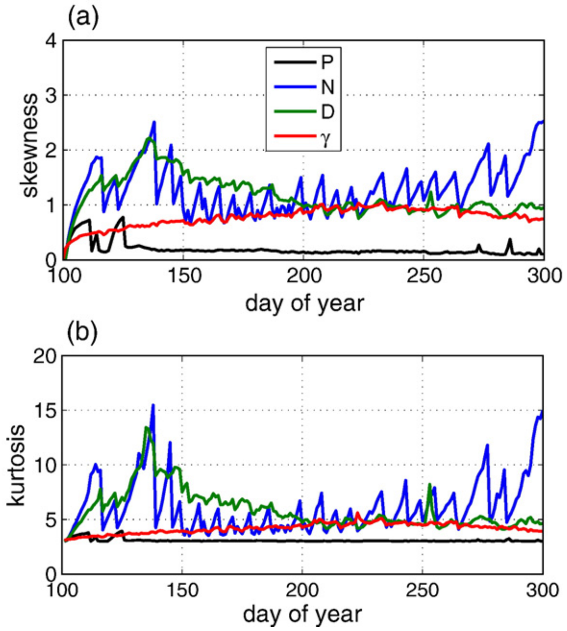

Fig. 5. Skewness (panel a) and kurtosis (panel b) for the ecosystem variables and the stochastic dynamic growth parameter.

Note that near the end of the integration period, mass is being added to the system (*via* the dynamical noise terms) to account for the observed increases in P and N.

Fig. 5 shows the time evolution of the skewness and kurtosis over the analysis period. This is reported for the nowcast ecosystem state variables and the stochastic dynamic growth parameter (note that the first two statistical moments, the mean and the variance, of the nowcast state are not reported since their values are evident from Fig. 4). These higher order statistical moments provide an indication of the extent of non-Gaussianity in the estimated pdf of the ecosystem state. For most of the analysis period, the state variable P has skewness near zero and a kurtosis near three. This indicates that the nowcast P distribution is not so far from a normal distribution. The reason for this near-normality is that P is observed at nearly all analysis times and the gamma distribution used to characterized these measurements itself resembles a normal. This feature influences the results of the Bayesian data assimilation with the nowcast pdf taking on features of the observation distribution. The remaining state variables N, D and γ are more non-Gaussian: they are left skewed, and the kurtosis is greater than three implying they have heavier tails (or are more outlier prone) than the normal distribution.

Another feature of Fig. 5 is the growth in the skewness and kurtosis between observation times, and its abrupt decrease at observation times (this is particularly evident for the observed N, but also found in other state variables). This feature has the same origin as the variance growth and collapse discussed above. These higher order moments grow between observation

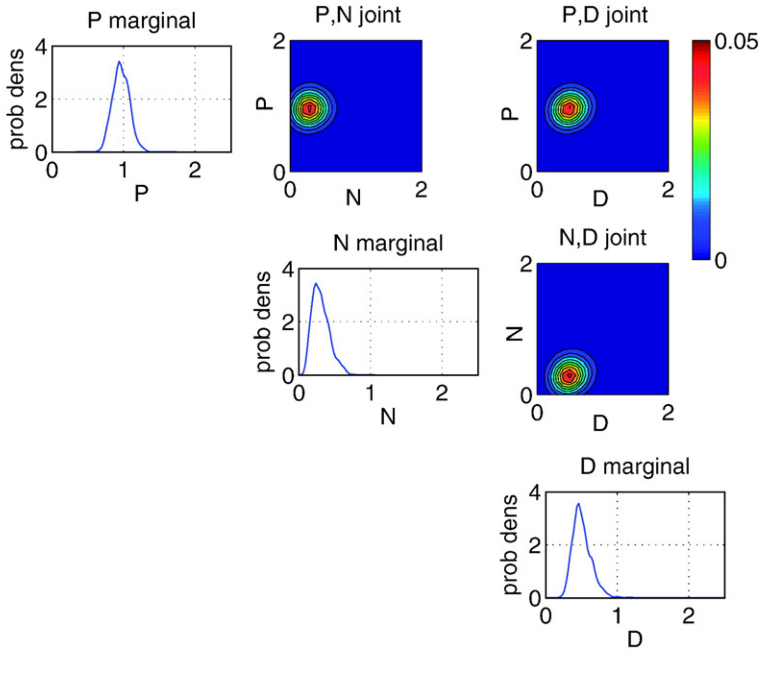

Fig. 6. Marginal (diagonal plots) and joint distribution (off diagonal plots) for the nowcast (filter) distributions, $p(x_t|y_{1:t})$, for day t=245. These are reported for the ecosystem state variables P, N and D and are based on kernel smoothed density estimates.

times due to nonlinear dynamical prediction (unconstrained by observations) that generates non-Gaussian distributions (*e.g.* Miller et al., 1999). At times when direct observations of the state variables are available, the Bayesian assimilation reduces these higher moments consistent with the postulated distributions of the observations.

The sequential data assimilation procedure can also be used to construct probability distributions for the ecosystem state at any given time. Fig. 6 shows the marginal and joint distributions for the nowcast ecosystem state, $p(x_t|y_{1:t})$, at a particular analysis time (t being day 245). These distributions are constructed by kernel smoothed density estimation. The marginal distributions for P, N and D can be compared to the time series results reported for day 245 in Figs. 4 and 5). At day 245, Fig. 6 indicates the marginal distribution for P is symmetric with skewness and kurtosis near that of a normal distribution. The marginal distributions for D and N are clearly non-Gaussian being left skewed and with light tails. The joint distributions are a statement of the relation between the ecosystem state variables; at this analysis time they suggest little in the way of dependence structure for the nowcast state variables.

To examine the role of nonlinear dynamical prediction in changing the probability distributions of the ecosystem state, forecast (predictive) distributions are next considered. Fig. 7 shows a predictive distribution for day 245 based on a 30 day forecast (*i.e.* starting from day 215). Compared to the nowcast distribution in Fig. 6, the marginals of the forecast ecosystem state all have a larger variance and are all left skewed. The joint distributions indicate a greater dependence structure (*e.g.* higher covariance). This latter feature is due to the fact that the forecasted state depends more on the linkages imposed on the state variables by the ecological dynamics, and less on observational information (a 30 day forecast suffices to effectively ‘forget’ the observational information).

To further examine the relative roles of the dynamics and the observations in setting the dependence structure amongst the ecosystem state variables consider the following. The strength of the relationship between two ecosystem state variables can be measured by their mutual information (*e.g.* Kantz and Schreiber, 2003), according to the formula,

$$I(x_{i,t},x_{j,t}) = \int\int p(x_{i,t},x_{j,t}|y_{1:t}) \times \log\frac{p(x_{i,t},x_{j,t}|y_{1:t})}{p(x_{i,t}|y_{1:t})p(x_{j,t}|y_{1:t})}\mathrm{d}x_{i,t}\mathrm{d}x_{j,t}.$$

Here $I(x_{i,t},x_{j,t})$ designates the mutual information between two of the ecosystem state variables $x_{i,t}$ and $x_{j,t}$ at

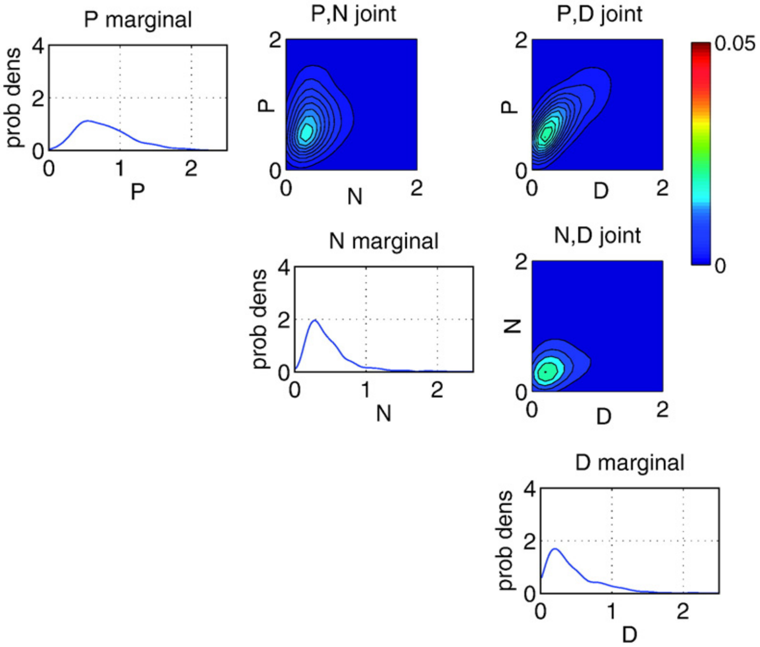

Fig. 7. Marginal (diagonal plots) and joint distribution (off diagonal plots) for the forecast (predictive) distributions, $p(x_{t+\tau}|y_{1:t})$, for day t=245 based on a τ=30 day forecast. These are reported for the ecosystem state variables P, N and D and are based on kernel smoothed density estimates.

time t (here i and j are one of P, N, or D, and $i \neq j$) . This measure may be thought of as a generalization of correlation (*e.g.* a correlation of 0.5 corresponds to a mutual information of 1.2 for variables that follow a bivariate normal). However, unlike correlation it is not restricted to measuring only linear relationships. The quantities on the right hand side involve both the joint and marginal distributions of the ecosystem state. These quantities are produced as part of the statistical data assimilation procedure, examples of which were given for a single analysis time in Figs. 6 and 7.

Fig. 8 reports the time evolution of the mutual information for pairwise combinations of the ecosystem state variables (P and N; P and D; N and D). For the nowcast (filter) density in panel (a) we see relatively weak relationships between the variables (note the vertical scale). The strongest relationship are between N and unobserved D, while the weakest is between P and N (which are both observed). The temporal pattern is one wherein the highest mutual information occurs at times with the fewest observations, consistent with the notion of the ecosystem dynamics imposing dependence structure. Panel (b) shows corresponding results for 30 day forecasts. In all cases, the mutual information is much higher than for the nowcasts as a consequence of the dynamical linkages between the variables playing a larger role (since observations have little influence on the ecosystem state for 30 day forecasts). Here P and D show the strongest relationship, and P and N the weakest.

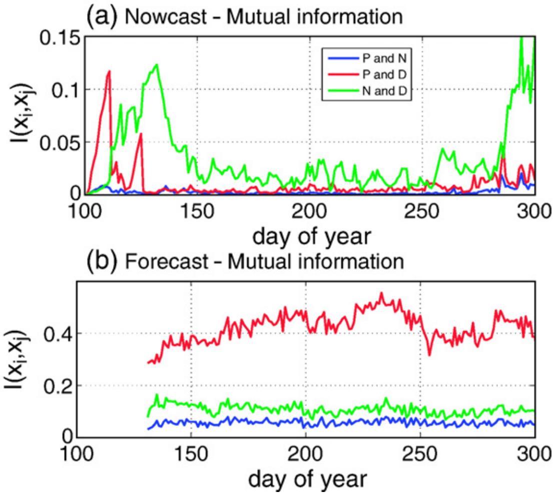

Fig. 8. Mutual information for pairwise combinations of the ecosystem state variables P, N and D. Panel (a) shows the case for the nowcast distributions, and panel (b) shows the case for distributions based on 30 day forecasts.

3.4. Computational aspects

Numerical experiments were carried out to assess the efficiency and effectiveness of the proposed MCMC sequential data assimilation procedure, and to compare it to the other candidate approaches (SIR and the EnKF). The purpose is to examine the performance of the algorithms for various ensemble sizes in terms of their ability to approximate the "exact" solution (*i.e.* the one reported in Section 3.3 and computed numerically using the large ensemble of $n=250{,}000$). Note that ensemble size is directly proportional to the number of model runs at each time step, and is the primary quantity determining the computational load (but recall that there is also a resampling step for SIR, and matrix inversions are required for the EnKF).

Fig. 9 shows convergence of the MCMC method as a function of ensemble size in terms of the statistical moments (the mean, variance, skewness and kurtosis). Each graph shows the root mean squared difference between the exact solution and those computed using various ensemble sizes. This is done for each of the ecosystem state variables (including the growth rate parameter), and for each of the four statistical moments. In every case there is a general pattern of the moments converging to the true solution with increasing ensemble size for all variables. P is well constrained by frequent observations and converges smoothly. However, in some cases (*e.g.* the kurtosis of N and D) the convergence does not always decrease with increasing ensemble size. This highlights the role of sample variation, as well as the difficulty in adequately sampling the tails of the distributions even for relatively large samples.

The performance of the sequential MCMC method was also compared to the other candidate approaches. This was done in terms of the convergence of the overall distributions to the exact solution. The measure for this was based on a time-averaged Kullback–Leibler divergence given as

$$\langle K{-}L\rangle = \left\langle \int \tilde{p}(x_t|y_{1:t}) \log \frac{\tilde{p}(x_t|y_{1:t})}{p(x_t|y_{1:t})} \mathrm{d}x_t \right\rangle. \tag{18}$$

The Kullback–Leibler divergence (Kullback and Leibler, 1951) is the quantity inside the angle brackets on the right hand side and measures the discrepancy between two probability density functions. Here, these two distributions are the "exact" target distribution $p(x_t|y_{1:t})$ (computed *via* sequential MCMC procedure with $n=250{,}000$) and the approximate distribution $\tilde{p}\ (x_t|y_{1:t})$

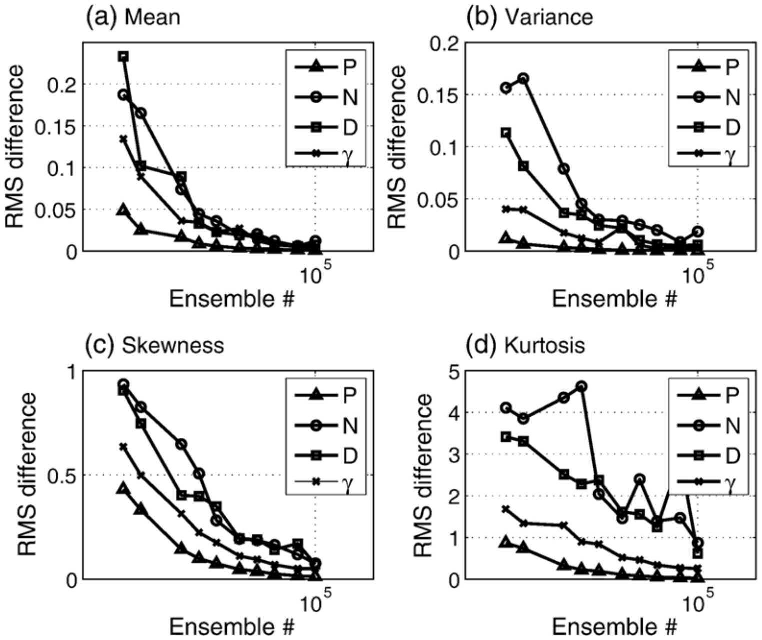

Fig. 9. Convergence to the exact solution of the statistical moments (mean, variance, skewness, kurtosis) for the sequential MCMC method for each of the model variables. Results are reported as root-mean-square difference between approximate solution (calculated for various ensemble sizes) and the exact solution.

(computed using various ensemble sizes and for each of the three candidate sequential data assimilation approaches). For reporting purposes, the Kullback–Leibler divergence has been time averaged over the entire analysis period, as denoted by the angle brackets in Eq. (18). Fig. 10 shows the convergence of the different

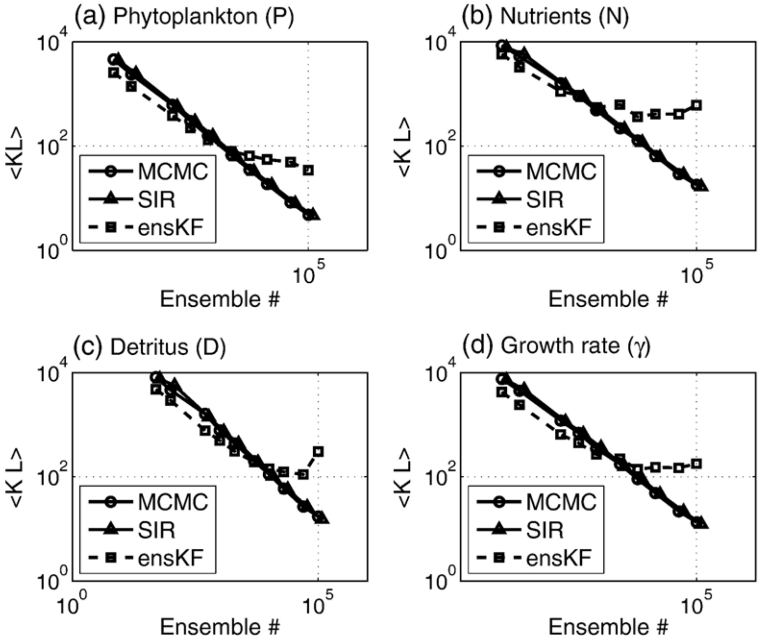

Fig. 10. Convergence to the exact solution of the overall distribution as computed by the sequential MCMC method. Results are reported in terms of the time averaged Kullback–Leibler divergence for various ensemble sizes. Each of the three candidate methods (the MCMC, SIR and EnKF) are considered for each of the four model variables.

sequential data assimilation procedures to the target distribution as a function of the ensemble size. Both SIR and MCMC converge to the true solution with increasing ensemble size in an almost identical fashion. As expected, the (approximate) ensemble Kalman filter does not converge to the true solution with increasing ensemble size. However, a notable point is that for $n < 1000$ the ensemble Kalman filter appears to be slightly closer to the true solution than the other methods. This is likely due to the smearing of the ensemble through use of the Kalman gain equations (van Leeuwen, 2003). Note that raw unsmoothed and normalized histograms (not kernel smoothed density estimates) of the SIR and MCMC ensembles were used in computing Eq. (18). However, by instead making use of kernel smoothed density estimation (Silverman, 1986), this result might change.

To further examine the performance of the ensemble Kalman filter, the convergence of the first two moments (the mean and variance) is shown in Fig. 11. Higher order moments are not reported here since the EnKF is not intended to produce such estimates (in fact, it was found that skewness and kurtosis from the EnKF in some cases actually diverged with increasing ensemble size). The mean and variance of state variable P is best estimated, while the variable N is estimated poorly. However, while the mean and variance do not converge to their true values with increasing ensemble size, they quickly reach an asymptotic equilibrium value. For the mean this occurs after $n = 100$, while for the variance it occurs after about $n = 1000$.

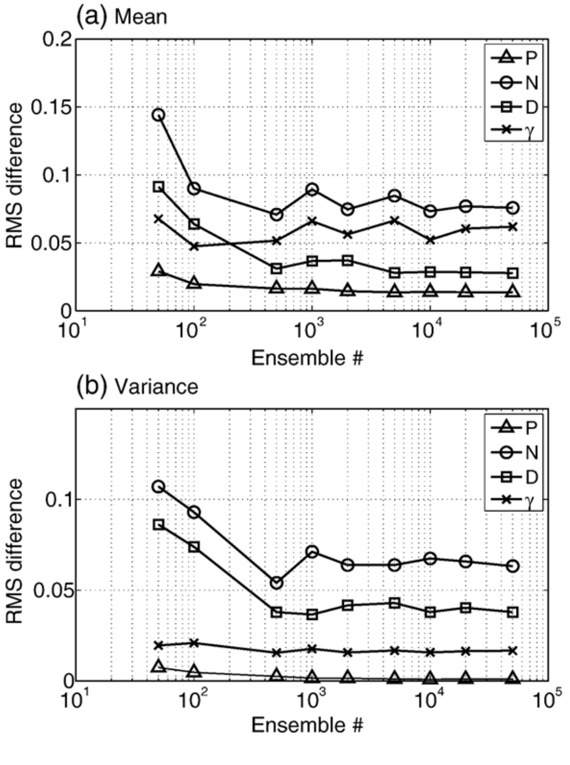

Fig. 11. Convergence to the exact solution of the mean (panel a) and variance (panel b) of the ensemble Kalman filter for each of the model variables. Results are reported as root-mean-square difference between approximate solution (calculated for various ensemble sizes) and the exact solution.

4. Discussion and conclusions

Sequential data assimilation is of central interest for problems in marine prediction, as well as the basis for emerging operational forecasting systems (Pinardi et al., 2003). This study examined advanced Bayesian statistical data assimilation for nowcasting and forecasting in marine ecological systems. These approaches are designed to give full consideration to nonlinear dynamics, as well as to emerging data types characterized by complex spatial and temporal structure and non-normal distributions. Existing sequential data assimilation techniques mainly rely on approximations based on the Kalman filter, usually involving linearization and/or dimension reduction (see Bertino et al., 2003). An alternative Monte Carlo based method is the ensemble Kalman filter (Evensen, 2003). The EnKF approximates the Bayesian measurement update step, and so cannot fully account for general non-Gaussian measurements. The more general case of strong dynamical nonlinearities and non-normal observational distributions requires a fully Bayesian approach (Dowd and Meyer, 2003; van Leeuwen, 2003). The state space model was put forth here as a comprehensive statistical framework for the data assimilation problem. Its general solution for the nonlinear and non-Gaussian case has a Bayesian probabilistic formulation (Jazwinski, 1970). Here, a novel and straightforward sequential data assimilation technique was offered based on a Markov Chain Monte Carlo (MCMC) approach. This provided for an exact solution in terms of a time evolving ensemble (or sample) that represents the probability density function governing the multivariate ecological state.

Markov Chain Monte Carlo (MCMC) methods are a class of approaches for computational Bayesian statistics (*e.g.* Gamerman, 1997). They have revolutionalized statistical applications, but have been little applied to the problem of oceanographic data assimilation. MCMC has, however, been used for parameter estimation in plankton models (Harmon and Challenor, 1997). Dowd and Meyer (2003) also considered it for hindcasting using a simple ecosystem model with general purpose Bayesian software (BUGS, Spiegelhalter et al., 1995). There it was found that this Gibbs sampling MCMC software was not readily tailored to time-dependent systems used for dynamical modelling. In this study, it was shown how the Metropolis-Hastings algorithm can be adapted for sequential data assimilation. The algorithm is straightforward to apply with complex ecosystem models, being highly modular and having the ecosystem model called simply as a subroutine. The MCMC algorithm works simply through stochastic dynamic prediction (ensemble forecasting), and a simple acceptance/rejection rule for the assimilation of observations. Unlike SIR, the approach does not rely on resampling and consequently does not exhibit sample (ensemble) degeneracy, a problem that itself has been the subject of a great deal of research effort (Arulampalam et al., 2002; Dowd, 2006). The MCMC data assimilation approach introduced in this paper produces an independence chain (Tierney, 1994), and so effectively eliminates the well-known MCMC issues of "burn-in", and facilitates mixing of the chain.

This study has repeatedly emphasized that ecological state variables do not, as a rule, follow a normal distribution, but rather have a non-Gaussian structure that is not readily described by parameteric probability distributions. It is well known that nonlinear models generate non-Gaussian distributions (Miller et al., 1999), as does the Bayesian assimilation of non-normal observations. To examine these features, a test application was undertaken using a 0-D nonlinear plankton ecosystem model and multivariate non-Gaussian observations from a coastal ocean observatory. Such a stochastic ecosystem model is a core element of the "weak constraint" approach, including stochastic environmental variations as well as model errors (Marion et al., 2000; Dowd, 2006). The application also treated an unknown parameter (the phytoplankton growth rate) as dynamic variable following a pre-defined stochastic process. The resultant ensemble that described the ecosystem state was used to construct marginal and joint distributions of the ecosystem state variables. It was also used to compute summary statistics like the median and confidence intervals, as well as measures of non-normality like skewness and kurtosis. It was shown how the ecosystem model imposed dependence structure on the ecosystem variables due to dynamical linkages (especially evident for forecasts). To effectively treat the non-Gaussian information that emerges from this data assimilation approach, some ideas from information theory have been suggested: mutual information provided a general measure for the dependence structure between state variables; and the Kullback–Leibler divergence was used to assess the convergence of the algorithms towards the true solution.

An important practical consideration for ensemble based statistical data assimilation methods is their efficiency and effectiveness in characterizing the ecosystem state. Towards this end, computational and convergence properties of various sequential Monte Carlo methods (MCMC, SIR and the EnKF) were examined in the context of a simple ecosystem model. It was found that MCMC and SIR had similar convergence towards the exact solution. The approximate ensemble Kalman filter also proved efficient with respect to recovering approximate values for the mean and the variance, at least until the ensemble size exceeded 100–1000, after which it could no longer approach the true solution. The ensemble size is the primary determinant of the computational load required for sequential data assimilation (SIR also requires resampling and the EnKF requires evaluating matrix expressions and inverses). An important general point, which is demonstrated in the analysis of this study, is that the ensemble size required for data assimilation will depend on the quantity desired. That is, an estimate for the mean state (or its variance) will require much smaller ensembles that if higher order moments (skewness or kurtosis) or full probability distributions are of interest.

An outstanding research question is the extent to which these ensemble based Bayesian data assimilation approaches can be adapted for problems of more realistic dimensions, such as multi-compartment ecosystem models coupled to three-dimensional ocean circulation models (*e.g.* Oschlies and Schartau, 2005). In contrast to the example in this study, these problems are described by PDE based biophysical models with the state vector containing spatial information on the prognostic ecosystem models, and typically having dimensions of 10^5 or greater (the effective degrees of freedom is much less). The challenge is to keep the ensemble size as small as possible, while maintaining adequate performance of the data assimilation algorithm (this study provided some general approaches to quantitatively measure this tradeoff.) For physical systems, Brusdal et al. (2003) applied an EnKF with a state vector of 10^6 and an ensemble size of 150, and van Leeuwen (2003) used a modified SIR procedure with an ensemble size of 495 and a state vector dimension of 2×10^5. Studies of coupled biophysical models using an EnKF have been successful with ensemble sizes of 100 to 200 (Allen et al., 2003; Natvik and Evensen, 2003), while SIR based applications have used ensemble sizes of 1000 (Losa et al., 2003). The MCMC approach operates on similar principles to SIR and should have comparable ensemble size requirements.

All of the ensemble based sequential data assimilation methods rely on forecasting *via* Monte Carlo integration. They differ only in the way in which observations are assimilated. The MCMC approach to Bayesian data assimilation uses stochastic dynamic prediction after which the candidate ensemble member is evaluated for inclusion in the final ensemble (characterizing the target pdf) using an accept/reject step. In general, this procedure will preferentially accept ensemble members that are closest to the observations. The acceptance probability is a key measure for the computational efficiency of Metropolis-Hastings methods (Chib and Greenberg, 1995). It will made highest when forecasts closely match new observations. The system noise is a key determinant for maximizing the acceptance probability as it dictates the spreading of the ensemble and how many particle will intersect with regions of state space where the measurement pdf takes on non-negligible values. To improve ensemble forecasts for Bayesian data assimilation other possibilities should also be considered. For example, Chorin and Krause (2004) develop an approach to populate the state space with particles (ensemble members) in the regions into which the dynamics are expanding. Another possibility is to smear the ensemble by adding 'jitter' to the ensemble members, or perhaps by sampling from kernel smoothed density estimates or fitted parameteric forms of distributions. Adaptive approaches which use a variable ensemble size to ensure that the ensemble well represents the ecosystem state are readily implemented *via* the MCMC algorithm.

In summary, advanced statistical approaches for the data assimilation problem are currently being widely applied in Statistics and Engineering. These treat general nonlinear dynamics and non-Gaussian measurements using modern Bayesian computational approaches. It is timely to consider the adaptation of these approaches to oceanographic data assimilation. They are ideally suited to problems in marine ecology and biogeochemistry due to their strong nonlinearities and structural uncertainty, along with the need to use ecological measurements having complex spatial and temporal structure. This study represents an initial step towards bringing these statistical approaches to bear on the data assimilation problem, and to allow researchers to consider non-Gaussian aspects of the problem. The implementation of these probabilistic approaches to data assimilation is surprisingly straightforward, and it is hoped that this work will encourage marine modelling community to experiment with these novel statistical data assimilation procedures.

Acknowledgements

The author would like to thank many colleagues in both Oceanography and Statistics for useful discussion including: John Cullen, Andy Edwards, Arnaud Doucet, Maud Guarracino, Chris Jones, Svetlana Losa, Renate Meyer, Bruce Smith, Keith Thompson, and Alain Vezina. This work was supported by an NSERC Discovery Grant, the National Program for Complex Data Structures, and the Canadian Foundation for Climate and Atmospheric Science.

Appendix A. Pseudo-code for Metropolis-Hastings MCMC

```
**Sequential MCMC data assimilation main program

FOR t=1 to T
    IF t is a measurement time
        $\{x_{t|t}^{(i)}\} = Assimilate(y_t, \{x_{t-1|t-1}^{(i)}\})$
    ELSE IF t is NOT a measurement time
        FOR i=1 to N
        $x_{t|t}^{(i)} = ModelForecast(x_{t-1|t-1}^{(i)}, \theta_t, n_t^{(i)})$
        END FOR
    END IF
END FOR

**Subroutine for M-H MCMC measurement update

SUBROUTINE $\{x_{t|t}^{(i)}\} = Assimilate(y_t, \{x_{t-1|t-1}^{(i)}\})$
- initialize $x_{t|t}^{(1)}$ (with, for example, $y_t$ or the median of
$\{x_{t-1|t-1}^{(i)}\}$)
    FOR i=2 to N
        - choose random pre-trial particle $x_{t-1|t-1}^{*}$ from
          ensemble $\{x_{t-1|t-1}^{(i)}\}$
        - generate trial particle
          $x_{t|t-1}^{*} = ModelForecast(x_{t-1|t-1}^{*}, \theta_t, n_t^{*})$
        - compute acceptance probability as:
        $\alpha = \min\left(1, \frac{p(y_t|x_{t|t-1}^{*})}{p(y_t|x_{t|t}^{(i-1)})}\right)$
        - draw $z$ from $Uniform(0,1)$ and set
        $x_{t|t}^{(i)} = \begin{cases} x_{t|t-1}^{*} & \text{if} \quad z \le \alpha \\ x_{t|t}^{(i-1)} & \text{if} \quad z > \alpha \end{cases}$

    END FOR
RETURN
```

References

Allen, J.I., Eknes, M., Evensen, G., 2003. An ensemble Kalman filter with a complex marine ecosystem model: hindcasting phytoplankton in the Cretan Sea. Annales Geophysicae 21, 399–411.

Arulampalam, M.S., Maskell, S., Gordon, N., Clapp, T., 2002. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. IEEE Transactions on Signal Processing 50 (2), 174–188.

Bailey, B., Doney, S.C., Lima, I.D., 2004. Quantifying the effects of dynamical noise on the predictability of a simple ecosystem model. Environmetrics 15, 337–355.

Bertino, L., Evensen, G., Wackernagel, H., 2003. Sequential data assimilation techniques in oceanography. International Statistical Review 71 (2), 223–241.

Brusdal, K., Brankart, J.M., Halberstadt, G., Evensen, G., Brasseur, P., van Leeuwen, P.J., Dombrowsky, E., Verron, J., 2003. Results from a comparison between the ensemble Kalman filter, the ensemble Kalman smoother, and the SEEK filter with real satellite observations in the North Atlantic, as performed in the DIADEM-project. Journal of Marine Systems 40–41, 253–289.

Chib, S., Greenberg, E., 1995. Understanding the Metropolis-Hastings algorithm. The American Statistician 49 (4), 327–335.

Chorin, A.J., Krause, P., 2004. Dimensional reduction for a Bayesian filter. Proceedings of the National Academy of Sciences 101 (42), 15013–15017.

Dickey, T.D., 2003. Emerging ocean observations for interdisciplinary data assimilation systems. Journal of Marine Systems 40, 5–48.

Doucet, A., de Freitas, N., Gordon, N., 2001. Sequential Monte Carlo Methods in Practice. Springer, New York. 581 pp.

Dowd, M., 2005. A biophysical coastal ecosystem model for assessing environmental effects of marine bivalve aquaculture. Ecological Modelling 183 (2–3), 323–346.

Dowd, M., 2006. A sequential Monte Carlo approach to marine ecological prediction. Environmetrics 17, 435–455.

Dowd, M., Meyer, R., 2003. A Bayesian approach to the ecosystem inverse problem. Ecological Modelling 169, 39–55.

Edwards, A.M., 2001. Adding detritus to a nutrient–phytoplankton–zooplankton model: a dynamical systems approach. Journal of Plankton Research 23, 389–413.

Edwards, A.M., Platt, T., Sathyendranath, S., 2004. The high nutrient, low chlorophyll regime of the ocean: limits on biomass and nitrate before and after iron enrichment. Ecological Modelling 171, 103–125.

Eknes, M., Evensen, G., 2002. An Ensemble Kalman filter with a 1-D marine ecosystem model. Journal of Marine Systems 36 (1–2), 75–100.

Evans, G.T., 2003. Defining misfit between biogeochemical models and data sets. Journal of Marine Systems 40–41, 49–54.

Evensen, G., 1994. Sequential data assimilation with a non-linear quasigeostrophic model using Monte Carlo methods to forecast error statistics. Journal of Geophysical Research 99, 10143–10162.

Evensen, G., 2003. The ensemble Kalman filter: theoretical formulation and practical implementation. Ocean Dynamics 53, 343–367.

Fennell, W., Neumann, T., 2004. Introduction to the Modelling of Marine Ecosystems. Elsevier, Amsterdam. 330 pp.

Gamerman, D., 1997. Markov Chain Monte Carlo. Chapman & Hall/CRC, New York. 245 pp.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2003. Bayesian Data Analysis. Chapman and Hall/CRC, Boca Raton. 696 pp.

Gilks, W.R., Berzuini, C., 2001. Following a moving target — Monte Carlo inference for dynamic Bayesian models. Journal of the Royal Statistical Society B 63 (1), 127–146.

Gordon, N.J., Salmond, D.J., Smith, A.F.M., 1993. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. IEE Proceedings-F 140 (2), 107–113.

Harmon, R., Challenor, P., 1997. A Markov chain Monte Carlo method for estimation and assimilation into models. Ecological Modelling 101, 41–59.

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

Huisman, J., Weissing, F.J., 2001. Biological conditions for oscillations and chaos generated by multi-species competition. Ecology 82, 2682–2695.

Huot, Y., Brown, C.A., Cullen, J.J., submitted for publication. Simultaneous inversion of reflectance, the diffuse attenu-ation coefficient, and sun-induced fluorescence in coastal waters. Journal of Geophysical Research.

Jazwinski, A.H., 1970. Stochastic Processes and Filtering Theory. Academic Press, New York. 376 pp.

Kantz, H., Schreiber, T., 2003. Nonlinear Time Series Analysis. Cambridge University Press, Cambridge. 386 pp.

Kitagawa, G., 1987. Non-Gaussian state-space modeling of nonstationary time series. Journal of the American Statistical Association 82, 1032–1041.

Kitagawa, G., 1996. Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. Journal of Computational and Graphical Statistics 5, 1–25.

Kitagawa, G., 1998. A self-organizing state space model. Journal of the American Statistical Association 93 (443), 1203–1215.

Kullback, S., Leibler, R.A., 1951. On information and sufficiency. Annals of Mathematical Statistics 22 (1), 7986.

Lawson, L.M., Sptiz, Y.H., Hofmann, E.E., Long, R.L., 1995. A data assimilation technique applied to a predator-prey model. Bulletin of Mathematical Biology 57, 593–617.

Lee, D.S., Chia, N.K.K., 2002. A particle algorithm for sequential Bayesian parameter estimation and model selection. IEEE Transactions on Signal Processing 50 (2), 326–336.

Losa, S.N., Kivman, G.A., Schroter, J., Wenzel, M., 2003. Sequential weak constraint parameter estimation in an ecosystem model. Journal of Marine Systems 43, 31–49.

Marion, G., Renshaw, E., Gibson, G., 2000. Stochastic modelling of environmental variation for biological populations. Theoretical Population Biology 57, 197–217.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E., 1953. Equation of state calculations by fast computing machines. Journal of Chemical Physics 21, 1087–1092.

Miller, R.N., Carter, E.F., Blue, S.T., 1999. Data assimilation into nonlinear stochastic models. Tellus 51A, 167–194.

Natvik, L.J., Evensen, G., 2003. Assimilation of ocean colour data into a biochemical model of the North Atlantic. Part 1. Data assimilation experiments. Journal of Marine Systems 40–41, 127–153.

Oschlies, A., Schartau, M., 2005. Basin-scale performance of a locally optimized marine ecosystem model. Journal of Marine Research 63 (2), 335–358.

Pham, D., Verron, J., Roubaud, M., 1997. Singular evolutive Kalman filter with EOF initialization for data assimilation in oceanography. Journal of Marine Systems 16 (3–4), 323–340.

Pinardi, N., Allen, I., Demirov, E., De Mey, P., Korres, G., Lascaratos, A., Le Traon, P.-Y., Maillard, C., Manzella, G., Tziavos, C., 2003. The Mediterranean ocean forecasting system: first phase of implementation (1998–2001). Annales Geophysicae 21, 3–20.

Ristic, B., Arulampalam, S., Gordon, N., 2004. Beyond the Kalman Filter: Particle Filters for Tracking Applications. Artech House, Boston.

Silverman, B.C., 1986. Density Estimation for Statistics and Data Analysis. Chapman and Hall, New York.

Spiegelhalter, D.J., Thomas, A., Best, N.G., Gilks, W.R., 1995. BUGS: Bayesian Inference Using Gibbs Sampling, Version 0.50. MRC Biostatistics Unit, Cambridge.

Tierney, L., 1994. Markov chains for exploring posterior distributions (with discussion). Annals of Statistics 22, 1710–1762.

Tsuji, S., Nakamura, M., 1973. Nonlinear Filter using a statistical processing technique. Transaction IEEJ (Institute of Electrical Engineers of Japan) 93-C (5), 109–115.

Vallino, J.J., 2000. Improving marine ecosystem models: use of data assimilation and mesocosm experiments. Journal of Marine Research 58, 117–164.

van Leeuwen, P.J., 2003. A variance-minimizing filter for large-scale applications. Monthly Weather Review 131, 2071–2084.

van Leeuwen, P.J., Evensen, G., 1996. Data assimilation and inverse methods in terms of a probabilistic formulation. Monthly Weather Review 124, 2898–2913.