Special Issue Paper

Environmetrics

Received: 6 October 2013, Revised: 22 January 2014, Accepted: 24 January 2014, Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/env.2264

A statistical overview and perspectives on data assimilation for marine biogeochemical models

Michael Dowd[a]*, Emlyn Jones[b] and John Parslow[b]

Marine biogeochemistry refers to the processes associated with the planktonic ecosystem of the ocean. These are central to nutrient, carbon, and energy cycling, as well as providing the basis of the marine food chain. The field is being revolutionized by new data types and observing platforms, as well as by improvements in ocean modelling brought about by increasing computer power. To further our understanding of these systems, statistical estimation and inference are needed to combine the information in these data with dynamic models to provide improved estimates for the ocean's biogeochemical (BGC) state and its parameters. Such methodologies are termed data assimilation (DA). This paper seeks to provide an overview of DA for the emerging area of marine BGC modelling. A statistical framework is offered, and DA methods that are applicable to the spatio-temporal dynamic models and data that define the BGC problem are reviewed. In addition to this primer on current BGC DA approaches, we offer our perspectives on the challenges and future work necessary to advance this field. This work emerged from a symposium on marine BGC DA that took place in Hobart, Australia, on 28–30 May 2013. Copyright © 2014 John Wiley & Sons, Ltd.

Keywords: data assimilation; ecological modelling; state space models; dynamic models; Bayesian hierarchical; particle filters

1. INTRODUCTION

Improved estimation and prediction of ocean biology is foundational for improving our understanding of how the oceans work, and how best to manage them. The lowest trophic levels of the ocean comprise the planktonic ecosystem, and this plays a major role in the cycling of matter and energy, as well as providing the basis of the marine food chain. The study of planktonic ecosystem dynamics and their influence on the nutrient and carbon cycles of the ocean is referred to as marine biogeochemistry. This includes processes such as photosynthesis, predation, the production and breakdown of organic matter, and the dynamics of inorganic nutrients. This biology and chemistry all take place within a dynamic fluid environment brought about by the circulation of the ocean waters.

There has been a revolution in ocean-observing technologies that allows us an unprecedented view of marine biogeochemistry. Satellites have, for more than two decades now, produced global surface fields for ocean chlorophyll (the photosynthetic pigment of phytoplankton), resulting in a manyfold increase in data volume compared with traditional data sources (Boyce *et al.*, 2012). Advances in ocean optics are allowing for automated measurements of other variables, such as nutrients (Johnson *et al.*, 2013). New observing platforms, such as autonomous underwater vehicles, allow us to transect through the interior of the ocean and measure biogeochemical (BGC) fields (Schofield *et al.*, 2013). Despite this, the ocean is still considered to be undersampled with respect to the scales of variability that characterize its biogeochemistry, but this is changing. Information technologies and advances in computing power are bringing these data online and propelling forward advances in ocean modelling. Marine BGC models are becoming an integral part of many ocean modelling platforms, such as the Regional Ocean Modelling System (Haidvogel *et al.*, 2008). The challenge is to bring together these data and dynamic models through data assimilation (DA) to improve estimation, prediction, and ultimately our understanding of marine biogeochemistry.

Data assimilation refers to the quantitative methods by which the information in dynamic models and data are combined to provide estimates of the ocean state and its key parameters (Evensen, 2009). The term DA was coined in the geosciences and is now a well-developed discipline for physical systems such as the ocean and atmosphere. Its more statistical counterpart is often termed physical–statistical modelling. The first studies in this area treated the spatio-temporal analysis of wind fields (Royle *et al.*, 1999; Wikle *et al.*, 2001) and has now advanced to encompass fully Bayesian approaches for DA (Wikle and Berliner, 2007; Cressie and Wikle, 2011). DA is emerging as a way to improve models of ocean biology and biogeochemistry (Brasseur *et al.*, 2009; Matear and Jones, 2011). However, marine BGC DA offers some unique statistical challenges distinct from its physical counterparts (Parslow *et al.*, 2013). For instance, there is considerable uncertainty in the form and structure of the governing equations, and there are wide ranges for reported parameters from field and laboratory

* *Correspondence to: Michael Dowd, Department of Mathematics and Statistics, Dalhousie University, 6316 Coburg Road, PO Box 15000 Halifax, Nova Scotia, Canada, B3H 4R2. E-mail: michael.dowd@dal.ca*

a *Dalhousie University, Halifax, Canada*

b *CSIRO, Hobart, Australia*

studies. In addition, BGC observations typically have a much more complex error structure, as they are often quite indirect measurements (e.g. using optical proxies to determine phytoplankton biomass).

The purpose of this paper is to carry out a statistically oriented overview of the current state of the art in marine BGC DA, provide a primer on the main approaches used, and most importantly to point out some outstanding issues and future directions necessary for progress of the field. We emphasize that the marine BGC DA problem is a subclass of more general estimation problems for dynamic systems. An excellent overview of physical–statistical modelling and DA is provided by Wikle and Berliner (2007). Here, we explore its specific application for BGC DA but emphasize the issues, and approaches taken, will inform the more general field of research.

2. PROBLEM STATEMENT

Our general goal is to estimate the BGC state of the ocean by making use of all available sources of information: dynamic models, observations, and prior knowledge on key parameters from field and laboratory studies. The system state describes the planktonic ecosystem and is comprised of a set of BGC variables, which are spatial fields defined over the ocean and evolving through time. Broadly speaking, the central variables of interest are the inorganic nutrients (e.g. nitrate), phytoplankton (microscopic plants), zooplankton (microscopic animals), and detritus (dead organic matter). This leads to the commonly used label 'NPZD', although other variables are typically also included, such as chlorophyll. Figure 1 shows a schematic diagram of the BGC system and describes the interactions between the system components. The BGC variables are embedded within the oceanic fluid environment and so redistributed by currents and influenced by external environmental drivers, or forcings, such as light, wind, currents, and heat.

Dynamic models are based on the accumulated scientific understanding of how the BGC variables interact according to processes such as photosynthesis, predator–prey interactions, physiology, and ocean chemistry, as determined from a century of laboratory and field studies. Models of BGC systems are based on the nonlinear differential equations of ecological population dynamics (see Gentleman, 2002, for a review and historical introduction to BGC models). For realistic studies, these are generally coupled to numerical ocean circulation models, rendering this a physical–biological system. Important new approaches in the field incorporate species diversity through selection-based models (Follows and Dutkiewicz, 2011) and individual variation through the use of agent-based simulation (Woods, 2005). An alternative approach to improve estimation and prediction for these system, is to consider statistical approaches that make use of dynamic models and available BGC observations (which take the form of point observations, time series, spatial fields, and spatio-temporal transects). Such DA approaches are the focus of this paper.

The state space model provides a useful statistical framework for considering dynamic models and data. The methodologies that arise from its solution form the basis for BGC DA. The state space model is comprised of two parts: (i) a dynamic model equation and (ii) an observation equation. These are each considered in turn below.

The dynamic model equation is given here as a discrete-time model. This formulation is simply meant to represent the transition of the dynamic system over an arbitrary time interval, $t-1$ to t. This time interval may be the time between observations or a predefined set of analysis times for which estimates are desired. Hence, the discrete-time framework can incorporate any continuous-time representation of a dynamic system and any numerical integration schemes that allow for forward prediction. It takes the following form

$$x_t = f\left(x_{t-1}, \theta, \gamma_t, e_t\right) \tag{1}$$

where x_t is a vector that defines the state of the system at time t and so contains the prognostic BGC variables of interest. The biological parameters contained within the BGC model are given by θ. The nonlinear operator f provides for the functional form of the governing

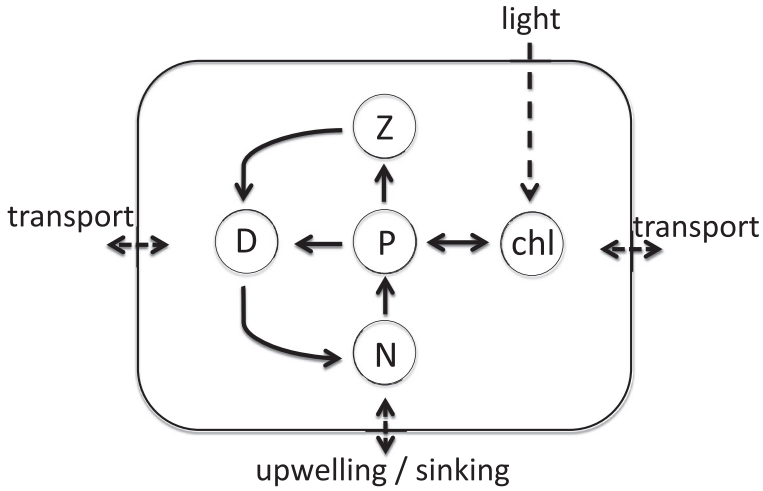

Figure 1. Schematic of a basic marine biogeochemical system. Here, N designates nutrients, P is phytoplankton, Z is zooplankton, D is detritus, and chl is chlorophyll. The system operates by sunlight driving photosynthesis through chl, which, together with inorganic N, produces P biomass. The P biomass is either consumed by Z or dies and goes to the organic D pool, along with Z. The organic D is remineralized into inorganic N. All components are subject to transport by ocean currents and can also sink or be upwelled into, or out of, the deep ocean.

equations that predict the temporal co-evolution of the BGC variables. The model is forced by a set of exogenous environmental inputs, γ_t. There are also model errors, e_t, due to uncertainty in the dynamics and forcings. For BGC problems, stochastic elements typically enter multiplicatively. Another alternative form for the state space model is to write the state transition probability distribution directly, that is, $[x_t|x_{t-1}, \theta]$, where the notation $[a|b]$ represents the conditional probability distribution of a given b. It is assumed here that any initial conditions, x_0, and forcing, γ_t, are given and, for simplicity, they are not shown in the distributions. For BGC models, this Markovian transition density is not generally available as a closed-form solution, and we rely on the representation (1) in practice. Features of the dynamic model specific to BGC modelling are taken up in Section 2.1.

A set of measurements, y_t (for $t = 1, \ldots, T$) is available on the BGC state. The complete observation set is designated as $y_{1:T}$. The measurement equation for the state space model can be written in the form of a data model as the distribution $[y_t|x_t, \phi]$, where y_t represents the observations and ϕ are parameters of the measurement model. An alternative form for the data model used in the state space representation is

$$y_t = h(x_t, \phi, v_t) \tag{2}$$

where h is a measurement operator and v_t represents the measurement error. This form is useful in that it makes explicit the fact that you may not be measuring the same quantity that is being dynamically modelled (e.g. you may be measuring chlorophyll but modelling phytoplankton nitrogen content). This conversion between measured and prognostic variables embodied in h involves a mathematical transformation of the state variable that is generally derived from an instrument calibration exercise. Note that the form (2) is, however, likely less useful in practice than writing down a parametric form for the data model directly. Further information on measurement models specific to the BGC problem is taken up in Section 2.2.

The goal of the state space model is to make inferences about the BGC state of the ocean, x_t, and the biological parameters, θ (and sometimes for the parameters, ϕ, of the measurement model). The main information used are the measurements, y_t, and the dynamic model (1). We also assume knowledge of the distributions of the model error, e_t, and the measurement error, v_t. Any conversion from measured quantities to prognostic variables, that is, the h operator, is considered known. For a complete Bayesian formulation of the inference problem, the state space model, as posed, should be complemented by a prior (statistical) model for the uncertain parameters (Section 3).

2.1. Dynamic models

The dynamic models used for BGC DA are often large-scale spatio-temporal models. The BGC variables ($B_i,\ i = 1, \ldots, m$) are essentially treated as a set of m interacting tracers embedded within an ocean circulation model. They are formulated as a set of partial differential equations of the general form

$$\frac{\partial B_i}{\partial t} + u \cdot \nabla B_i - \nabla \cdot (\kappa \nabla B_i) = \tilde{f_i}\,(B_1, \ldots, B_m, \theta, \gamma),\ \ i = 1, \ldots, m. \tag{3}$$

Here, B_i represents a three-dimensional spatial field (3-D: latitude s_x, longitude s_y, and depth s_z) of the ith prognostic BGC variable, u is the 3-D velocity or current field, and κ is the matrix of diffusion coefficients. The symbol '$\cdot$' represents the inner product of two vectors, and $\nabla = (\partial/\partial s_x, \partial/\partial s_y, \partial/\partial s_z)'$ is the 3-D spatial gradient operator. The equations are applied over the spatial domain of interest (which may be a bay, the continental shelf, an ocean basin, or the global ocean).

Typical prognostic BGC variables include nutrients, phytoplankton, zooplankton, and detritus (note that these plankton groupings may even be further subdivided into size fractions or functional groupings). The last two terms on the left-hand side of the equation are, respectively, the advection of the BGC variable by the flow field and its mixing. The right-hand side (RHS) of the equation represents the governing equations where the $\tilde{f_i}$ are the continuous analogue to f in the discrete dynamics (1). These allow the BGC variables to interact with one another and so change their concentrations in the ocean environment (e.g. Fennel *et al.*, 2006). They include processes such as photosynthesis of phytoplankton, predator–prey dynamics, conversion of living matter to organic detritus, and conversion of organic detritus to inorganic nutrients. These governing equations depend on a set of biological parameters, θ, and forcing fields, γ, such as light. Appropriate initial and boundary conditions are required. The velocity and diffusion processes are determined from a physical model based on a numerical approximation to the Navier–Stokes equations for fluid motion and, to a first approximation, are not affected by changes in the biogeochemistry.

There are a number of practical considerations imposed by the dynamic model that are relevant for BGC DA. The dynamic model (3) generally admits no closed-form, or analytic, solutions owing to its nonlinearities and complex geometry (e.g. coastlines). Instead, discretized versions of these equations are implemented as numerical models. There are a number of community ocean models that are used to facilitate application (e.g. Regional Ocean Modelling System, Haidvogel *et al.*, 2008). The discrete time-stepping model (1) can be used as a reasonable representation of the dynamic system, and the state, x_t, will contain all the prognostic variables, B_i, and their values on the spatial grid of the numerical model. Nitrogen is often the limiting nutrient in the ocean and so often acts as the model currency (rather than carbon). It is emphasized that there is no single definitive set of governing equations and that a generally accepted set of equations is chosen that approximates biological reality. Another issue is that the state dimension is large. A typical regional model of the mid-Atlantic Bight (e.g. Mattern *et al.*, 2013) has seven prognostic BGC variables, 130-by-82 grid cells in the horizontal (latitude, longitude) layer and 36 vertical (depth) layers. This yields a state vector with 2 686 320 elements. This high dimensionality is a key feature of the BGC DA problem.

Various levels of complexity for marine BGC models are possible. For the BGC dynamics, the number of compartments must be chosen, and governing equations assigned. The guiding principle is how to make the model 'fit to purpose'. That is, a scientific choice is made as to how suitable the particular dynamic model is for the application at hand. Simplifications are also made in the physical part of the model.

One may ignore spatial variation, and so (3) then corresponds to a system of coupled ordinary differential equations, which reduces the state dimension to the number of prognostic BGC variables. These 0-D models have utility as test beds to develop DA methods (Jones *et al.*, 2010; Parslow *et al.*, 2013; Weir *et al.*, 2013). Another common simplification is to consider variation only in the depth dimension. This is often a reasonable approximation because BGC dynamics are, to a first order, a vertical process rather than a horizontal one. Such 1-D models provide for reduced complexity but a high degree of realism and have been widely used, including for BGC DA studies (Mattern *et al.*, 2010). However, the ultimate target problem for BGC DA is still the fully 3-D one.

The basic formulation of BGC dynamics is usually as a deterministic model. That is, given the initial and boundary conditions, a set of parameters, and the required exogenous forcing, these equations produce a single prediction for the BGC variables. However, there are many sources of uncertainty and error, including the following:

1. The BGC parameters, θ, within any given set of governing equations, are uncertain. There are a wide range of reported values in the literature as determined from laboratory and field studies (Parslow *et al.*, 2013). There are also regional and seasonal differences in parameters (Mattern *et al.*, 2012; Toyoda *et al.*, 2013).
2. The BGC governing equations are only approximations of reality, and alternative formulations are possible for the governing equations (Friedrichs *et al.*, 2007). That is, we aggregate individuals into populations and populations into functional groups, simplify ecophysiological processes, avoid modelling certain unknown/difficult components (such as heterotrophic bacteria), and dynamically model spatial averages of point processes.
3. There are uncertainties in the ocean circulation models that drive BGC models (Toyoda *et al.*, 2013). For instance, limitations in the spatial resolution of physical models impose errors on the transport and mixing processes that affect BGC dynamics.
4. The forcing fields are uncertain. Deterministic model runs typically use physical forcing (e.g. wind fields) or biological forcing (e.g. light fields) derived directly from observations. But these are, in reality, only a single realization of a random field.

It is therefore argued here that a stochastic formulation of the dynamic model is desirable and indeed necessary. This can take the general form (1). There are various ways to incorporate stochasticity into numerical models. These are often based on generating realizations using different configurations of the model and includes ensemble methods (Gneiting and Raftery, 2005) and using stochastic BGC parameters (Dowd, 2011). However, a full incorporation of all sources of uncertainty is needed.

2.2. Observations

The observations for BGC variables, y_t, are many and varied. Traditional observations take the form of water samples at various depths, and these yield direct information on BGC variables such as nutrients and phytoplankton. Some variables, however, remain difficult to observe and interpret, for example, zooplankton and detritus. There are now some remarkable ocean BGC time series spanning multiple decades, such as the Bermuda Atlantic Time Series (Steinberg *et al.*, 2001), and public databases for phytoplankton now span a century or more (Boyce *et al.*, 2012). Ocean optics and remotely sensed imagery from satellites have revolutionized measurement of ocean chlorophyll and provided for global-scale estimates of the temporal and spatial variability of phytoplankton. New types of observation platforms and sensors for BGC variables are being developed. The main point is that there are now many data types available that provide information on BGC variables. Their sampling strategies are often complex, and an understanding of the distributional properties of their errors is important. For instance, inter-calibration experiments for satellite measurements are undertaken where water samples are compared with optical proxies, and these provide for transfer functions that map between measured quantities (e.g. the light spectrum) and BGC quantities (e.g. chlorophyll) (Hooker and McClain, 2000). We rely on the output of these complex calibration processes but make little reference to the errors. Another key issue is the errors of representativeness (Oke and Sakov, 2008). This relates to the uncertainty due to measurements that are point observations of BGC variables and the fact we use as spatial averages in estimation and dynamic modelling. In other words, it is a problem in change of support (Gotway and Young, 2002). In spite of the observation revolution underway, BGC processes still remain poorly resolved, and we rely heavily on dynamic models to help interpret them.

3. SOLUTION TECHNIQUES

The general solution for the BGC DA problem is the joint distribution of the state and parameters conditional on all available observations. That is, we wish to obtain the target, or posterior, density $[x_{1:T}, \theta | y_{1:T}]$, where $x_{1:T}$ is the state over the complete analysis period $t = 1, \ldots, T$. (Note that, for simplicity, we do not explicitly include the statistical parameters, ϕ, as their treatment is conceptually the same as for the BGC parameters θ.) The hierarchical Bayesian formulation allows us to express the posterior as

$$[x_{1:T}, \theta | y_{1:T}] \propto [y_{1:T} | x_{1:T}, \theta] \cdot [x_{1:T} | \theta] \cdot [\theta]. \tag{4}$$

The terms on the RHS side of the equation are the following:

1. $[y_{1:T} | x_{1:T}, \theta]$ is the data model and conforms to the measurement equation of the state space model (2). We are often able to write down this likelihood directly as a parametric distribution.
2. $[x_{1:T} | \theta]$ is the model state conditioned on the parameters and arises from the transition density associated with the dynamic model. There is no closed-form solution for this, and we must rely on samples, or realizations, derived from (1) for its evaluation.
3. $[\theta]$ represents the prior information on parameters, and for the BGC problem, this is based on information gleaned from the scientific literature and expert knowledge.

The decomposition in (4) provides a useful framework for viewing the problem of DA in terms of its constituent elements: data, a dynamic model, and parameters (Wikle and Berliner, 2007). There is an implicit conditioning, however, on other elements that were identified earlier as additional sources of uncertainty (i.e. the dynamic model chosen, errors in the physical model, and the environmental forcing). However, these could readily be incorporated into the hierarchical framework and are a target for future work.

For the remainder of this section, we provide an overview of current approaches being used for BGC DA for state and parameter estimations. These are all special cases of the Bayesian hierarchical model (4). A salient feature of BGC DA is its inherent nonlinearity and non-Gaussianity, and we only emphasize approaches that can treat these aspects. It is important to keep in mind that the choice of methodology for carrying out inference is often dictated by the practical goals of the BGC study in the context of the computational resources that can be brought to bear on the problem. For instance, an online prediction system may emphasize computational efficiency and approximate but fast methods for state estimation, whereas a retrospective off-line analysis may focus on comprehensive joint state and parameter estimation with less concern for computational requirements. The choice of dynamical complexity used (e.g. 0-D vs 3-D models) is dictated by similar considerations.

3.1. Variational approaches

Variational DA methods are so named because their derivation and solution relies on the calculus of variations (Bennett, 2002). They are optimization approaches that minimize a least-squares-based cost function subject to the constraint that the state estimates conform to the dynamical model. If the state is required to satisfy a deterministic dynamical model exactly, they are called strong constraint problems; otherwise, they are termed weak constraint problems.

Strong constraint variational DA is perhaps the most common approach used in BGC DA. It assumes that the model error terms, e_t, in (1) are zero, so that the dynamics embodied in $[x_{1:T}|\theta]$ becomes a deterministic map (i.e. given θ, we know $x_{1:T}$). The forcing, γ_t, is assumed known. No prior information is used, so $[\theta]$ is ignored. The observation errors, v_t, are usually assumed to be additive and to follow a zero-mean multivariate normal distribution with error covariance matrix Σ_v. The strong constraint formalism yields a nonlinear least-squares regression problem (Thompson *et al*., 2000). The goal is to minimize

$$J = \sum_{t=1}^{T} \|v_t\|^2_{\Sigma_v^{-1}} = \sum_{t=1}^{T} \|y_t - h(x_t)\|^2_{\Sigma_v^{-1}} \tag{5}$$

with respect to θ and subject to the constraint $x_t = f(x_{t-1}, \theta, \gamma_t) \ \forall \ t$. (The notation $\|a\|^2_B$ is defined as $a'Ba$.) The cost function, J, measures the weighted squared discrepancy between the observations and their dynamic model counterparts (recall h from (2)). We note that in variational DA, the 'parameter' is often the initial conditions for the dynamic model. With a parameter estimate for θ available, one can recover an estimate for associated state $x_{1:T}$. Second-order properties of the estimated θ and $x_{1:T}$ (i.e. the variances and covariances) can be obtained with the Hessian matrix. Many studies in BGC DA have used such variational approaches to estimate parameters (e.g. Matear, 1995; Vallino, 2000; Freidrichs *et al.*, 2007).

Weak constraint variational approaches, in contrast, do not assume that model errors are zero. They instead minimize

$$J = \sum_{t=1}^{T} \left(\|v_t\|^2_{\Sigma_v^{-1}} + \|e_t\|^2_{\Sigma_e^{-1}} \right) \tag{6}$$

with respect to $x_{1:T}$ and θ and so consider both observation errors, v_t, and model errors, e_t. The estimates for the state corresponds to the mode of the posterior distribution where the data model has covariance Σ_v and the state equation has covariance Σ_e. Solutions are considerably more complex than for the strong constraint problem (Bennett, 2002). Weak constraint approaches have been applied to ocean circulation models (Moore *et al.*, 2011) but have been relatively little used for BGC DA (but see, Losa *et al.*, 2004). Note that for the linear and additive Gaussian state space model, the problem (6) conforms to random-beta regression (Duncan and Horne, 1972) and is also the solution for the Kalman smoother (Jazwinski, 1970).

Strong constraint variational methods are well established for DA. Their key attribute is that they can successfully treat estimation in large-scale and nonlinear systems and have improved skill in oceanic and atmospheric predictions. However, minimization of J is challenging: the dynamic model is computationally costly, there are a large number of parameters, and there may be multiple minima. Optimization of J in (5) and (6) requires its gradient to be computed, and this in turn requires a tangent linear approximation to the dynamic model, as well as its adjoint equations. These equations, and their corresponding computer codes, are of comparable complexity to the dynamic model and hence require considerable effort to develop and maintain. However, it has been demonstrated that approximate gradients can be successfully used (Thompson *et al*., 2000). For BGC DA, an important shortcoming of strong constraint methods is their restrictive assumption of no model error. Another problematic area is that the cost functions generally take a simple least-squares form, often making only basic assumptions about Σ_v (such as it being an identity or a diagonal matrix). This ignores the effect of non-normality in both the observations and that generated by the strong nonlinearities of the dynamics. Another issue is the non-uniqueness of solutions frequently seen in variational BGC DA, which is in part due to parameter dependence (Vallino, 2000).

3.2. Sampling-based estimation

Sampling-based estimation allows us to relax the restrictive assumptions inherent in many variational approaches. The target density for the BGC DA problem is $[x_{1:T}, \theta|y_{1:T}]$. However, while (4) provides for a hierarchical Bayesian solution, it cannot easily be evaluated because

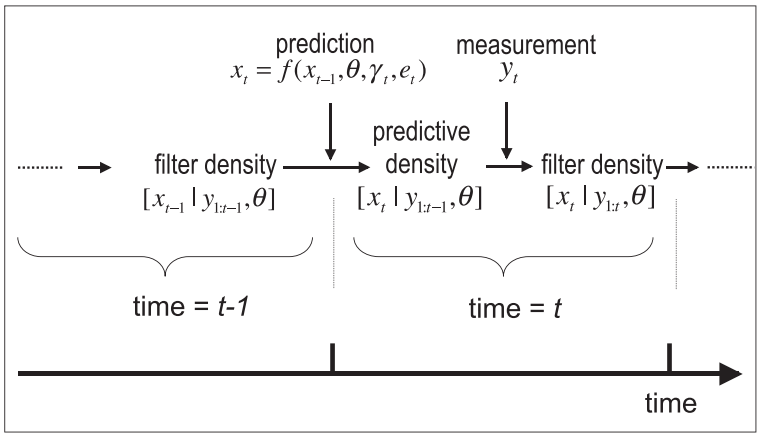

Figure 2. The data assimilation cycle, or the single state transition of the system from time $t-1$ to t. The emphasis in this figure is on sequential estimation with fixed parameters for online estimation, or filtering. See text for a detailed explanation.

there is no closed-form solution for $[x_{1:T}|\theta]$ for BGC problems. Hence, we rely on sampling-based solutions that make use of samples $\left\{x_{1:T}^{(i)}, \theta^{(i)}\right\}$ that are approximately distributed following the target posterior distribution $[x_{1:T}, \theta|y_{1:T}]$. Here, the notation $\left\{x_{1:T}^{(i)}, \theta^{(i)}\right\}$ for $i = 1, \ldots, n$ designates a sample (or ensemble) for the state and its parameters with n members; a similar notation is used throughout.

Application of standard Markov chain Monte Carlo (MCMC) can, in principle, provide for a sample-based computational solution to (4) but is computationally costly for realistic BGC DA problems. Note that the target density, $[x_{1:T}, \theta|y_{1:T}]$, corresponds to the smoothing problem where all available measurements, $y_{1:T}$, are used to determine the full state, $x_{1:T}$, and the parameters, θ. The filtering problem is also of interest for online estimation and prediction. It seeks to estimate $[x_t, \theta|y_{1:t}]$, or state and parameter estimates at time t using data up to and including time t, that is, $y_{1:t}$. Sample-based methods for filtering, or sequential Monte Carlo (SMC) methods, are in fact the foundation for addressing many Bayesian formulations of the DA problem. The goal of sample-based DA is to identify efficient and effective algorithms to draw the required samples from the target distributions, or reasonable approximations thereof. Note that the state and parameter estimation problems are interlinked but are treated separately here because the parameter estimation is often built on a foundation of state estimation.

The application of SMC methods relies on two steps: prediction and measurement update. These correspond to using the dynamical model (1) and the measurement equation (2). To explain the basic idea, we focus on state estimation for a single stage transition of our system from time $t-1$ to time t (we are fixing the parameters, θ, for now). A diagram of this DA cycle is given in Figure 2 and explained further later. Assume we have a sample, $\left\{x_{t-1|t-1}^{(i)}|\theta\right\}$, that is approximately distributed following $[x_{t-1}|y_{1:t-1}, \theta]$ and describes the system state at time $t-1$. (The notation used here for the sample emphasizes that parameters are now fixed, to distinguish it from the more general case of joint estimation of the state and parameters.) The first task is to use the dynamic model (1) to predict the state forward in time. That is, the sample members from the predictive density are formed by generating one-step-ahead realizations as

$$x_{t|t-1}^{(i)} = f\left(x_{t-1|t-1}^{(i)}, \theta, \gamma_t, e_t^{(i)}\right), \quad i = 1, \ldots, n \tag{7}$$

where $e_t^{(i)}$ is a realization of the model error. This yields a sample, $\left\{x_{t|t-1}^{(i)}|\theta\right\}$, that is an approximate draw from the predictive density, $[x_t|y_{1:t-1}, \theta]$. All SMC approaches use predictions generated in this way from a stochastic simulation.

The measurement step (also referred to as the 'analysis step' in DA) relies on Bayes' theorem to obtain the probability density for the state after measurement. That is, we want to update the predictive density, $[x_t|y_{1:t-1}, \theta]$, with the new information contained in the observation, y_t. This is performed via Bayes' theorem,

$$[x_t|y_{1:t}, \theta] \propto [y_t|x_t, y_{1:t-1}, \theta] \cdot [x_t|y_{1:t-1}, \theta] \tag{8}$$

and yields the filter density, $[x_t|y_{1:t}, \theta]$. The RHS makes use of the likelihood associated with the measurement equation (2), and the predictive density acts as a prior. For computational purposes, we seek a corresponding algorithm to update the sample from the predictive density, $\left\{x_{t|t-1}^{(i)}|\theta\right\}$, using measurements, y_t, to yield a sample, $\left\{x_{t|t}^{(i)}|\theta\right\}$, that is approximately distributed following the filter density, $[x_t|y_{1:t}, \theta]$. The candidate algorithms for this, particle filters and the ensemble Kalman filter (enKF), are discussed in the following.

3.2.1. Particle filters

Particle filters provide a means to solve the filtering problem and by extension can aid in parameter estimation (Section 3.2.3) and in MCMC (Section 3.2.4). The most basic particle filtering algorithm is sequential importance resampling (Gordon *et al.*, 1993; Kitagawa, 1996). Here, one simply performs the prediction step (7) to obtain $\left\{x_{t|t-1}^{(i)}|\theta\right\}$ (termed the proposal). The new measurement information, y_t, is then incorporated by weighted resampling of $\left\{x_{t|t-1}^{(i)}|\theta\right\}$, with the weights being proportional to the likelihood $\left[y_t|x_{t|t-1}^{(i)}, \theta\right]$. This yields a

sample, $\left\{x_{t|t}^{(i)}|\theta\right\}$, that is an approximate draw from the filter density, $[x_t\,|\,y_{1:t},\theta]$. Besides its simplicity, the primary advantage of a particle filter is that in the limit of large sample sizes, it can be shown to converge to the target (filter) density. No restrictive assumptions need be placed on the form of the model or measurement errors. Its central problem for high-dimensional applications in DA is that the computed sample is often a poor representation of the target distribution, as non-zero weights become concentrated on just a few sample members, and the filter becomes degenerate (Snyder *et al.*, 2008). More complex particle filters focus on obtaining optimal proposals that avoid such a weight collapse, such as the lookahead filter of van Leeuwen (2010). Smoothing algorithms are also built on the foundation of particle filtering (Godsill *et al.*, 2004). Applications of particle filters for BGC DA are limited and confined to simple examples (Dowd, 2006, Weir, 2013), with some notable exceptions (Mattern *et al.*, 2013).

3.2.2. *Ensemble Kalman filter*

The enKF is a widely applied, but approximate, approach for filtering (Evensen, 2009). There are many variants of the enKF available, and we outline a basic one to illustrate the main idea. Suppose that a sample from the predictive density, $\left\{x_{t|t-1}^{(i)}|\theta\right\}$, is available, say from (7). The measurement update step follows the Kalman filter updating equations. Assuming, for simplicity, that we have an additive measurement error, v_t, and that the measurement operator, h, is linear and corresponds to the matrix H, the form of the enKF measurement update is

$$x_{t|t}^{(i)} = x_{t|t-1}^{(i)} + K\left(y_t^{(i)} - Hx_{t|t-1}^{(i)}\right), \quad i = 1,\ldots,n \tag{9}$$

where the pseudo-observations are $y_t^{(i)} = y_t + v_t^{(i)}$, $i = 1,\ldots,n$ and $K = \Sigma_{t|t-1}H'\left(H\Sigma_{t|t-1}H' + \Sigma_v\right)^{-1}$. (Recall that $\Sigma_{t|t-1}$ and H are dependent on the parameters θ, which are assumed prespecified and fixed here.) It can been seen that the sample, or ensemble, is obtained by taking each sample member from the predictive density and updating it based on the discrepancy between the pseudo-observation and the prediction. This updating is controlled by the gain matrix, K, which is computed from the following: (i) the observation error covariance Σ_v (which is specified) and (ii) the prediction error covariance, $\Sigma_{t|t-1}$ (which is estimated as a sample covariance matrix using $\left\{x_{t|t-1}^{(i)}|\theta\right\}$; note that this sampling variability is ignored in the enKF).

The sample $\left\{x_{t|t}^{(i)}|\theta\right\}$ generated from (9) is meant to be a draw from the target filter density, $[x_t\,|\,y_{1:t},\theta]$. However it will, in general, only be a rough approximation because (9) only matches the Bayesian update (8) for linear models with additive Gaussian model and observation errors. However, the emphasis in enKF applications is typically only on estimating the first two moments of the state (the mean and covariance matrix), and moment corrections have been developed to improve accuracy (Anderson, 2001). Gaussian anamorphosis approaches are also being used to extend the enKF (Béal *et al.*, 2010). The main disadvantage of the enKF is the fact that it is an approximation; it will under some circumstances produce results far from the target density, and it is difficult to assess these types of approximation errors and predict when they will occur. The main advantage of the enKF is that it is straightforward to implement and has been successfully applied to many large-scale estimation problems in the geosciences and appears robust to outliers. It requires much less effort to apply than adjoint-based variational methods, while accounting for model error. Its successful implementation does, however, require some *ad hoc* tricks, such as suppression of spurious long-range correlations in $\Sigma_{t|t-1}$, also known as localization. It has been applied fairly extensively in BGC DA; early studies dealt with simple 1-D models (Eknes and Evensen, 2002), while more recent applications use fully 3-D BGC models (Hu *et al.*, 2012).

3.2.3. *Parameter estimation*

In the preceding development of SMC, no provision has been made for parameter estimation, nor for the use of prior information on parameters. There are two main approaches to parameter estimation (in the absence of priors): (i) state augmentation and (ii) likelihood methods. State augmentation (Kitagawa, 1998) is based on an expansion of the state vector to include the parameters, that is, $\tilde{x}_t = \left(x_t', \theta_t'\right)'$, and then assuming that the parameter evolves in time according some simple model, usually a random walk. This means that any method that can provide for state estimation (i.e. the particle filter and the enKF) can also be used for parameter estimation. Doron *et al.* (2013) provide a 3-D BGC example using the enKF. Note that the parameters have been considered as (artificially) time varying in this formulation. The multiple iterative filter of Ionides *et al.* (2006) extends the state augmentation procedure to allow for estimation of static, or time-invariant, parameters. For likelihood based methods, we note that the likelihood can be approximated as

$$L\left(\theta|y_{1:T}\right) \approx \prod_{t=1}^{T}\left(\frac{1}{n}\sum_{i=1}^{n}\left[y_t|x_{t|t-1}^{(i)},\theta\right]\right) \tag{10}$$

The computation of this composite likelihood relies on samples from the predictive density that can be obtained from the particle filter. Parameter estimates are determined by maximizing L. One problem is that sample-based likelihoods suffer from Monte Carlo variation and so may be difficult to maximize. The EM algorithm can also be used to estimate θ (Xu and Wikle, 2007). Another key issue for both parameter estimation approaches in BGC DA is the inherent parameter dependence (Matear, 1995; Dowd, 2011).

3.2.4. *Markov chain Monte Carlo*

Thus far, we have treated special cases of the general DA problem wherein we have ignored any prior information on the parameters. However, the fully Bayesian formulation of (4) and its hierarchical decomposition use prior information embodied in $[\theta]$, and this is the real target problem in marine BGC DA. MCMC approaches are computational algorithms that allow for the solution of Bayesian problems

(Gelman *et al.*, 2003). They generate long realizations using ergodic Markov chains that provide samples from the target distribution, in this instance the joint distribution of the parameters and the state, $[x_{1:T}, \theta | y_{1:T}]$. There are two general algorithmic classes: Gibbs sampling and the Metropolis–Hastings (M-H) algorithm (which are often combined in practice, e.g. M-H within Gibbs). Application of standard MCMC algorithms is possible for 0-D marine BGC models (Harmon and Challenor, 1997; Dowd and Meyer, 2003). However, they are computationally challenging for realistic, higher-dimensional BGC models.

Specialized MCMC algorithms that deal more effectively with marine BGC problems in the Bayesian hierarchical framework have been proposed. One approach is particle MCMC (Jones *et al.*, 2010; Andrieu *et al.*, 2010). It relies on a particle filter operating in conjunction with an M-H algorithm. The basic idea can be explained as follows. Note that in (10), we showed that for any parameter, θ, the particle filter is able to approximate the likelihood $L(\theta | y_{1:T})$. (And with a specific parameter available, one can then obtain samples from the prediction, filter and smoother densities for the state.) The elements needed for an M-H MCMC algorithm are then in place: (i) we have a user specified prior for the parameters, $[\theta]$; and (ii) we have a means to evaluate the likelihood, $L(\theta | y_{1:T})$. Our goal is to use these building blocks to determine a sample from the joint posterior of the state and parameters, $\left\{x_{1:T}^{(i)}, \theta^{(i)}\right\}$.

The M-H algorithm builds the sample via an iterative algorithm. Suppose we are at iteration i and have obtained the ith sample member for the parameter posterior, $\theta^{(i)}$. We want to determine the next sample member, $\theta^{(i+1)}$. A very simple version of the M-H algorithm proceeds by proposing a candidate parameter, θ^*, for inclusion in the sample—this candidate could be drawn from $[\theta]$ (although this is not the best choice), and with acceptance probability α, θ^* is either accepted into the sample, so $\theta^{(i+1)} = \theta^*$, or it is assumed that the value from the last iteration remains, so $\theta^{(i+1)} = \theta^{(i)}$. The M-H rule for this case uses an acceptance probability based on ratios of the likelihoods, that is, $\alpha = L(\theta^* | y_{1:T}) / L\left(\theta^{(i)} | y_{1:T}\right)$. Note that for any sample parameter, $\theta^{(i)}$, we also have an associated sample of the state, $x_{1:T}$, from the particle filter/smoother. Hence, this procedure provides a means to obtain the desired joint posterior.

The main message is that particle MCMC operates with a particle filter embedded within an MCMC algorithm so that each iteration of the MCMC algorithm requires a particle filter to be run as a means for likelihood evaluation. The number of MCMC iterations needed over θ is typically large ($\gg$1000), and each particle filter pass may require hundreds to thousands of particles to provide a useful likelihood estimate. The inherently sequential nature of MCMC algorithms also means that a 'burn-in' period must be discarded, a problem that is further multiplied when parallel chains are being run. This means that many expensive dynamic model evaluations are not being used, and the computational load is increased. Approaches combining multiple particle filters with MCMC may help alleviate this problem (Chopin *et al.*, 2013). Another major issue is ensuring that the chain mixes, in the sense that it properly explores the regions of the posterior $[x_{1:T}, \theta | y_{1:T}]$, having non-negligible probability density. For marine BGC models, this dimension can be very large, the dynamic system can have a strong memory, and mixing can be very slow.

Early application of MCMC approaches to the BGC DA problem emphasized the gains achieved by incorporating prior information on parameters (Harmon and Challenor, 1997), something not considered by variational DA. The study of Malve *et al.* (2007) extended these ideas and provided improvements in the computational methods (e.g. adaptive MCMC) for BGC DA while confronting real ocean data. The most recent studies have aimed towards an explicit Bayesian hierarchical representation. Fiechter *et al.* (2013) applied the framework to a 1-D BGC model with an emphasis of how sampling strategies affect parameter estimation. Both Hooten *et al.* (2011) and Leeds *et al.* (2014) emphasize the use of dimension reduction via emulators as a way to apply Bayesian computational approaches to realistic problems, which are otherwise rendered difficult because of the computational demands of numerical BGC models. A novel approach is offered by Leeds *et al.* (2013) who treat the BGC DA problem by using a statistical emulator in the depth dimension and linking a set of these 1-D estimates using a statistical model of the spatial field in the horizontal, thereby treating the full 3-D BGC DA problem. Such approaches that blend mechanistic and statistical models represent an important direction forward. The recent study of Parslow *et al.* (2013) applies a particle MCMC approach to 0-D BGC model and discusses in detail many of the philosophical and practical issues that must be considered for realistic application of the Bayesian hierarchical framework.

4. FUTURE DIRECTIONS AND CHALLENGES

In our exploration of the BGC DA problem and solution techniques, a number of issues were identified. Our emphasis has been on emerging statistical approaches, and less on the more widely used variational methods. Below, we summarize these and identify some challenges ahead for statistical DA for BGC models.

1. *Statistical methods for high-dimensional problems.* The state vector, x_t, that characterizes the BGC state of the ocean is of a large dimension, often $>10^6$ elements. Approaches such as variational methods and the ensemble Kalman filter have proven able to deal with this high dimensionality in practice. However, they also have restrictive assumptions or approximations built into them. The exact methods, which use sampling-based inference (i.e. particle filters and MCMC), are quite limited in the dimensionality that can be handled. The challenge is to address the larger and more realistic problems by extending sampling-based approaches to operate in much higher dimensions. Some success in this area for particle filters has been achieved by van Leeuwen (2010) by carefully treating the proposal distributions and also by Briggs *et al.* (2013) who implement a particle filter/smoother that runs sequentially through space as well as time.
2. *Dimension reduction.* The computational cost of dynamic BGC models is high owing to their large dimension, which results from their fine resolution in space and time (required mainly for numerical solution to the underlying differential equations), combined with the number of prognostic BGC variables. This means that, in practice, sampling-based inference using realistic numerical models is restricted to small ensemble sizes (<1000). Because BGC variables are highly temporally and spatially correlated, the effective degrees of freedom are much less than the state dimension. Considerable opportunity thus exists for dimension reduction. There is no particular reason to use

the computational grid of the numerical model for statistical analyses, and, for example, a spectral representation could be used (Cressie and Wikle, 2011). One promising avenue for dimension reduction is surrogates, or emulators, which are low-dimensional statistical representations of complex numerical models. These offer the potential for very large sample sizes, which is a key to making inference in any Monte Carlo-based method (Leeds *et al.*, 2014). Emulators can include a non-specific random component to capture the extra variation missed by the emulator, and this emulator error can also be built into the hierarchy of the Bayesian model of (4). Emulators can be based on a variety of approaches including neural networks (van der Merwe *et al.*, 2007), polynomial chaos (Mattern *et al.*, 2012), and Gaussian processes (Margvelashvili *et al.*, 2012). Kalman filter applications also often focus on dimension reduction, sometimes in conjunction with emulators (Frolov *et al.*, 2009). In these cases, inference is carried out using only local spatial regions (Ott *et al.*, 2004) or by operating in a reduced dimension subspace (Doron *et al.*, 2013).

3. *Incorporating priors on parameters.* While the general Bayesian framework considers prior distributions for parameters, most of the DA techniques examined do not use explicit prior information on BGC parameters. In fact, the standard state space model makes it clear that priors are only required for the statistical parameters that describe the distributions of the model error and measurement errors. It has been well demonstrated that the structural form of BGC models imposes a high degree of parameter dependence to the likelihood surface and leads to identifiability problems in estimation. This problem could be alleviated by using a considerable amount of prior information on parameters available from field and laboratory studies. We argue that this prior information can, and should, be used in BGC DA and so advocate for fully Bayesian approaches. Note that it might be necessary to define truncated priors to avoid biologically inadmissible values (or use log-transformations as in Parslow *et al.*, 2013) and to also consider the joint structure of the priors in addition to their marginals. Geographic variation must also be incorporated (Toyoda *et al.*, 2013).
4. *Dynamic model complexity and model selection.* A central question for BGC models is as follows: how complex a dynamic model do we require? The trade-off is between biological realism and understanding and computational tractability, all the while recognizing that there is not one definitive set of governing equations. Marine BGC models presently range widely in their number of compartments, and complexity has been shown not to necessarily relate to predictive skill (Friedrichs *et al.*, 2007). Dynamic marine ecological models of intermediate complexity have been advocated as a reasonable compromise (Hannah *et al.*, 2010). Because the DA problem is formulated as one of statistical estimation, we also have the possibility of considering model selection and multi-model inference. A key point is that a fully Bayesian hierarchical modelling framework allows us, at least in principle, to answer questions about model complexity rigorously and quantitatively, while accounting properly for reliance on prior knowledge.
5. *Improved stochastic simulation and quantification of prediction uncertainty*. The prediction step is an integral part of sample-based DA methods. It uses realizations generated from the stochastic dynamic system to obtain a sample (or ensemble) that characterizes the variability in the BGC system. (However, Briggs *et al.* (2013) suggest that we could use flexible multivariate distributions based on copulas to replace an ensemble-based characterization of uncertainty.) In any case, we must account for all sources of uncertainty affecting stochastic dynamic prediction. Uncertainties that are generally ignored include the following: (i) the functional form of the governing equations; (ii) the physical model that drives the BGC model; and (iii) the environmental forcing functions. This implies that in the current practice of BGC DA, the predictive sample is clearly under-dispersed. This is not just a theoretical problem: when we fail to account for a source of uncertainty, we have false confidence in our predictions.
6. *Specifying the data model.* Another neglected area in BGC DA is the treatment of measurements and their uncertainty. Many observed BGC variables are actually measured in a very indirect manner (they are often optical measurements converted to BGC quantities). The actual measurement operator is thus itself a complex mathematical model with many uncertain elements. Considerable effort by field scientists has been put into the development of these conversion models. However, they are not systematically used by the BGC DA community, who rely mainly on the final data products. Bayesian hierarchical approaches offer a formalism for bridging this gap between observationalists and modellers and treating the actual measurements more directly and with a full characterization of uncertainty. Another outstanding issue is representation error, or the uncertainty associated with using point observations to represent spatial averages. This is a ubiquitous feature of BGC DA, and we need to better understand and statistically characterize it within the data model.
7. *Sampling design*. A DA system can provide guidance on which variables to measure, and where and how often to collect them. Design for observing arrays has been termed Observation System Simulation Experiments (Sakov and Oke, 2008). These have been undertaken for physical systems in the ocean and atmosphere but have been little applied for BGC DA systems. A notable exception is Fiechter *et al.* (2013), who used a Bayesian approach to assessing sampling strategies for a BGC model. As BGC DA becomes more mature, this sampling design aspect should be developed so as to optimize the use of new and existing observation technologies.

In summary, this paper has offered an overview of marine BGC DA and current state-of-the-art approaches. Outstanding issues and promising directions were also highlighted. The development of general approaches for estimation and inference for dynamic systems has been advancing rapidly (Poole and Raftery, 2000; Cressie and Wikle, 2011; Leeds and Wikle, 2012). Novel techniques such as functional data analysis are being applied to estimation for ordinary (Cao *et al.*, 2011) and partial differential equations (Xun *et al.*, 2013). Integrated nested Laplace approximations are being used for some state space models (Ruiz-Cardenas *et al.*, 2012). MCMC is being adapted for strongly nonlinear chaotic dynamics (Wood, 2010), and methods for treating of non-Gaussianity in large-scale nonlinear systems are being investigated (Bocquet, 2010). The BGC DA problem shares many attributes with other problems in physical–statistical modelling, but there are also some unique features. We emphasize that candidate methods must be nearly always adapted to effectively treat the specific problem at hand. For BGC DA, this means accounting for structural uncertainty in the governing equations, and more effectively using prior information (Parslow *et al.*, 2013). It is our opinion that an important direction for BGC DA is the further development of sampling-based Bayesian approaches for DA in conjunction with novel approaches for dimension reduction (Leeds *et al.*, 2013). An obvious trade-off in the current practice of BGC DA is between using sophisticated statistical estimation methods and using complex dynamical models. This must,

in turn, take account whether one is pursuing online prediction (which may favour fast sequential state estimation) or doing hindcasting for scientific understanding (and so favour fully Bayesian MCMC methods for joint state and parameter estimation). One thing is clear however: progress in BGC DA will require a highly collaborative effort between statisticians and marine scientists.

Acknowledgements

The authors wish to thank all the participants of the Marine BGC DA symposium held in Hobart, Tasmania on May 2013. The symposium funding was supplied by the CSIRO office of the Chief Executive through the Cutting Edge Symposium fund; additional funds were supplied by the CSIRO Computational and Simulation Sciences transformational Capability Platform and also the Wealth from the Oceans Flagship. M. Dowd was supported by an NSERC Discovery Grant. The two anonymous reviewers are also gratefully acknowledged for their role in both shaping and improving the manuscript.

REFERENCES

Anderson JL. 2001. An ensemble adjustment Kalman filter for data assimilation. *Monthly Weather Review* **129**:2884–2903.
Andrieu C, Doucet A, Holenstein R. 2010. Particle Markov chain Monte Carlo. *Journal of the Royal Statistical Society B* **72**(3):269–342.
Béal D, Brasseur P, Brankart JM, Ourmieres Y, Verron J. 2010. Characterization of mixing errors in a coupled physical biogeochemical model of the North Atlantic: implications for nonlinear estimation using Gaussian anamorphosis. *Ocean Science* **6**(1):247–262.
Bennett A. 2002. *Inverse Modeling of the Ocean and Atmosphere*. Cambridge University Press: Cambridge, Massachusetts.
Bocquet M, Pires CA, Wu L. 2010. Beyond Gaussian statistical modeling in geophysical data assimilation. *Monthly Weather Review* **138**(8):2997–3023.
Boyce DG, Lewis M, Worm B. 2012. Integrating global chlorophyll data from 1890 to 2010. *Limnology and Oceanography: Methods* **10**:840–852.
Brasseur P, Gruber N, Barciela R, Brander K, Doron M, El Moussaoui A, Hobday AJ, Huret M, Kremeur A-S, Lehodey P, Matear R, Moulin C, Murtugudde R, Senina I, Svendsen E. 2009. Integrating biogeochemistry and ecology into ocean data assimilation systems. *Oceanography* **22**(3):206–215.
Briggs J, Dowd M, Meyer R. 2013. Data assimilation for large scale spatio-temporal systems using a location particle smoother. *Environmetrics* **24**(2):81–97.
Cao J, Wang L, Xu J. 2011. Robust estimation for ordinary differential equation models. *Biometrics* **67**(4):1305–1313.
Chopin N, Jacob PE, Papaspiliopoulos O. 2013. SMC2: an efficient algorithm for sequential analysis of state space models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75**(3):397–426.
Cressie N, Wikle CW. 2011. *Statistics for Spatio-temporal Data*. Wiley: Hoboken, New Jersey.
Doron M, Brasseur P, Brankart JM, Losa SN, Melet A. 2013. Stochastic estimation of biogeochemical parameters from Globcolour ocean color satellite data in a North Atlantic 3D ocean coupled physical–biogeochemical model. *Journal of Marine Systems* **117-118**:81–95.
Dowd M. 2006. A sequential Monte Carlo approach to marine ecological prediction. *Environmetrics* **17**:435–455.
Dowd M. 2011. Parameter estimation for nonlinear stochastic dynamic systems with application to ocean biogeochemistry. *Environmetrics* **22**(4):501–515.
Dowd M, Meyer R. 2003. A Bayesian approach to the ecosystem inverse problem. *Ecological Modelling* **168**:39–55.
Duncan DB, Horn SD. 1972. Linear dynamic recursive estimation from the viewpoint of regression analysis. *Journal of the American Statistical Society* **67**:815–821.
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.
Evensen G. 2009. *Data Assimilation: The Ensemble Kalman Filter*. Springer: Berlin.
Fennel K, Wilkin J, Levin J, Moisan J, OReilly J, Haidvogel D. 2006. Nitrogen¡ cycling in the Middle Atlantic Bight: results from a three-dimensional model and implications for the North Atlantic nitrogen budget. *Global Biogeochemical Cycles* **20**(3):GB3007, doi: 10.1029/2005GB002456.
Fiechter J, Herbei R, Leeds W, Brown J, Milliff R, Wikle C, Moore A, Powell T. 2013. A Bayesian parameter estimation method applied to a marine ecosystem model for the coastal Gulf of Alaska. *Ecological Modelling* **258**:122–133.
Follows MJ, Dutkiewicz S. 2011. Modeling diverse communities of marine microbes. *Annual Review of Marine Science* **3**:427–451.
Friedrichs MAM, Dusenberry JA, Anderson LA, Armstrong RA, Chai F, Christian JR, Doney SC, Dunne J, Fujii M, Hood R, McGillicuddy DJ, Moore JK, Schartau M, Spitz YH, Wiggert JD. 2007. Assessment of skill and portability in regional marine biogeochemical models: role of multiple planktonic groups. *Journal of Geophysical Research: Oceans* **112**(C8):1978–2012.
Frolov S, Baptista AM, Leen TK, Lu Z, van der Merwe R. 2009. Fast data assimilation using a nonlinear Kalman filter and a model surrogate: an application to the Columbia River estuary. *Dynamics of Atmospheres and Oceans* **48**(1):16–45.
Gelman A, Carlin JB, Stern HS, Rubin DB. 2003. *Bayesian Data Analysis*. CRC Press: Boca Raton, Florida.
Gentleman W. 2002. A chronology of plankton dynamics in silico: how computer models have been used to study marine ecosystems. *Hydrobiologia* **480**(1-3):69–85.
Gneiting T, Raftery AE. 2005. Weather forecasting with ensemble methods. *Science* **310**(5746):248–249.
Godsill SJ, Doucet A, West M. 2004. Monte Carlo smoothing for nonlinear time series. *Journal of the American Statistical Association* **99**(465):156–168.
Gordon N, Salmond D, Smith A. 1993. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEEE Proceedings F, Radar and Signal Processing* **140**(2):107–113.
Gotway CA, Young LJ. 2002. Combining incompatible spatial data. *Journal of the American Statistical Association* **97**(458):632–648.
Haidvogel DB, Arango H, Budgell WP, Cornuelle BD, Curchitser E, Di Lorenzo E, Fennel K, Geyer WR, Hermann AJ, Lanerolle L, Levin J, McWilliams JC, Miller AJ, Moore AM, Powell TM, Shchepetkin AF, Sherwood CR, Signell RP, Warner JC, Wilkin J. 2008. Ocean forecasting in terrain-following coordinates: formulation and skill assessment of the regional ocean modeling system. *Journal of Computational Physics* **227**(7):3595–3624.
Hannah C, Vezina A, John MS. 2010. The case for marine ecosystem models of intermediate complexity. *Progress in Oceanography* **84**(1):121–128.
Harmon R, Challenor P. 1997. A Markov chain Monte Carlo method for estimation and assimilation into models. *Ecological Modelling* **101**:41–59.
Hooker SB, McClain CR. 2000. The calibration and validation of SeaWiFS data. *Progress in Oceanography* **45**(3):427–465.
Hooten MB, Leeds WB, Fiechter J, Wikle CK. 2011. Assessing first-order emulator inference for physical parameters in nonlinear mechanistic models. *Journal of Agricultural, Biological, and Environmental Statistics* **16**(4):475–494.
Hu J, Fennel K, Mattern JP, Wilkin J. 2012. Data Assimilation with a local ensemble Kalman filter applied to a three-dimensional biological model of the Middle Atlantic Bight. *Journal of Marine Systems* **94**:145–156.
Ionides EL, Breto C, King AA. 2006. Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences* **103**:18438–18443.
Jazwinski AH. 1970. *Stochastic Processes and Filtering Theory*. Academic Press: New York.
Johnson KS, Coletti LJ, Jannasch HW, Sakamoto CM, Swift DD, Riser SC. 2013. Long-term nitrate measurements in the ocean using the in-situ ultraviolet spectrophotometer: Sensor integration into the APEX profiling float. *Journal of Atmospheric and Oceanic Technology* **30**:1854–1866.

Jones E, Parslow J, Murray L. 2010. A Bayesian approach to state and parameter estimation in a phytoplankton–zooplankton model. *Australian Meteorological and Oceanographic Journal* **59**:7–16.
Kitagawa G. 1996. Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics* **5**(1):1–25.
Kitagawa G. 1998. A self-organizing state-space model. *Journal of the American Statistical Association* **93**(443):1203–1215.
Leeds WB, Wikle CK. 2012. Science based parameterizations for dynamical spatiotemporal models. *Wiley Interdisciplinary Reviews: Computational Statistics* **4**(6):554–560.
Leeds WB, Wikle CK, Fiechter J, Brown J, Milliff RF. 2013. Modeling 3D spatio-temporal biogeochemical processes with a forest of 1D statistical emulators. *Environmetrics* **24**(1):1–12.
Leeds WB, Wikle CK, Fiechter J. 2014. Emulator-assisted reduced-rank ecological data assimilation for nonlinear multivariate dynamical spatio-temporal processes. *Statistical Methodology* **17**:126–138.
Losa SN, Kivman GA, Ryabchenko VA. 2004. Weak constraint parameter estimation for a simple ocean ecosystem model: what can we learn about the model and data? *Journal of Marine Systems* **45**(1):1–20.
Malve O, Laine M, Haario H, Kirkkala T, Sarvala J. 2007. Bayesian modelling of algal mass occurrences using adaptive MCMC methods with a lake water quality model. *Environmental Modelling & Software* **22**(7):966–977.
Margvelashvili N, Campbell EP. 2012. Sequential data assimilation in fine-resolution models using error-subspace emulators: theory and preliminary evaluation. *Journal of Marine Systems* **90**(1):13–22.
Matear RJ. 1995. Parameter optimization and analysis of ecosystem models using simulated annealing: a case study at station P. *Journal of Marine Research* **53**(4):571–607.
Matear RJ, Jones E. 2011. Marine biogeochemical modelling and data assimilation. *In: Operational Oceanography in the 21st Century.* Springer. 2011, *295–317*.
Mattern JP, Dowd M, Fennel K. 2010. Sequential data assimilation applied to a physical–biological model for the Bermuda Atlantic Time Series station. *Journal of Marine Systems* **79**:144–156.
Mattern JP, Fennel K, Dowd M. 2012. Estimating time-dependent parameters for a biological ocean model using an emulator approach. *Journal of Marine Systems* **96-97**:32–48.
Mattern JP, Dowd M, Fennel K. 2013. Particle filter-based data assimilation for a 3-dimensional biological ocean model and satellite observations. *Journal of Geophysical Research: Oceans* **118**(5):2746–2760.
Moore AM, Arango HG, Broquet G, Powell BS, Weaver AT, Zavala-Garay J. 2011. The Regional Ocean Modeling System (ROMS) 4-dimensional variational data assimilation systems: part I system overview and formulation. *Progress in Oceanography* **91**(1):34–49.
Oke PR, Sakov P. 2008. Representation error of oceanic observations for data assimilation. *Journal of Atmospheric and Oceanic Technology* **25**:1004–1017.
Ott E, Hunt BR, Szunyogh I, Zimin AV, Kostelich EJ, Corazza M, Kalnay E, Patil DJ, Yorke JA. 2004. A local ensemble Kalman filter for atmospheric data assimilation. *Tellus A* **56**(5):415–428.
Parslow J, Cressie N, Campbell EP, Jones E, Murray L. 2013. Bayesian learning and predictability in a stochastic nonlinear dynamical model. *Ecological Applications* **23**(4):679–698.
Poole D, Raftery AE. 2000. Inference for deterministic simulation models: the Bayesian melding approach. *Journal of the American Statistical Association* **95**(452):1244–1255.
Royle JA, Berliner LM, Wikle CK, Milliff R. 1999. A hierarchical spatial model for constructing wind fields from scatterometer data in the Labrador Sea. In *Case Studies in Bayesian Statistics*, Gatsonis G, Kass RE, Carlin B, Carriquiry A, Gelman A, Verdinelli I, West M (eds). Springer: New York; 367–382.
Ruiz-Cardenas R, Krainski ET, Rue H. 2012. Direct fitting of dynamic models using integrated nested Laplace approximations, INLA. *Computational Statistics & Data Analysis* **56**(6):1808–1828.
Sakov P, Oke PR. 2008. Objective array design: application to the tropical Indian Ocean. *Journal of Atmospheric and Oceanic Technology* **25**:794–807.
Schofield O, Glenn SM, Moline MA, Oliver M, Irwin A, Chao Y, Arrott M. 2013. Ocean observatories and information: building a global ocean observing network. In *Earth System Monitoring*, Orcutt J (ed.). Springer: New York; 319–336.
Snyder C, Bengtsson T, Bickel P, Anderson J. 2008. Obstacles to high-dimensional particle filtering. *Monthly Weather Review* **136**(12):4629–4640.
Steinberg D, Carlson C, Bates N, Johnson R, Michaels A, Knap A. 2001. Overview of the US JGOFS Bermuda Atlantic Time-series Study (BATS): a decade-scale look at ocean biology and biogeochemistry. *Deep-Sea Research Part II* **48**(8-9):1405–1447.
Thompson KR, Dowd M, Lu Y, Smith B. 2000. Oceanographic data assimilation and regression analysis. *Environmetrics* **11**:183–196.
Toyoda T, Awajib T, Masuda S, Sugiura N, Igarashi H, Sasaki Y, Hiyoshi Y, Ishikawa Y, Saitoh S-I, Yoon S, In T, Kishi MJ. 2013. Improved state estimations of lower trophic ecosystems in the global ocean based on a Greens function approach. *Progress in Oceanography* **119**:90–107.
Vallino JJ. 2000. Improving marine ecosystem models: use of data assimilation and mesocosm experiments. *Journal of Marine Research* **58**:117–164.
van der Merwe R, Leen TK, Lu Z, Frolov S, Baptista AM. 2007. Fast neural network surrogates for very high dimensional physics-based models in computational oceanography. *Neural Networks* **20**(4):462–478.
van Leeuwen PJ. 2010. Nonlinear data assimilation in geosciences: an extremely efficient particle filter. *Quarterly Journal of the Royal Meteorological Society* **136**(653):1991–1999.
Weir B, Miller RN, Spitz YH. 2013. Implicit estimation of ecological model parameters. *Bulletin of Mathematical Biology* **75**(2):I223–257.
Wikle CK, Milliff RF, Nychka D, Berliner LM. 2001. Spatiotemporal hierarchical Bayesian modeling tropical ocean surface winds. *Journal of the American Statistical Association* **96**(454):382–397.
Wikle CK, Berliner LM. 2007. A Bayesian tutorial for data assimilation. *Physica D* **230**:1–16.
Wood SN. 2010. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature* **466**(7310):1102–1104.
Woods JD. 2005. The Lagrangian ensemble metamodel for simulating plankton ecosystems. *Progress in Oceanography* **67**(1):84–159.
Xu K, Wikle CK. 2007. Estimation of parameterized spatio-temporal dynamic models. *Journal of Statistical Planning and Inference* **137**(2):567–588.
Xun X, Cao J, Mallick B, Carroll RJ, Maity A. 2013. Parameter estimation of partial differential equation models. *Journal of the American Statistical Association* **503**:1009–1020.

