

University of Wollongong

## Research Online

---

Faculty of Engineering and Information  
Sciences - Papers: Part A

Faculty of Engineering and Information  
Sciences

---

1-1-2013

### Bayesian learning and predictability in a stochastic nonlinear dynamical model

John Parslow

*Csiro, Marine And Atmospheric Research*

Noel Cressie

*University of Wollongong, [ncressie@uow.edu.au](mailto:ncressie@uow.edu.au)*

Edward P. Campbell

*Csiro, Mathematics, Informatics & Statistics*

Emlyn Jones

*Csiro, Marine And Atmospheric Research*

Lawrence Murray

*Csiro, Mathematics, Informatics & Statistics*

Follow this and additional works at: <https://ro.uow.edu.au/eispapers>



Part of the [Engineering Commons](#), and the [Science and Technology Studies Commons](#)

---

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: [research-pubs@uow.edu.au](mailto:research-pubs@uow.edu.au)

---

# Bayesian learning and predictability in a stochastic nonlinear dynamical model

## Abstract

Bayesian inference methods are applied within a Bayesian hierarchical modelling framework to the problems of joint state and parameter estimation, and of state forecasting. We explore and demonstrate the ideas in the context of a simple nonlinear marine biogeochemical model. A novel approach is proposed to the formulation of the stochastic process model, in which ecophysiological properties of plankton communities are represented by autoregressive stochastic processes. This approach captures the effects of changes in plankton communities over time, and it allows the incorporation of literature metadata on individual species into prior distributions for process model parameters. The approach is applied to a case study at Ocean Station Papa, using Particle Markov chain Monte Carlo computational techniques. The results suggest that, by drawing on objective prior information, it is possible to extract useful information about model state and a subset of parameters, and even to make useful long-term forecasts, based on sparse and noisy observations.

## Keywords

stochastic, nonlinear, bayesian, learning, dynamical, predictability, model

## Disciplines

Engineering | Science and Technology Studies

## Publication Details

Parslow, J., Cressie, N., Campbell, E. P., Jones, E. & Murray, L. (2013). Bayesian learning and predictability in a stochastic nonlinear dynamical model. *Ecological Applications*, 23 (4), 679-698.

# Bayesian learning and predictability in a stochastic nonlinear dynamical model

JOHN PARSLow,<sup>1</sup> NOEL CRESSIE,<sup>2</sup> EDWARD P. CAMPBELL,<sup>3,4</sup> EMLYN JONES,<sup>1</sup> AND LAWRENCE MURRAY<sup>3</sup>

<sup>1</sup>*CSIRO Computational and Simulation Science, Marine and Atmospheric Research, GPO Box 1538, Hobart, Tasmania 7001 Australia*

<sup>2</sup>*National Institute of Applied Statistics Research Australia, School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, NSW 2522 Australia*

<sup>3</sup>*CSIRO Computational and Simulation Science, Mathematics, Informatics and Statistics, Private Bag 5, Western Australia 6913 Australia*

**Abstract.** Bayesian inference methods are applied within a Bayesian hierarchical modeling framework to the problems of joint state and parameter estimation, and of state forecasting. We explore and demonstrate the ideas in the context of a simple nonlinear marine biogeochemical model. A novel approach is proposed to the formulation of the stochastic process model, in which ecophysiological properties of plankton communities are represented by autoregressive stochastic processes. This approach captures the effects of changes in plankton communities over time, and it allows the incorporation of literature metadata on individual species into prior distributions for process model parameters. The approach is applied to a case study at Ocean Station Papa, using particle Markov chain Monte Carlo computational techniques. The results suggest that, by drawing on objective prior information, it is possible to extract useful information about model state and a subset of parameters, and even to make useful long-term forecasts, based on sparse and noisy observations.

**Key words:** *Bayesian hierarchical modeling; data model; inference in nonlinear models; parameter (prior) model; prediction; stochastic process model; uncertainty.*

## INTRODUCTION

The last century has seen major advances in the ecological and earth sciences, both in the development of theoretical understanding, encapsulated in mechanistic process models, and in the development of sophisticated statistical theories and models for the interpretation and analysis of observations. However, as Berliner (2003) has pointed out, until recently the development of process models and the statistical analysis of observations have occurred in parallel and somewhat at arms length. Over the last two decades, there has been increasing effort devoted to the integration of observations and process models, so that model–data comparison and data assimilation are now key research topics.

There are a number of drivers for this increased emphasis on the integration of models and observations. The scientific community increasingly insists on the use of more objective and quantitative measures or metrics to evaluate model predictions against observations (e.g., Allen et al. 2007). But ecological and earth-system models are increasingly used for practical purposes, from short-term environmental forecasting to local issues of pollution, conservation, and renewable re-

sources, to global issues of climate change. Users of model outputs would like more accurate predictions and increasingly demand formal assessments of the uncertainty in model predictions, to inform decision-making and risk management.

Techniques for the integration of models and observations are intended to quantify model performance and allow intercomparison of alternative models, to improve performance or skill in model predictions, and to provide error estimates or confidence/credible intervals around those predictions. Errors enter into an integrated model–data system from at least three sources. First, there are errors in the process of making observations, which typically provide a distorted and/or fragmented glimpse of the underlying reality. One consequence is that we do not know the exact state of the system when we initialize dynamic models. Second, process models make simplifying assumptions and approximations, so that model simulations cannot be expected to reproduce reality exactly. Many ecological and earth system models are dynamic models, predicting the evolution of system trajectories over time, and model errors are typically stochastic, leading to divergence of simulated trajectories over time. Finally, process models typically incorporate a number of parameters, assumed to be constant over time, whose values are uncertain.

The term “data assimilation” has been used broadly to describe model–data integration (e.g., Gregg 2008,

Manuscript received 27 February 2012; revised 23 August 2012; accepted 29 August 2012. Corresponding Editor: D. S. Schimel.

<sup>4</sup> Corresponding author. E-mail: eddy.campbell@csiro.au

Luo et al. 2011). In practice, approaches and applications have tended to fall into one of two categories. In the first, attention has focused on the estimation of uncertain parameters in deterministic process models (e.g., Matear 1995). Parameters are often estimated by minimizing some kind of cost function based on model–data mismatches, typically a sum of squared errors. In some cases, the cost function is constructed and interpreted as a negative log-likelihood based on a formal error model but, in other cases, the cost function is ad hoc. The second class of applications typically involves short-term environmental forecasting or hindcasting, where errors are believed to be dominated by uncertainty about the true value of the system state. Sequential data assimilation techniques are used to update estimates of the state based on current or recent observations. In these approaches, there tends to be a strong emphasis on building realistic observation models, while the stochastic model error is often modeled as simple additive white noise and adjusted to achieve convergence of the assimilation procedure. Very sophisticated data assimilation schemes are now widely adopted and routinely used in weather and ocean forecasting.

The last decade especially has seen increasing advocacy of Bayesian approaches to data assimilation (e.g., Link et al. 2001, Berliner 2003, Calder et al. 2003, Cressie et al. 2009, Zobitz et al. 2011). Bayesian methods typically yield posterior distributions for the inferred state and parameters, most often summarized using large samples from these distributions. These can be particularly useful in applied contexts, where users may be interested in the probability distribution of performance measures derived from model predictions. A key attraction of the Bayesian approach is its ability to formally incorporate prior information about models and parameters. Given that the rationale for using mechanistic, process-based models is that they build on prior scientific knowledge about the structure and function of system components, it makes sense to use methods that allow this knowledge to be formally represented in model–data comparisons. It is of course possible to use the Bayesian formalism, while discounting or ignoring prior information, through uninformative priors or empirical Bayes methods. In these cases, Bayesian methods can generally be shown to be equivalent to classical methods (e.g., Ver Hoef 1996, Cressie et al. 2009).

Within the broader Bayesian tradition, Bayesian hierarchical modeling (BHM) offers a particularly attractive framework for the integration of mechanistic process models and observations. BHM provides a consistent, formal probabilistic framework combining error or uncertainty in model parameters, model state, model processes, and observations (Berliner 2003, Wikle 2003, Cressie et al. 2009). This framework encourages the modeler to think carefully and systematically about the approximations and assumptions involved in process

model formulation, about the observation process and the relationship between model state variables and observations, and about the relationship between model parameters and independent prior knowledge. One can think of BHM not just as an integration of models and data, but as a deep integration of mechanistic and statistical modeling; Berliner (2003) describes this as “physical-statistical” modeling.

The last decade has seen a rapid growth of Bayesian applications in ecology and the earth sciences, ranging from population dynamics and dispersal (e.g., Link et al. 2001, Calder et al. 2003, Wikle 2003, Clark and Bjornstaad 2004, Clark and Gelfand 2006, Barber and Gelfand 2007, Hooten et al. 2007) to plant ecology and terrestrial surface fluxes (e.g., Ogle et al. 2004, Baker et al. 2006, Sacks et al. 2006, Xu et al. 2006, Zobitz et al. 2007, 2008) to ocean circulation and climate (e.g., Berliner et al. 2000, Berliner 2003). Encouragingly, Bayesian approaches are now widely and successfully used for stock assessment and fisheries management (Maunder 2004).

In this paper, we focus on the application of Bayesian methods, specifically BHM, to aquatic biogeochemical (BGC)/ecological models. Model–data integration in this field has paralleled the broader trajectory we have outlined. Earlier studies focused on the problem of parameter estimation in deterministic models (Matear 1995). Over the last decade, and following developments in data assimilation into physical ocean circulation models, there has been considerable progress in implementing sequential data assimilation techniques for state estimation in three-dimensional biogeochemical models (Gregg 2008). Examples of Bayesian approaches in this area fall into two streams. The first uses a Bayesian approach to obtain posteriors for parameters and state estimation in (effectively) deterministic eutrophication models (Arhonditsis et al. 2007, 2008, Zhang and Arhonditsis 2009). The second, in contrast, uses sequential Bayesian assimilation to obtain posteriors for current and forecast state in stochastic models in which the underlying parameters are assumed constant and known (Dowd and Meyer 2003, Dowd 2006, 2007). More recently, Dowd (2011) has extended this work to obtain joint posteriors for the state and a subset of parameters. These examples all embed the ecological dynamics physically within a zero-dimensional box model setting, but Mattern et al. (2010) extend this to a one-dimensional setting.

The study presented here aims to build on previous work by using the BHM probabilistic framework to underpin enhancements in several areas:

- 1) The process models used here include stochastic errors in a way that accounts for key simplifying approximations made in replacing communities of species by a single biomass variable. These approximations are widely used in ecological and biogeo-

chemical models, and the approach seems likely to find broader application.

- 2) Our approach also allows prior distributions for model parameters to be more directly and objectively related to prior information obtained from field and laboratory studies, and from in literature meta-data. This prior information makes a valuable contribution to state estimation and forecasting in the application considered here, where observations are severely limited.
- 3) The process model has been modified to include a diagnostic variable, chlorophyll *a* (chl *a*), to support a simpler and more rigorous observation model.
- 4) Bayesian inference in nonlinear problems is generally analytically intractable, and computationally intensive simulation-based methods, such as Markov chain Monte Carlo, are used to obtain large random samples from the posterior. Our study exploits new methods for Bayesian inference (Andrieu et al. 2010) to derive a joint posterior for parameters and state in nonlinear dynamical models. This allows us to simultaneously address problems of parameter estimation, state estimation, short-term forecasting, and long-term projections in a unified probabilistic framework.

The remainder of this paper is organized as follows. In *General methodology*, we provide a brief introduction to BHM and its application to dynamical state-space models. *Reformulating a marine BGC model as a BHM* presents a reformulation of a conventional deterministic model as a stochastic process model within the BHM framework. Uncertainty in the parameters is captured through a collection of time-varying stochastic processes. In *Learning and predictability given observations*, we provide a case study of this generic model applied to a time series of observations at Ocean Station Papa. Bayesian inference procedures are used to extract information in the form of posteriors for state and parameters from a set of observations that are sparse and patchy in time, and include only a subset of state variables. Twin experiments are used to test the performance and consistency of the inference procedures, and to draw some preliminary conclusions about the effect of observation intensity on posteriors. *Discussion and conclusions* discusses the results obtained in the context of the enhancements listed above, and we make some observations about the strengths and weaknesses of this approach for marine biogeochemical modeling, and ecological modeling more broadly. Links are included to mathematical, statistical, and computing appendices.

#### GENERAL METHODOLOGY

##### *Bayesian hierarchical models (BHMs)*

The physical-statistical models described by Berliner (2003), formulated as BHMs, are models that explicitly represent three sources of uncertainty: (1) the data

model expresses uncertainty arising from observations subject to measurement error and bias; (2) the process model expresses uncertainty arising from scientific (here, biophysical) processes that are not completely understood or are approximated; (3) the parameter (prior) model expresses uncertainty arising from parameters not known exactly.

BHMs are probabilistic models, constructed from conditional probability distributions. The data are treated as conditional on the process and some parameters, and the process is treated as conditional on other parameters. Hence, the three components, data, processes, and parameters, can be thought of as hierarchical levels in a chain of conditional dependence, which we now formalize.

Let the data (observations), process(es), and parameters be represented by the vectors  $\mathbf{Y}$ ,  $\mathbf{W}$ , and  $\boldsymbol{\theta}$ , respectively. In some models, the process has a continuous index in time or space; for the purpose of computations it is enough to consider  $\mathbf{W}$  as a high-dimensional vector. The joint uncertainty is denoted  $[\mathbf{Y}, \mathbf{W}, \boldsymbol{\theta}]$ , where the notation  $[\mathbf{A}]$  represents “the probability distribution of  $\mathbf{A}$ .” It makes sense to partition the parameters into biophysical parameters and so-called statistical parameters arising from the observation process. Therefore, we write  $\boldsymbol{\theta} = \{\boldsymbol{\theta}_Y, \boldsymbol{\theta}_W\}$ .

Applying the rules of conditional probability, we can factorize the joint probability distribution as

$$[\mathbf{Y}, \mathbf{W}, \boldsymbol{\theta}] = [\mathbf{Y} | \mathbf{W}, \boldsymbol{\theta}_Y, \boldsymbol{\theta}_W][\mathbf{W}, \boldsymbol{\theta}_Y, \boldsymbol{\theta}_W] \quad (1)$$

where  $[\mathbf{A} | \mathbf{B}]$  denotes “the conditional probability of  $\mathbf{A}$  given  $\mathbf{B}$ .” Repeating this for the second component of Eq. 1, we find

$$[\mathbf{Y}, \mathbf{W}, \boldsymbol{\theta}] = [\mathbf{Y} | \mathbf{W}, \boldsymbol{\theta}_Y, \boldsymbol{\theta}_W][\mathbf{W} | \boldsymbol{\theta}_Y, \boldsymbol{\theta}_W][\boldsymbol{\theta}_Y, \boldsymbol{\theta}_W]. \quad (2)$$

The components of Eq. 2 may be simplified a little by noting that the biophysical parameters,  $\boldsymbol{\theta}_W$ , are not needed in the data model when we also condition on the process; similarly, the statistical parameters,  $\boldsymbol{\theta}_Y$ , are not needed in the second component when we also condition on the biophysical parameters. Hence, we obtain

$$[\mathbf{Y}, \mathbf{W}, \boldsymbol{\theta}] = [\mathbf{Y} | \mathbf{W}, \boldsymbol{\theta}_Y][\mathbf{W} | \boldsymbol{\theta}_W][\boldsymbol{\theta}_Y, \boldsymbol{\theta}_W]. \quad (3)$$

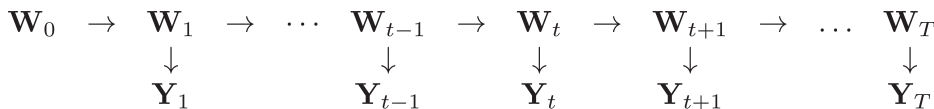
We see that the three probability distributions on the right-hand side correspond to the BHM hierarchy of sources of uncertainty identified above, representing a data model, a (stochastic) biophysical process model, and a parameter model, respectively. The parameter model is often referred to as the *prior distribution*.

Of key interest is how one can make inferences about the unobserved process state  $\mathbf{W}$  and the parameters  $\boldsymbol{\theta}$ , given the observations  $\mathbf{Y}$  on the biogeochemical process. Appealing to Bayes’ Theorem (e.g., Cox and Hinkley 1986:365–367), we may write

$$[\mathbf{W}, \boldsymbol{\theta} | \mathbf{Y}] \propto [\mathbf{Y} | \mathbf{W}, \boldsymbol{\theta}_Y][\mathbf{W} | \boldsymbol{\theta}_W][\boldsymbol{\theta}_Y, \boldsymbol{\theta}_W] \quad (4)$$

where the constant of proportionality is a function of  $\mathbf{Y}$



FIG. 1. A graphical representation for the process  $\mathbf{W}$  and observations  $\mathbf{Y}$ .

only and guarantees that the right-hand side of statement 4 is a proper joint probability distribution. This so-called *posterior distribution* is proportional to the product of the three levels of the BHM (data model, process model, parameter model) that we have developed above. We return later in this section to the issue of making inferences based on statement 4.

The use of the three levels of conditional probability models via Bayes' Theorem to learn from data is precisely the BHM framework we alluded to at the beginning of this section. Examples of its use have been growing in the last decade. It was introduced in a climate-modeling and climate-prediction context by Berliner et al. (2000), in an introductory geophysical context by Berliner (2003), and in an ecological context by Wikle (2003); see also the review by Cressie et al. (2009).

#### A state-space representation

We are interested here in the application of BHM to dynamical systems, in which the state evolves as a function of time (discrete or continuous), and the data are collected by sampling (potentially irregularly and coarsely) in time, while the process evolves at a relatively fine time step. We write the time-evolving process  $\mathbf{W}$  as  $(\mathbf{W}_0, \mathbf{W}_1, \dots, \mathbf{W}_T)$  with corresponding observations  $(\mathbf{Y}_1, \dots, \mathbf{Y}_T)$  taken after the initial value of the process  $\mathbf{W}_0$ . We use subscript  $t$  to index time, such that  $\mathbf{W}_t$  is coincident with  $\mathbf{Y}_t$ , for  $t = 1, \dots, T$ . A graphical depiction of the dependencies is shown in Fig. 1.

We remark that, in practice, observations will be missing at some times, which the BHM framework can readily handle.

We henceforth assume that the forward evolution of the process  $\mathbf{W}$  depends only on the current state; that is,  $\mathbf{W}$  is a Markov-process model described by  $[\mathbf{W}_t | \mathbf{W}_{t-1}, \boldsymbol{\theta}_W]$ , for  $t = 1, \dots, T$ . This form of conditional independence implies that  $[\mathbf{W} | \boldsymbol{\theta}_W] = \prod_{t=1}^T [\mathbf{W}_t | \mathbf{W}_{t-1}, \boldsymbol{\theta}_W]$ . Further, observations at time  $t$  are assumed to be independent of observations at other times, conditional on the state  $\mathbf{W}_t$ . Thus, the data model has the form,  $[\mathbf{Y} | \mathbf{W}, \boldsymbol{\theta}_Y] = \prod_{t=1}^T [\mathbf{Y}_t | \mathbf{W}_t, \boldsymbol{\theta}_Y]$ .

#### Statistical inference

The focus of our statistical inference is the calculation of the posterior distribution described by Eq. 4, which is rarely amenable to analytic solutions. As a result, modern Bayesian inference has harnessed efficient algorithms deployed on contemporary computing architectures to simulate samples from the posterior distribution. Statistics calculated for these samples, such as

means and quantiles, can be shown to converge to the appropriate quantities for the posterior distribution (Tierney 1994).

Suppose for instance that we are interested in estimating some function  $f(\cdot)$  of the state and parameters. We obtain a simulated sample  $\{(\mathbf{W}^{(\ell)}, \boldsymbol{\theta}^{(\ell)}) : \ell = 1, \dots, L\}$  from the posterior distribution  $[\mathbf{W}, \boldsymbol{\theta} | \mathbf{Y}]$ , and we use the transformed sample  $\{f(\mathbf{W}^{(\ell)}, \boldsymbol{\theta}^{(\ell)}) : \ell = 1, \dots, L\}$  to calculate summary statistics. For example, we can produce an estimate of its expectation, denoted  $\hat{E}$ , as

$$\hat{E}(f(\mathbf{W}, \boldsymbol{\theta}) | \mathbf{Y}) \equiv (1/L) \sum_{\ell=1}^L f(\mathbf{W}^{(\ell)}, \boldsymbol{\theta}^{(\ell)})$$

so sampling from the posterior distribution over states and parameters is key to the success of Bayesian hierarchical modeling in this context. The computational approach adopted must also be able to cope with the nonlinear behavior of the process model, noting that the state transition density function is not available in closed form.

Particle Markov chain Monte Carlo (PMCMC) was developed for exactly this situation, and so we have applied it in our case study. In particular, we use the particle marginal Metropolis-Hastings (PMMH) sampler (Andrieu et al. 2010), which we have previously applied successfully to a simple Lotka-Volterra type model (Jones et al. 2010). Details of PMMH are given in Appendix C.

#### REFORMULATING A MARINE BGC MODEL AS A BHM

A general description of the BHM framework and its use for scientific inference was given in *General methodology*. We now show how these ideas can be applied in a marine BGC setting.

#### The process model

Recall from *General methodology* that the biogeochemical process model is at the second level of the BHM hierarchy. We present the model first in terms of a deterministic model, and then we derive a stochastic version of it.

*A deterministic biogeochemical process model.*—One of the advantages of the BHM framework is that it allows us to build on existing scientific understanding, typically incorporated in deterministic process models. We can draw here on a long and rich history of (deterministic) marine BGC models that describe the cycling of nutrients (e.g., nitrogen) and/or carbon through living and nonliving organic and inorganic compartments, in simplified marine ecosystems. Open-ocean models typically deal only with pelagic planktonic systems, while

coastal models may deal with coupled pelagic-benthic systems. In this article, we deal with the simpler case of pelagic models.

In the general case, the state variables in marine BGC models are expressed as component concentrations (mass per unit volume) as functions of space  $\mathbf{x}$  and time  $t$ . These components are subject to physical transport (advection and mixing), as well as local biological and chemical reactions. If  $\mathbf{c}(\mathbf{x}, t)$  is a vector of state variables, we can write the general reaction-transport equation as

$$\frac{\partial \mathbf{c}}{\partial t} = \mathbf{R}(\mathbf{c}, \mathbf{x}, t) + \mathbf{T}(\mathbf{c}, \mathbf{x}, t) \quad (5)$$

where  $\mathbf{R}$  represents local biological and chemical reactions and  $\mathbf{T}$  is a transport operator; see Appendix A: *The transport operator* for the specific form of Eq. 5 used in the case study in *Learning and predictability given observations*. In this paper, we consider the highly simplified physical setting of a mixed-layer one-box model and, for the moment, we ignore the transport operator and focus on the local reactions  $\mathbf{R}$ . This setting allows us to formulate a BHM most clearly. However, we do include a simple transport term to account for vertical mixing in the case study and this is presented in Appendix A: *The transport operator*.

Pelagic planktonic ecosystems are complex systems that involve many species of phytoplankton and zooplankton, multiple (potentially limiting) nutrients, and dissolved and particulate organic matter pools comprised of complex mixtures. All models of these systems require simplifying approximations, and the level of detail varies across models and depends on the purpose of the model. Model detail and complexity have tended to increase over the last decade, as scientific understanding and computational power have increased. However, this in turn has led to concern about the identifiability of complex models with many uncertain parameters (Hood et al. 2006).

We have chosen a relatively simple, classic NPZD model formulation, which represents the cycling of a limiting nutrient (nitrogen) through four compartments: dissolved inorganic nitrogen or DIN ( $N$ ), phytoplankton nitrogen ( $P$ ), zooplankton nitrogen ( $Z$ ), and detrital nitrogen ( $D$ ). We can write the equations for the local rate of change of the state variables as

$$\frac{dP}{dt} = g \times P - gr \times Z \quad (6)$$

$$\frac{dZ}{dt} = E_Z \times gr \times Z - m \times Z \quad (7)$$

$$\frac{dD}{dt} = (1 - E_Z) \times f_D \times gr \times Z + m \times Z - r \times D \quad (8)$$

$$\frac{dN}{dt} = -g \times P + (1 - E_Z)(1 - f_D) \times gr \times Z + r \times D. \quad (9)$$

Notice that  $dP/dt + dZ/dt + dD/dt + dN/dt = 0$ , which is a

consequence of “mass balance” in the currency of nitrogen. In Eqs. 6–9,  $g$  is the phytoplankton-specific growth rate (per day, or  $\text{d}^{-1}$ ),  $gr$  is the zooplankton-specific grazing rate ( $\text{mg } P \text{ grazed} \cdot \text{mg } Z^{-1} \cdot \text{d}^{-1}$ ),  $m$  is the zooplankton-specific mortality rate ( $\text{d}^{-1}$ ), and  $r$  is the specific breakdown rate of detritus ( $\text{d}^{-1}$ ). A fraction  $E_Z$  of zooplankton ingestion is converted to zooplankton growth and, of the remainder, a fraction  $f_D$  is allocated to detritus, with the rest released as dissolved inorganic nitrogen,  $N$ . The fractions,  $E_Z$  and  $f_D$ , are treated as constant, independent of ingestion rates. This is a common simplifying assumption in biogeochemical models (e.g., Wild-Allen et al. 2010).

The process rates  $g$ ,  $gr$ ,  $m$ , and  $r$  are all functions of state variables and/or exogenous forcing variables, and hence they are functions of time. As we shall see below, a multiplicative temperature correction  $T_c$  is applied to all rate processes; to define  $T_c$ , we use a so-called “ $Q_{10}$  formulation” for dependence on temperature  $T$ :

$$T_c = Q_{10}^{(T - T_{\text{ref}})/10}. \quad (10)$$

Notice that  $T$  depends on time and, hence, so does  $T_c$ , where  $T_{\text{ref}}$  is a reference temperature, and  $Q_{10}$  is a prescribed parameter.

We use a flexible formulation for the dependence of zooplankton’s grazing rate on phytoplankton concentration (zooplankton functional response):

$$gr = \frac{T_c \times I_Z \times A^v}{(1 + A^v)} \quad (11)$$

where  $v$  is a given power; the relative availability of phytoplankton  $A$  is

$$A = \frac{Cl_Z \times P}{I_Z} \quad (12)$$

where  $A$  depends on time because  $P$  does. In Eq. 12,  $I_Z$  is the maximum zooplankton ingestion rate ( $\text{mg } P \cdot \text{mg } Z^{-1} \cdot \text{d}^{-1}$ ); and  $Cl_Z$  is the maximum clearance rate ( $\text{m}^3 \text{ swept clear} \cdot \text{mg } Z^{-1} \cdot \text{d}^{-1}$ ). Both are constant in the deterministic formulation. This is a standard rectangular hyperbola or Type-2 functional response (Holling 1965) when  $v = 1$ , and a Type-3 sigmoid functional response when  $v > 1$ .

We follow Steele (1976) and Steele and Henderson (1992) in adopting a time-dependent quadratic formulation for zooplankton mortality:

$$m = T_c \times m_Q \times Z \quad (13)$$

where the constant quadratic mortality rate  $m_Q$  has units of  $(\text{mg } Z/\text{m}^3)^{-1} \cdot \text{d}^{-1}$ . The detrital remineralization rate is assumed to depend only on temperature (which is time dependent):

$$r = T_c \times r_D \quad (14)$$

where the constant parameter  $r_D$  prescribes the rate at the reference temperature and has units of  $\text{d}^{-1}$ .

Finally, the phytoplankton specific growth rate  $g$  depends on temperature  $T$ , available light or irradiance  $E$  (see Appendix A: *The light model*) and dissolved inorganic nitrogen  $N$ . The submodel given below for  $g$  is somewhat more elaborate than the submodels used for the other rate processes. We shall see that it predicts changes in phytoplankton composition (nitrogen:carbon ratio and chlorophyll  $a$ :carbon ratio) as well as the phytoplankton specific growth rate, as phytoplankton adapt to changes in available light and nutrients.

In the BHM framework, we are encouraged to pay careful attention to the relationship between process model variables and what we can observe. For example, the process model predicts phytoplankton biomass  $P$  in the currency of  $\text{mg N/m}^3$ , but we typically measure phytoplankton as a pigment ( $\text{mg chl a/m}^3$ ). The submodel given in the following paragraphs allows us to relate these chlorophyll observations ( $\text{chl } a$ ) more rigorously to the state variable  $P$ . Our formulation represents a variant on models proposed by Geider et al. (1998), and details of our derivation are given in Appendix A: *A simple adaptive model of phytoplankton growth and composition in response to light, nutrient, and temperature*.

The phytoplankton-specific growth rate  $g$  is expressed in terms of  $g^{\max}$  (in units of  $\text{d}^{-1}$ ), a constant maximum specific growth rate at the reference temperature,  $T_{\text{ref}}$ , a light-limitation term  $h_E$ , and a nutrient-limitation term,  $h_N$ . That is,

$$g = Tc \times g^{\max} \times h_E \times h_N / (h_E + h_N). \quad (15)$$

The light limitation term is given by

$$h_E = 1 - \exp(-\alpha \times \lambda^{\max} \times E / g^{\max}) \quad (16)$$

where  $\alpha$  is the initial slope of the photosynthesis vs. irradiance curve ( $\text{mg C} \cdot \text{mg chl } a^{-1} \cdot \text{mol photon}^{-1} \cdot \text{m}^{-2}$ ) and  $\lambda^{\max}$  is the maximum chlorophyll  $a$ :carbon ratio ( $\text{mg chl } a / \text{mg C}$ ). The parameter  $\alpha = a_{\text{Ch}} \times Q$  is the product of the chlorophyll-specific absorption coefficient for phytoplankton,  $a_{\text{Ch}}$  ( $\text{m}^2 / \text{mg chl } a$ ), and the maximum quantum yield for photosynthesis,  $Q$  ( $\text{mg C} / \text{mol photons}$ ).

The nitrogen limitation term is given by

$$h_N = N / [(g^{\max} \times Tc / a_N) + N] \quad (17)$$

where  $a_N$  is the maximum specific affinity for nitrogen uptake ( $\text{mg N}^{-1} \cdot \text{m}^3 \cdot \text{d}^{-1}$ ).

The phytoplankton nitrogen:carbon ratio,  $\chi$ , predicted by the model is given by

$$\chi = \frac{\chi^{\min} \times h_E + \chi^{\max} \times h_N}{h_E + h_N} \quad (18)$$

where  $\chi^{\min}$  and  $\chi^{\max}$  are the minimum and maximum nitrogen:carbon ratios ( $\text{mg N/mg C}$ ).

The model predicts the phytoplankton chlorophyll  $a$ :carbon ratio  $\lambda$ , and this can be combined with the nitrogen:carbon ratio  $\chi$  to convert phytoplankton

biomass  $P$  ( $\text{mg N/m}^3$ ) to a predicted chl  $a$  concentration as

$$\text{chl } a = P(\lambda^{\max} / \chi^{\max}) \times h_N \times Tc / (R_N \times h_E + h_N) \quad (19)$$

where  $R_N = \chi^{\min} / \chi^{\max}$ . This growth model involves six parameters ( $g^{\max}$ ,  $\alpha$ ,  $\lambda^{\max}$ ,  $a_N$ ,  $\chi^{\max}$ ,  $R_N$ ). The parameters  $\alpha$ ,  $\lambda^{\max}$ , and  $\chi^{\max}$  appear only in terms of the ratios  $\alpha / \lambda^{\max}$ , and  $\lambda^{\max} / \chi^{\max}$ , but since  $\chi^{\max}$  is fixed based on the Redfield ratio, this does not result in redundant parameters in our inference procedure.

While this completes the specification of the local reactions  $\mathbf{R}$  given in Eq. 5, in the simple one-box, mixed-layer (i.e., zero-dimensional) model adopted here, we do need to allow for effects of physical exchanges between the mixed layer and the underlying water mass. These exchanges add additional source-sink terms to the right-hand sides of Eqs. 6–9, and these are specified in Appendix A: *The transport operator*.

*From a deterministic to a stochastic BGC process model.*—The BHM framework encourages us to formulate the state or process model in probabilistic or stochastic terms, in order to capture the effects of approximations and errors in the process representation. Note that a stochastic-model formulation is not equivalent to recognizing prior uncertainty in the (constant) parameters in a deterministic model. A deterministic model effectively asserts that, given the initial state and the parameters, the future state can be predicted exactly at all future times. A stochastic model asserts that, given the model state and parameters at the current time, we can make statements only about the probability distribution of the state at future times.

A deterministic model of the kind described in *Reformulating a marine BGC model as a BHM: The process model* can be converted to a stochastic model in a number of ways. The simplest approach is to introduce an additive error term on the right-hand side of equations, either as a continuous Wiener process for the differential equations Eqs. 6–9, or as a Gaussian error term at each time step in the discretized version. We have not adopted that approach here; we have tried instead to introduce randomness into the process model in a way that better reflects the approximations we make in formulating such models, and that preserves mass balance. Specifically, we replace the constant ecophysiological parameters in the deterministic model with stochastic processes that change as the underlying plankton community composition changes. In the remainder of this section, we provide motivation for, and a detailed explanation of, this approach.

A key approximation made in formulating models like the one given in *The process model*, involves biological aggregation. Phytoplankton and zooplankton communities, which consist of many different species, are each represented in the model by a single compartment. More complicated models may divide phytoplankton or zooplankton biomass into two or more functional groups with different ecological roles, but each group



still constitutes an aggregation of diverse species. The model formulations used in *The process model* are largely derived from many, many, laboratory studies of individual species or isolated samples, which give us reason for confidence in the structural form of the models. However, these studies also show very large levels of variation in many of these ecophysiological parameters, across individual species, or across field samples. Hence, the properties assigned to functional groups in these models must be thought of as representing some kind of average across the community of species making up the functional group.

The key point here is not just that variation exists, and so there is uncertainty in specifying these community properties, but that community composition varies over time, and so the community parameters must also be expected to vary over time. In models like those given in *The process model*, we do not attempt to explain or predict these changes in community composition (and consequently in community properties) mechanistically, but we can account for them by treating them as stochastic processes. Now, we expect some level of persistence in community composition, so it does not seem realistic to treat community properties as being drawn independently from some underlying distribution at each time step. Instead, we allow for community persistence by treating community properties as the outcome of a first-order autoregressive stochastic process.

This means that if  $b$  is a generic biogeochemical *parameter* in the deterministic model, we replace  $b$  by a stochastic BGC process  $B$  in the model, with

$$B(t + \Delta t) = B(t) \times (1 - \Delta t/\tau) + \zeta_B(t) \times \Delta t/\tau \quad (20)$$

for  $|1 - \Delta t/\tau| < 1$ . Here,  $\Delta t$  is the discrete time step (assumed to be 1 day in our example),  $\tau$  is the characteristic time of the autoregressive process (that is, the time scale on which community composition changes), and  $\{\zeta_B(t)\}$  represents a sequence of independent and identically distributed random variables with distribution  $[\zeta_B]$ . Detailed properties of this process required for our study are provided in Appendix B.

We can obtain prior information on the distribution  $[\zeta_B]$  by considering past laboratory and field studies. In fact, meta-analyses of past studies for many ecophysiological parameters have been conducted by researchers looking to establish systematic relationships between these parameters and individual size. These analyses show that parameters typically vary over orders of magnitude, so there is good reason to propose log-normal distributions for  $[\zeta_B]$  (i.e., normal distributions for  $\log(\zeta_B(t))$ ), for most parameters.

There are some further complications we need to consider in making the step from a meta-analysis of laboratory studies to specifying a prior for distributions like  $[\zeta_B]$ . The meta-analyses summarize results of measurements on individual species drawn from a wide variety of locations, but the processes  $B$  refer to means

over the community of species present at a particular location. We would expect the variance of the community mean to be less than the variance over the constituent species; this effect is dealt with explicitly in Appendix B. It is also possible that the species comprising a functional group at a particular location will be less diverse, and may exhibit lower variance, than the species represented in meta-analyses. We denote the ratio of the coefficient of variation (CV) of community mean parameters to the CV of species parameters by PDF for phytoplankton, and ZDF for zooplankton. In Appendix B, we relate these ratios to measures of community diversity.

Because of the lognormal nature of the autoregressive error  $\zeta_B(t)$  in Eq. 20, we consider the mean of  $B$ ,  $E(B)$ , and the coefficient of variation of  $B$ , namely  $CV(B) \equiv \sqrt{\text{Var}(B)}/E(B)$ . Appendix B shows how it is possible to choose the mean and variance of  $\log \zeta_B(t)$  such that  $E(B)$  and  $CV(B)$  are consistent with the mean and variance of individual species properties, given the values of PDF and ZDF. We treat PDF, ZDF, and the expected value  $E(\mathbf{B}) \equiv \mu_{\mathbf{B}}$ , where  $\mathbf{B}$  is the set of all BGC autoregressive processes, as parameters in  $\theta_{\mathbf{W}}$ . We also assume characteristic time scales for changes in phytoplankton community composition ( $\tau_P$ ), and likewise for zooplankton community composition ( $\tau_Z$ ).

We need to establish priors for the parameters controlling the behavior of the autoregressive processes: PDF, ZDF, and  $\mu_{\mathbf{B}}$ . We set broad, relatively uninformative, priors for PDF and ZDF. We also set relatively uninformative priors for the components of  $\mu_{\mathbf{B}}$ , by assigning them the same distribution (mean and variance) used to describe the individual species parameters, based on the meta-data (Appendix B). This means that the prior distribution allows the community parameter to take on the most extreme values revealed by individual species. For further information on priors and their derivations, see Appendix A: *The parameter (prior) model* and *Reformulating a marine BGC model as a BHM: The parameter (prior) model*.

We can now translate the stochastic BGC process model into the BHM formalism presented in *General methodology*. The process  $\mathbf{W}$ , as defined in *Bayesian hierarchical models (BHMs)*, can be split into the state vector  $\mathbf{X}$  and a vector  $\mathbf{B}$  that recall is the set of autoregressive BGC processes. That is,

$$[\mathbf{W}] = [\mathbf{X}, \mathbf{B}] \quad (21)$$

where the state is  $\mathbf{X} = \{N, P, Z, D\}$  and the (random) BGC processes are  $\mathbf{B} = \{g^{\max}, \lambda^{\max}, R_n, a_N, I_Z, \text{Cl}_Z, E_Z, r_D, m_Q\}$ . Similarly,  $\theta_{\mathbf{W}}$  in Eq. 3 can be split into two parameter sets, those appearing explicitly in the equations updating  $\mathbf{X}$ , namely,  $\theta_{\mathbf{X}} = \{K_W, a_{\text{ch}}, S_D, f_D\}$ , and those appearing in the autoregressive equations for the BGC processes  $\mathbf{B}$ , namely  $\theta_{\mathbf{B}} = \{\text{PDF}, \text{ZDF}, \mu_{g^{\max}}, \mu_{\lambda^{\max}}, \mu_{R_n}, \mu_{a_N}, \mu_{I_Z}, \mu_{\text{Cl}_Z}, \mu_{E_Z}, \mu_{r_D}, \mu_{m_Q}\}$  taking note that PDF and ZDF effectively scale the coefficient of variation,

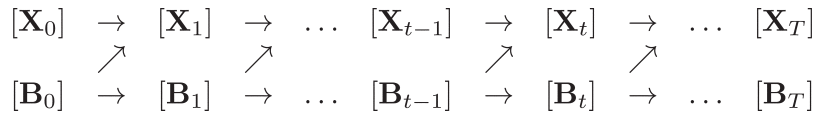


FIG. 2. Evolution of the state ( $\mathbf{X}$ ) and the biogeochemical (BGC,  $\mathbf{B}$ ) processes. Recall that  $\mathbf{W} = [\mathbf{X}, \mathbf{B}]$  and that Fig. 1 shows how observations  $\mathbf{Y}$  are related to the process  $\mathbf{W}$ .

$\text{CV}(\mathbf{B})$ , given in Appendix A: Table A1. The state-space representation is now as given in Fig. 2.

In terms of conditional probabilities, the formulation developed in this section means that

$$\begin{aligned} [\mathbf{W}_t, \mathbf{W}_{t-1}, \boldsymbol{\theta}_W] &= [\mathbf{X}_t, \mathbf{B}_t | \mathbf{X}_{t-1}, \mathbf{B}_{t-1}, \boldsymbol{\theta}_W] \\ &= [\mathbf{X}_t | \mathbf{X}_{t-1}, \mathbf{B}_{t-1}, \boldsymbol{\theta}_X] [\mathbf{B}_t | \mathbf{B}_{t-1}, \boldsymbol{\theta}_B] \quad (22) \end{aligned}$$

where the last equality expresses the fundamental evolution of the process model (*Reformulating a marine BGC model as a BHM: The process model: From a deterministic to a stochastic BGC process model*).

#### The parameter (prior) model

The priors assigned to the parameters specified in this study were drawn from a meta-analysis of the literature. A summary of the prior information available for the BGC parameters and processes, and the sources of this information, is given in Appendix A: *The parameter (prior) model*. Each component of the prior is assumed independent of the other components, and no attempt has been made to introduce any dependence structure between the parameters.

#### The data model

The data model explicitly links the process model with the observations. The parameters  $\boldsymbol{\theta}_Y$  in Eq. 2 control the observation process, and we consider the following two broad classes of observation error: (1) Analytical measurement errors should reflect the precision of in situ instruments or laboratory analyses. For example, laboratory determinations of chlorophyll *a* pigment concentration might be expected to have a precision of a few percent. (2) Representation errors can arise from mismatches in scale (we may model a large volume of ocean, many kilometers across, but make measurements on bottle samples comprising a few liters) and mismatches in type (we may predict zooplankton concentration in the currency of nitrogen, but measure volume or wet mass of biomass).

In most real-world situations, errors associated with mismatches in scale and type outweigh analytical measurement errors. The use of a simple one-box mixed-layer model here introduces an additional ambiguity. We are neglecting horizontal advection, which might be thought of as an additional process-model error. The significance of horizontal advection compared with local processes depends on the area of ocean represented by the box. If we regard the box as representing an ocean area several hundred kilometers

in extent, we might hope that the errors involved in neglecting advection are small. But we must then expand the observation error to account for the spatial variability observed on these length scales.

In *Learning and predictability given observations*, the data model for our application to data from Ocean Station Papa is given by

$$[\mathbf{Y} | \mathbf{W}, \boldsymbol{\theta}_Y] = \prod_{t=1}^T [\mathbf{Y}_t | \mathbf{X}_t, \boldsymbol{\theta}_Y]. \quad (23)$$

Treatment of  $\boldsymbol{\theta}_Y$  for our case study is discussed in *Learning and predictability given observations*. Recall that  $\mathbf{W}$  is made up of  $\mathbf{X}$  and  $\mathbf{B}$ ; note that if we had direct observations of the ecophysiological properties represented in  $\mathbf{B}$ , these could be incorporated into the data model.

#### LEARNING AND PREDICTABILITY GIVEN OBSERVATIONS

We demonstrate the application of the BHM framework to a marine BGC model using the historical Ocean Station Papa (OSP) data set as a case study. This site was chosen over alternative subtropical time series sites because the simple mixed layer model is believed to be a better approximation at OSP. Two experiments were conducted. First, a twin experiment was run using climatological forcing at OSP, with synthetic observations of all state variables assimilated daily. The synthetic observations were generated by adding noise to a known “true” trajectory through the state space. Second, a subset of the historic OSP data set comprising observations of chlorophyll *a* (chl *a*) and nitrate (*N*) was assimilated for the period January 1971–November 1974. This corresponds to part of a sustained observing campaign, and we found that the marginal posteriors for parameters did not change greatly if additional years were included.

#### Ocean Station Papa site description

Ocean Station Papa (OSP) is located at 50° N, 145° W (Fig. 3), in 1500 m of water in the subarctic region of the north east Pacific Ocean. It experiences a strong seasonal cycle in temperature, wind stress, and incident solar radiation (Whitney and Freeland 1999). During winter and spring, a mixed layer of depth 80–120 m is sustained by a high wind stress with the low incident solar radiation unable to induce any persistent stratification of the water column. During summer, the thermocline shallows in response to increased surface heating and a reduction in the wind stress. Consequent-

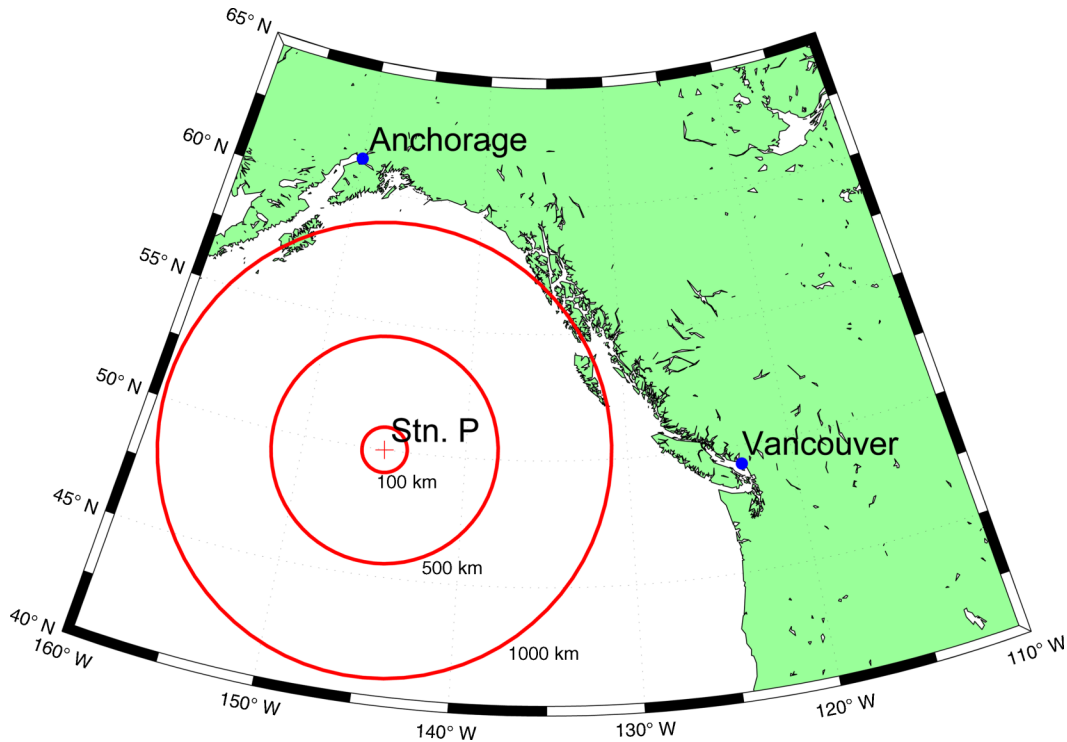


FIG. 3. A map of the northeast Pacific Ocean displaying the location of Ocean Station Papa (Stn. P) with range circles at 100, 500, and 1000 km.

ly, a relatively shallow mixed layer is maintained of typical depth 25–40 m.

It has been noted that there are persistently high macronutrient concentrations in the mixed layer and the phytoplankton biomass is typically low. This phenomenon is observed throughout much of the open subarctic Pacific ocean. While the concentration of dissolved inorganic nitrogen (DIN) is lower in summer than in winter, it is rarely if ever depleted to levels that may cause nutrient limitation in primary producers (Harrison 2002). There is no discernible seasonal cycle in chlorophyll *a*. Previous modeling studies of Matear (1995), Denman and Pena (1999), and Denman (2003) discuss the likely controls on phytoplankton biomass and the seasonal variation in primary productivity and zooplankton biomass.

*Learning from observations: twin experiment with climatological forcing*

Twin experiments in a setting like that of OSP have been conducted to compare samples from the posterior,  $[\mathbf{W}, \boldsymbol{\theta} | \mathbf{Y}]$ , produced by Bayesian inference, with known “true” values of the state and parameters. The term “twin,” borrowed from the data-assimilation literature, refers to experiments where the model used for inference, and the model from which synthetic observations are generated, are the same. Model forcing and boundary conditions are taken from Matear (1995) and

are climatological in nature; details are given in Appendix D.

*Twin experiment: design.*—To generate the synthetic observations, we select a parameter set  $\boldsymbol{\theta}^*$  (the “true” parameters) and take a single realization of the stochastic model  $\{\mathbf{W}_t^*; t = 0, 1, \dots, T\}$  to produce the trajectory  $\{\mathbf{X}_t^*; t = 0, 1, \dots, T\}$  through state space (again referred to as the truth). We have chosen a set of “true” parameters in the twin experiment that are shifted away from the prior means (to provide a clearer test of the inference procedure), but that nevertheless yield state-variable trajectories qualitatively consistent with OSP observations (e.g., high-nutrient low-chlorophyll (HNLC) conditions). The (synthetic) observations  $\mathbf{Y}$  are generated by

$$\mathbf{Y}_t = \mathbf{X}_t^* \exp(\boldsymbol{\xi}_t) \quad t = 0, 1, \dots, T \quad (24)$$

where  $\boldsymbol{\xi}_t$  are independent and identically distributed (IID) as the normal distribution  $N(0, \sigma_{\text{obs}}^2)$ . The standard deviation,  $\sigma_{\text{obs}}$ , was 0.1 for DIN observations and 0.2 for observations of the remaining state variables. The log-normal error model was adopted because errors in the estimates of plankton density are typically better represented by log-normal multiplicative error than by additive normal error (Campbell 1995), and the log-normal multiplicative-error model delivers synthetic observations that are nonnegative. The observation errors are assumed to be independent over time,

reflecting either analytical error or (more likely) uncorrelated small-scale variation in concentrations.

*Twin experiment: results.*—We first generate an ensemble of model trajectories by sampling from the prior distribution for parameters and running the stochastic model forward through the period January 1971–November 1974, without assimilating any observations. This so-called free-run process-model ensemble is precisely a sample from the prior distribution over the state (Fig. 4, blue shading), which expresses the uncertainty in the state based only on the prior knowledge of the parameters gained from a meta-analysis of the literature. In spite of the large prior uncertainty in some of the process-model parameters, the median values of the (marginal) prior distributions over state variables show surprisingly similar qualitative behavior to the observed climatology at OSP (Fig. 4, dark blue line). The median DIN values remain elevated, and median chlorophyll *a* values remain low. However, the 95% contours of the prior ensemble include unrealistic behaviors not observed at OSP, involving near-complete depletion of DIN and intense phytoplankton blooms.

When the synthetic observations described in *Twin experiment: design* are assimilated, using the methodology described in *Reformulating a marine BGC model as a BHM*, the 95% credibility intervals for the posterior distribution of the state are very tightly constrained about the true trajectory (Fig. 4, red shading), compared with the prior intervals and with the observations. Despite the 20% observation error, the dynamical BHM implemented through the PMCMC described in Appendix C, accurately tracks the true state (Fig. 4, green line).

The case for N deserves additional explanation. The seasonally varying N concentration, prescribed below the mixed layer as a boundary condition, imposes a sharp upper limit to the predicted mixed-layer N concentrations. Provided grazing control keeps phytoplankton biomass and N utilization small, the predicted concentration is very close to this upper bound. In most prior trajectories, grazing control is effective, so the prior median is close to the upper limit. Some prior parameter combinations allow phytoplankton blooms and N depletion, resulting in the drawdown of N to near-zero levels seen in the prior lower 95th percentile for N. The truth is chosen to be OSP-like, and so produces N concentrations close to the upper bound. Since we add noise to the truth, a significant fraction of the observations lie above the upper bound.

The prior distributions over the parameters given in Appendix A: Table A1 are the blue curves in Fig. 5. These priors are discussed in *Reformulating a marine BGC model as a BHM: The parameter (prior) model* and are considered “global” in that they represent experimental results encompassing a wide range of species and domains. For some model parameters ( $a_{Ch}$ ,  $s_D$ , PDF, ZDF,  $\mu_{g^{max}}$ ,  $\mu_{\lambda^{max}}$ ,  $\mu_{Cl_Z}$ ,  $\mu_{E_Z}$ ,  $\mu_{r_D}$ , and  $\mu_{m_Q}$ ) the marginal posteriors in Fig. 5 show evidence of learning in that the

posterior mode has moved toward the truth and the posterior variances have contracted compared with the prior. However, for others, the inference procedure appears to extract little or no information from the data, and the marginal posteriors appear to merely recover the prior distributions. This is true for the parameters controlling light attenuation due to water ( $K_W$ ), the fraction of zooplankton waste diverted to detritus ( $f_D$ ), the parameters related to nitrogen uptake and nitrogen:carbon ratios ( $a_N$  and  $R_N$ ), and the maximum zooplankton ingestion rate ( $I_Z$ ). In the case of  $a_N$ , the posterior variance is slightly reduced, but the posterior median remains centered at the prior mean.

The inference procedure generates posterior distributions for time series of the autoregressive processes  $\mathbf{B}(t)$ , and could provide information about changes over time in the ecophysiological properties they represent. However, the results from this twin experiment are only mildly encouraging in this regard. In cases where the observations are uninformative about the parameters underlying the autoregressive processes, one can hardly expect to obtain information about the temporal variation in the processes themselves. Indeed, in those cases, the posteriors for the stochastic-process trajectories are the same as the priors. In two cases ( $g^{max}$  and  $Cl_Z$ ), the posterior median trajectories appear to track the truth, although with consistent bias in the case of  $Cl_Z$  (Fig. 6). But for these, and all other autoregressive processes, the 95% credibility interval for the posterior exceeds the amplitude of the temporal variation in the truth by some margin. The inference procedure does not allow us to conclude that there are significant changes in these processes over time.

These results reflect the particular nature of the climatological forcing and system behavior at OSP. Given that concentrations of dissolved inorganic nitrogen at OSP remain well above levels expected to limit phytoplankton growth, it is unsurprising that parameters controlling nitrogen limitation of growth rates are poorly constrained. Similarly, phytoplankton biomass remains at levels well below those required to saturate zooplankton grazing, and zooplankton growth rates are controlled by the clearance rate  $Cl_Z$ , not by the maximum ingestion rate.

#### *Learning from observations: Ocean Station Papa data set*

To demonstrate the application of the BHM approach to a real data set, we have used a subsample of historical OSP data.

*Ocean Station Papa: data model.*—Observations of nitrate (DIN) and chlorophyll *a* taken between January 1971 and November 1974 are used. Observation errors are large and dominated by spatial sampling errors, because we neglect horizontal advection and assume a large model domain with high levels of within-domain variability. The presence of larger observation errors means that the data will be less informative. We draw on



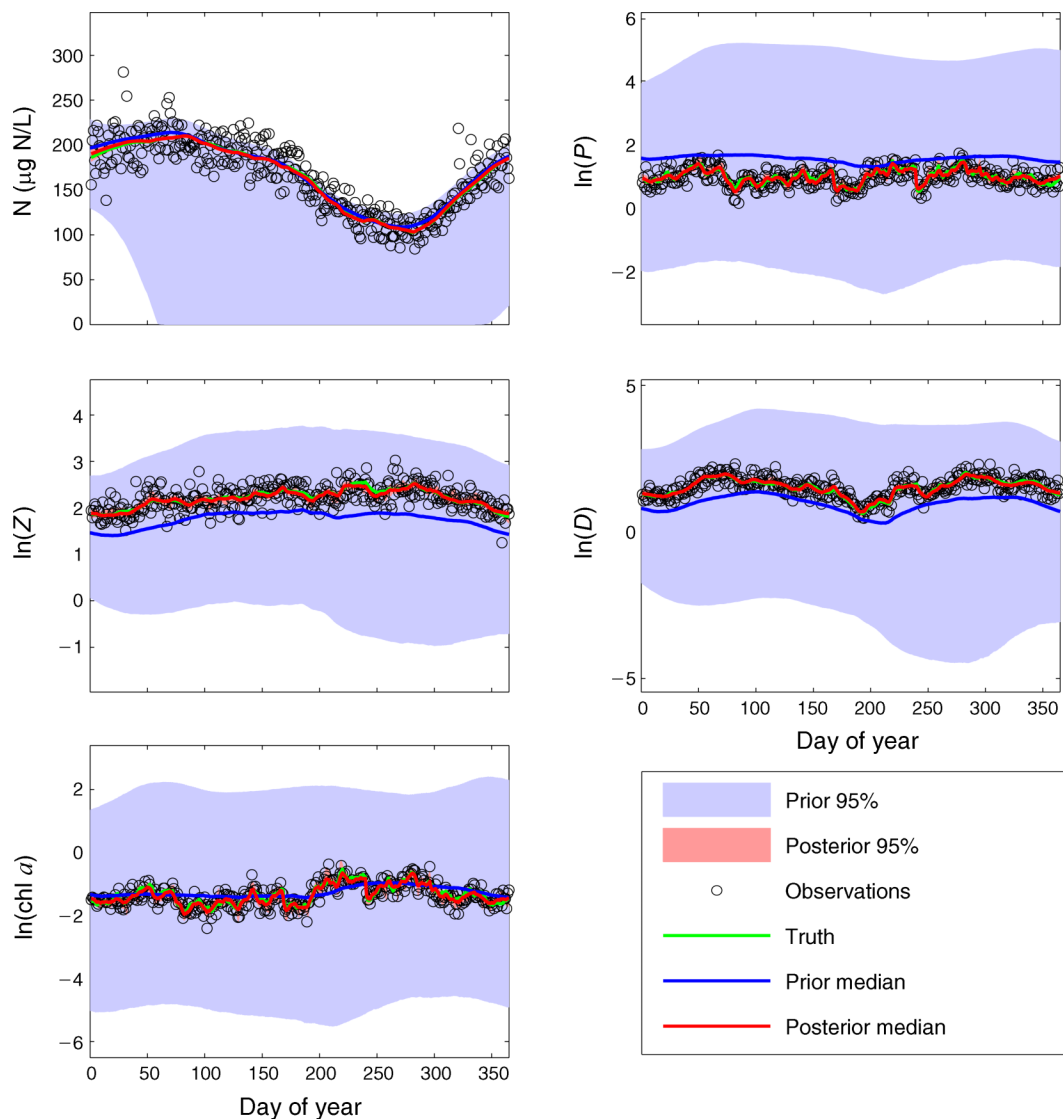


FIG. 4. Twin experiment: a time series of the prior and posterior distributional properties of the state variables for phytoplankton ( $P$ ), zooplankton ( $Z$ ), nutrients ( $N$ ), and detritus ( $D$ ) in the common currency of nitrogen (measured as  $\mu\text{g N/L}$ ) and the instantaneous concentration of chlorophyll  $a$  pigment ( $\text{chl } a$ ;  $\mu\text{g chl } a/\text{L}$ ). Note that the posterior credibility intervals remain so close to the posterior median that they are difficult to distinguish.

a number of studies below for estimates of the appropriate levels of spatial variability.

The spatial and temporal variability of particulate organic carbon (POC) in this region has been investigated at a number of scales (Bishop et al. 1999). The spatial variability in the vertical and horizontal directions was calculated from the beam-attenuation coefficient obtained from a transmissometer. Small-scale horizontal variability (1–10 km) of POC appears to be 5–10%, which is deemed negligible in comparison to the ocean scale and temporal variability. Large-scale horizontal variability (100–300 km) of POC appears to range from 10% to 40%, however, we attribute some of this variability to the passage of weather systems on time

scales of 5–10 days. Ocean-basin-scale variability (800–2000 km) exceeds both the large-scale and small-scale spatial variability, but this is due to the change from HNLC conditions in the deep ocean to a more typical temperate seasonal cycle on the continental shelf.

Bishop et al. (1999) also noted significant interannual variability that may be linked to El Niño events. Nitrate data collected along the Line P transect (a 1425 km long transect between the coast adjacent to the Juan de Fuca Strait and Ocean Station Papa [e.g., Pena and Bograd 2007]) from 1992 to 1997 display a similar pattern to the POC data. Again, it appears that on the scale of 100–300 km around OSP, variability in total concentrations of nitrates and nitrites appears to be 10–30%, with



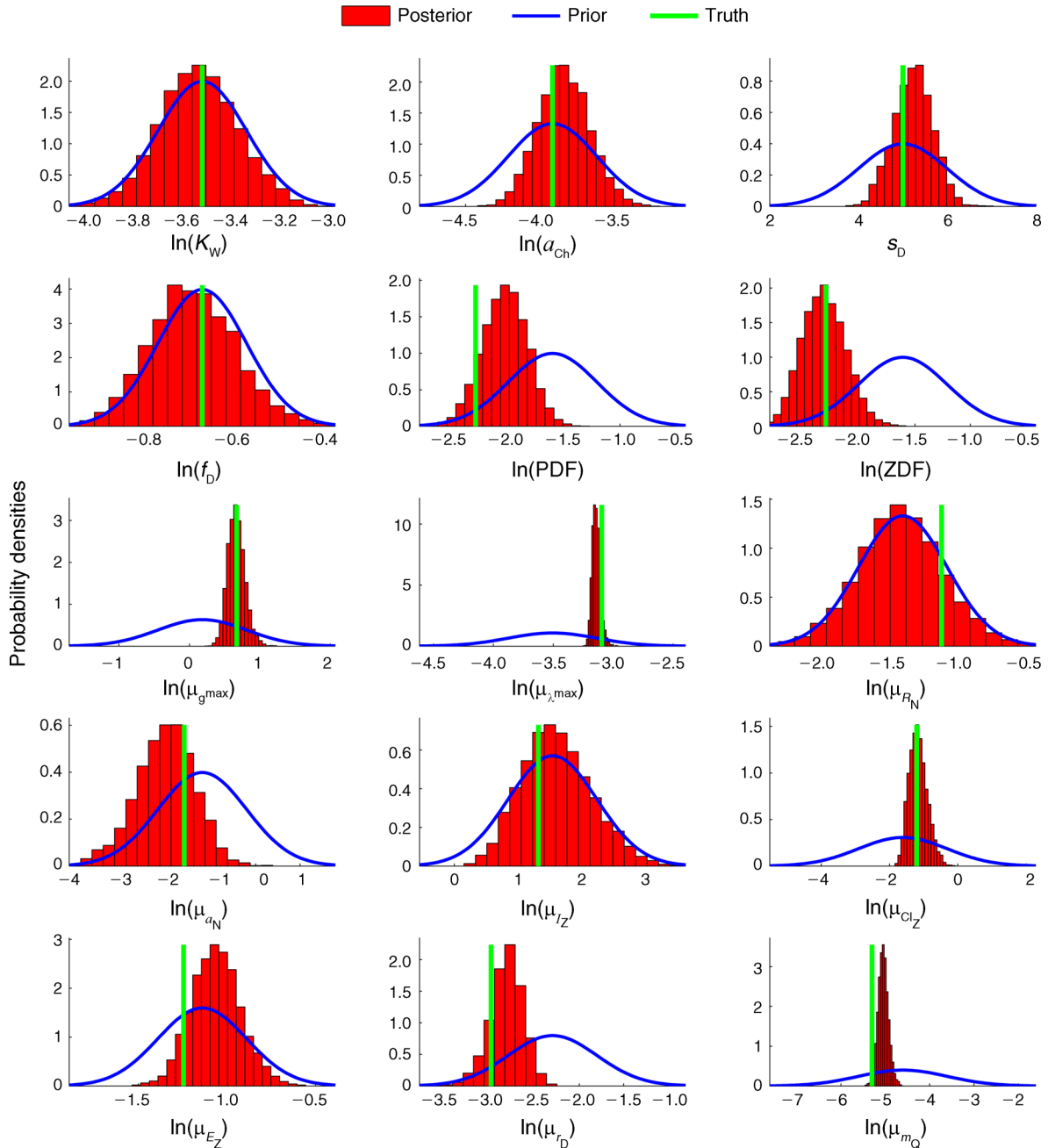


FIG. 5. A comparison between the prior (blue curve) and the posterior (red histogram) for all parameters  $\theta_X$  and  $\theta_B$ . The true value is given by the vertical green line. The vector  $\theta_X = \{K_W, a_{Ch}, s_D, f_D\}$  and contains parameters affecting the state, while  $\theta_B = \{PDF, ZDF, \mu_{gmax}, \mu_{\lambda max}, \mu_{RN}, \mu_{aN}, \mu_{Iz}, \mu_{Clz}, \mu_{Ez}, \mu_{rD}, \mu_{mQ}\}$  and contains parameters controlling the distributional properties of the autoregressive processes. For a description of the parameter symbols, please refer to Appendix A: Table A1. Graphs show probability densities; the area under each curve and of each histogram is normalized to 1.

interannual variability exceeding the large-scale spatial variability (Whitney and Freeland 1999).

Taking all these sources of information into account, we have assigned a CV ( $\sigma_{obs}$ ) of 0.5 to the observation error for both DIN and chlorophyll *a*. This is a conservative (upper) estimate, representing an upper bound to spatial variation, and allowing for other

nonspatial contributions, including analytical measurement error.

*Ocean Station Papa: results from hindcast.*—A prior ensemble over the state was constructed in a similar manner to the twin experiment, using real forcing from January 1971 to November 1974. Model parameters were sampled from the prior distributions described

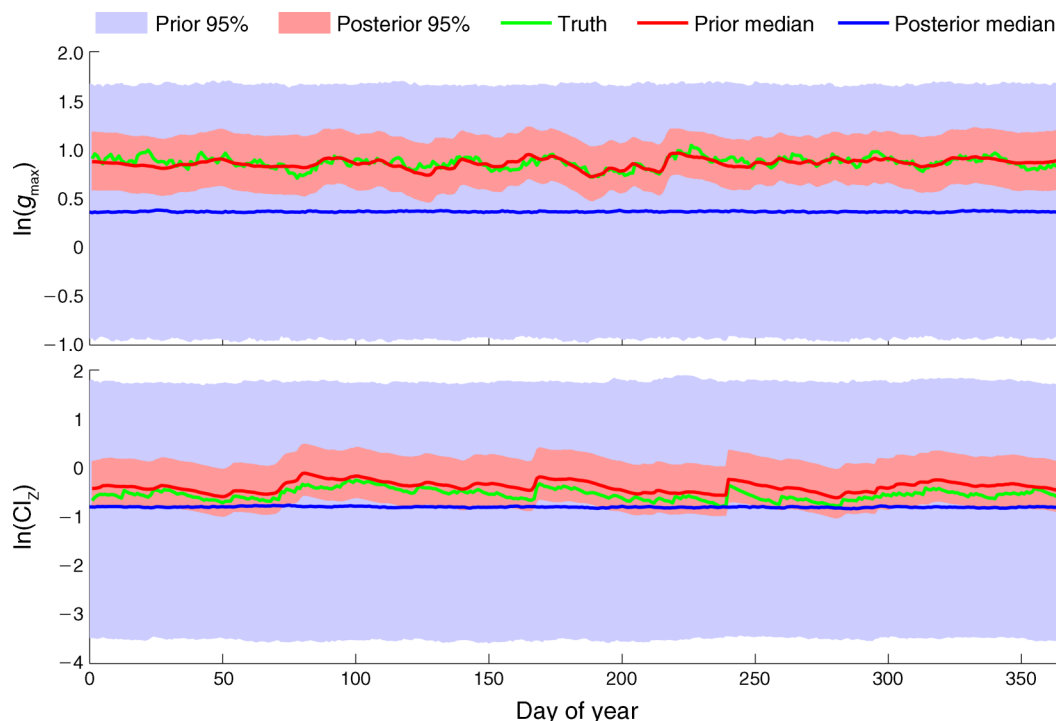


FIG. 6. Twin experiment: a time series of the prior and posterior distributional properties of the autoregressive for the maximum carbon-specific growth rate of phytoplankton ( $g_{\max}$ , measured as  $\text{d}^{-1}$ ) and the maximum zooplankton clearance rate ( $\text{Cl}_z$ , measured as  $[\text{m}^3 \text{ swept clear}] \cdot [\text{mg N}]^{-1} \cdot \text{d}^{-1}$ ). These processes are a subset of  $\mathbf{B}(t)$ .

earlier. As in the twin experiment, a wide range of model behaviors was observed (Figs. 7 and 8), ranging from near-complete depletion of DIN during summer, to year-round grazing control. As in the twin experiment, the median of the prior over the state based on 1971–1974 forcing qualitatively agreed with observed OSP behavior, in that DIN was never limiting, and there were no strong phytoplankton blooms as zooplankton grazing maintained relatively constant phytoplankton biomass (Matear 1995, Denman and Pena 1999, Denman 2003).

When observations of chlorophyll  $a$  and DIN are assimilated, the 95% credibility interval is dramatically reduced. Due to the relatively large observation error prescribed (see *Reformulating a marine BGC model as a BHM: The parameter (prior) model*), the transient, low-magnitude increase in chlorophyll  $a$  seen in the summer of 1972 is absorbed into the observation error and not tracked in the state. While the three individual observations of this anomalous bloom do not fall within the posterior 95% credibility interval, this cannot be interpreted immediately as lack of model fit. This is because the credibility interval depicted is over the latent chlorophyll  $a$  state variable, not over the “noisier” observed chlorophyll  $a$ ; this distinction is important and is discussed by Cressie and Wikle (2011: Section 2.2.2). Although short-lived transient features are not tracked by the model, slow seasonal and intra-seasonal variations are well captured. The methods described in

*General methodology: Statistical inference* not only condition the state on observations from previous times, as do filtering approaches, but also on future times. This is referred to as *smoothing* in the Bayesian filtering literature (Briers et al. 2010, Fearnhead et al. 2010). The advantage of such smoothing is evident in time periods where there are very few observations (e.g., mid-1973).

Through the process model, Bayesian methods allow inference on the unobserved state variables  $P$ ,  $Z$ , and  $D$ ; see Fig. 8. Notice that there is a substantial reduction in the uncertainty expressed through the posterior compared with that expressed through the prior, even for unobserved state variables. For example, there is a strong seasonal cycle in the zooplankton biomass, which has been observed in a number of studies (Harrison 2002). The peak in the zooplankton biomass occurs during mid summer, which coincides with a peak in primary production (not shown).

The marginal posteriors for model parameters shown in Fig. 9 demonstrate that the sparse and limited OSP observations carry very little information about many of the parameters. This was not unexpected; previous studies have also experienced difficulty in using the OSP data set to estimate parameters in deterministic models (Matear 1995). The large observation variances used here, which compensate for effects of advection, reduce the effective information content of the data, but we believe this is realistic, given the model structure. The

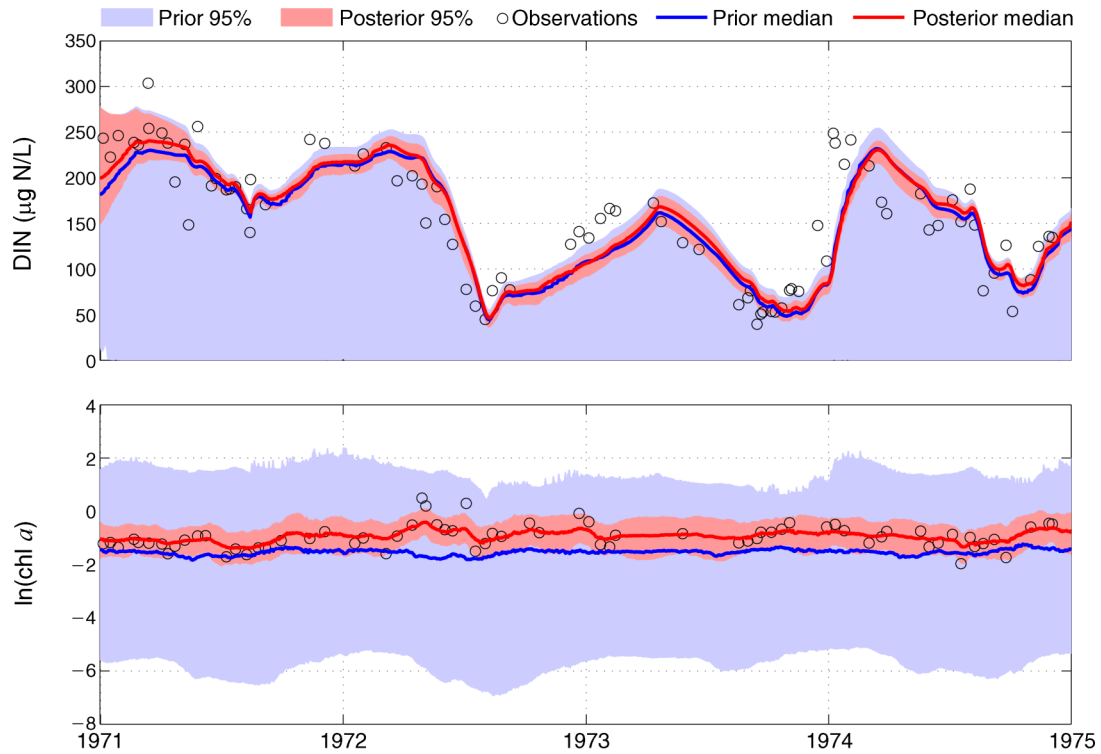


FIG. 7. A time series of prior and posterior distributional properties of observed state variables, comparing observations (open black circles), prior (blue), and posterior (red). DIN stands for dissolved inorganic nitrogen. Note that chl *a* was measured in  $\mu\text{g/L}$ .

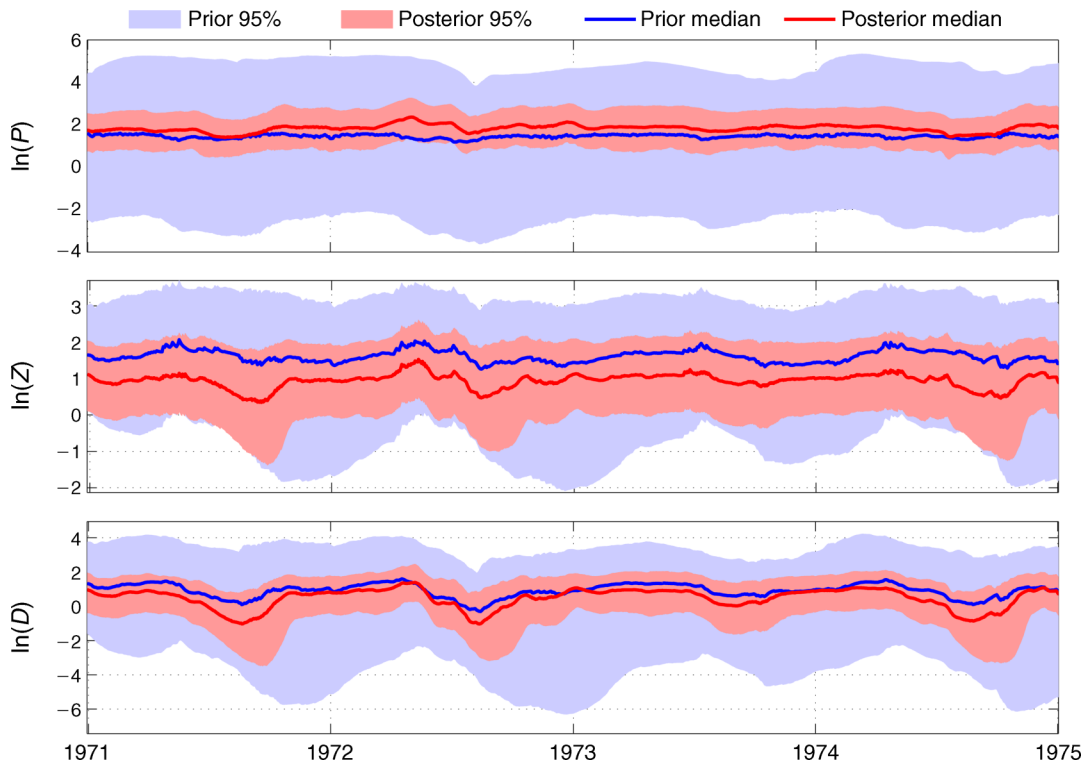


FIG. 8. A time series of prior and posterior distributional properties of unobserved state variables for phytoplankton (*P*), zooplankton (*Z*), and detritus (*D*) (all measured as  $\mu\text{g N/L}$ ), comparing the prior (blue) and posterior (red).

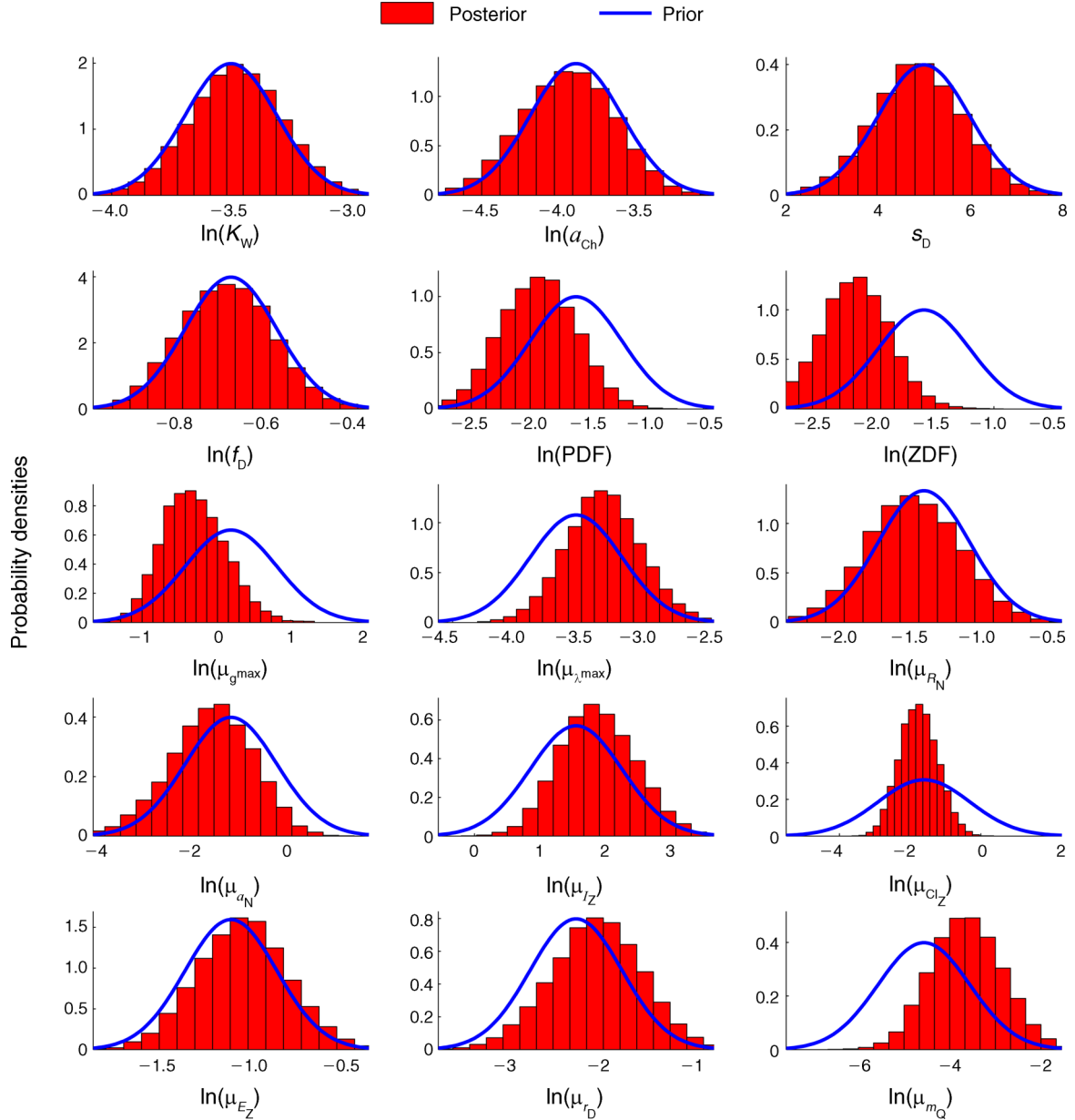


FIG. 9. A comparison between the prior (blue curve) and the posterior (red histogram) for all parameters  $\theta_{\mathbf{x}}$  and  $\theta_{\mathbf{B}}$ , where  $\theta_{\mathbf{x}} = \{K_W, a_{Ch}, s_D, f_D\}$  and contains parameters affecting the state, while  $\theta_{\mathbf{B}} = \{PDF, ZDF, \mu_{g^{\max}}, \mu_{\lambda^{\max}}, \mu_{R_N}, \mu_{a_N}, \mu_{I_Z}, \mu_{Cl_Z}, \mu_{E_Z}, \mu_{r_D}, \mu_{m_Q}\}$  and contains parameters controlling the distributional properties of the autoregressive processes. For a description of the parameter symbols, please refer to Appendix A: Table A1. Graphs show probability densities; the area under each curve and of each histogram is normalized to 1.

posterior marginals show evidence of learning for four parameters:  $ZDF$ ,  $g^{\max}$ ,  $\mu_{Cl_Z}$ , and  $\mu_{I_Z}$ .

Perhaps unsurprisingly, given the high noise levels and sparse observations, the OSP data do not allow us to derive useful information about temporal variation in the autoregressive processes  $\mathbf{B}(t)$ . Even for those parameters,  $g^{\max}$  and  $Cl_Z$ , where the observations appear to inform the posteriors for the underlying parameters, the posteriors for the autoregressive trajec-

tories show no significant variation over time (not shown).

One advantage of the BHM framework is that we can use the sample generated from the joint posterior of the state and parameters, conditioned on past observations, to assess the uncertainty in model forecasts and scenarios. In this case, we have used the posterior conditional on observations from January 1971 to November 1974 to make a probabilistic forecast for

1975. We do this simply by propagating all posterior trajectories forward the additional year, using the boundary and forcing fields for that year. The results from this forecast ensemble (median and 95% credibility intervals) are shown in Fig. 10. Agreement with the (non-assimilated) observations in the forecast period is very good.

#### DISCUSSION AND CONCLUSIONS

A key consideration in building BHMs is the treatment of model error. In our study, we used the fact that the aggregation of communities of species into single trophic levels or functional groups, and the replacement of well defined ecophysiological parameters for individual species by community-average parameters, is an important source of model error. Consequently, we have replaced the constant community parameters used in most biogeochemical models by stochastic autoregressive processes that vary slowly in time. This is in contrast to the common approach of simply adding white noise to the rate equations.

A potential drawback of this approach is that it increases the complexity and dimensionality of the model and the inference problem. We have augmented the four-dimensional primary state space ( $N, P, Z, D$ ) with nine additional state variables (**B**). Instead of estimating nine constant parameters, we must estimate nine means and nine variances controlling the evolution of the stochastic processes **B**( $t$ ). We have mitigated this problem by using prior information to set the relative magnitudes of the variances of phytoplankton and zooplankton community parameters, and using stochastic factors related to community diversity to set the absolute magnitude. One advantage of the process model, as formulated, is that it allows a strong and direct connection to literature meta-data on the distributions of ecophysiological parameters across species. This allows us to set informative objective priors for most of the parameters, exploiting a key advantage of Bayesian approaches, and partially counterbalancing the increase in unknowns.

The inference procedure was designed to derive joint posteriors for system parameters and the (augmented) system state. Many examples of data assimilation in dynamical models concentrate on either state estimation or parameter estimation. Joint inference is particularly difficult in nonlinear models with sparse data, and it has typically required strong simplifying approximations, such as the replacement of nonlinear dynamics by approximating linear models. The underlying deterministic NPZD model is highly nonlinear, displaying two qualitatively different modes of behavior or local stability domains, and the observed behavior at OSP correspond to only one of these domains. The new particle MCMC techniques employed here are able to cope with this nonlinear, threshold behavior, but are computationally expensive.

Given these challenges, the results of the OSP case study offer a number of grounds for encouragement. First, the stochastic process model allows the construction of priors over the model state, by drawing random samples from the prior distribution for model parameters and initial conditions, and running ensembles of model simulations. We can think of this prior ensemble as encapsulating our ability to predict system behavior at OSP, given independent scientific knowledge about BGC processes, and local environmental forcing, but no other local knowledge. Encouragingly, the state median in these prior distributions bears a strong qualitative and even quantitative resemblance to OSP observations (Figs. 4 and 7), even though the priors were chosen to reflect the full range of species attributes reported in the literature. But the 95% credibility intervals for the prior distribution also include trajectories involving phytoplankton blooms and nitrate depletion, which are incompatible with observations at OSP.

Data assimilation into dynamical process models can serve a variety of different diagnostic and prognostic purposes (see, e.g., Gregg 2008, Luo et al. 2011). One class of diagnostic applications targets the hindcasting or nowcasting of system state, given limited observations. Despite sparse observations with large sampling errors on one state variable ( $N$ ) and one diagnostic variable (chl  $a$ ), the Bayesian inference procedure recovers quite tight posteriors for these observed variables (Fig. 7). The Bayesian inference procedure is also able to transfer information from observed to unobserved state variables, reducing the uncertainty in the unobserved state variables ( $P, Z$ , and  $D$ ) by about one-half (Fig. 8).

A second class of diagnostic applications focuses on learning about, and interpretation of, model parameters. Here, the parameters describe the ecological characteristics of the plankton communities present at OSP. To the extent that these parameters have smaller variances a posteriori, we can conclude that the observations have provided information about the parameters and the communities they represent. The results for OSP are informative and cautionary. In the twin experiment, with observations on all state variables, the posteriors for some parameters are essentially identical to the priors, so provide no additional information (Fig. 5). These results can be explained in terms of model dynamics. Since nutrients under OSP conditions are always saturating, and phytoplankton concentrations remain low, the parameters affecting phytoplankton growth at low nutrient concentrations, and zooplankton ingestion at high phytoplankton concentrations, have negligible effect on model predictions, and are not identifiable. This pattern of identifiability is an intrinsic characteristic of the environmental forcing and dynamics at OSP. In other ocean conditions, such as oligotrophic mixed layers where nutrient concentrations are always low and limiting, one would expect different sets of model parameters to be identifiable.



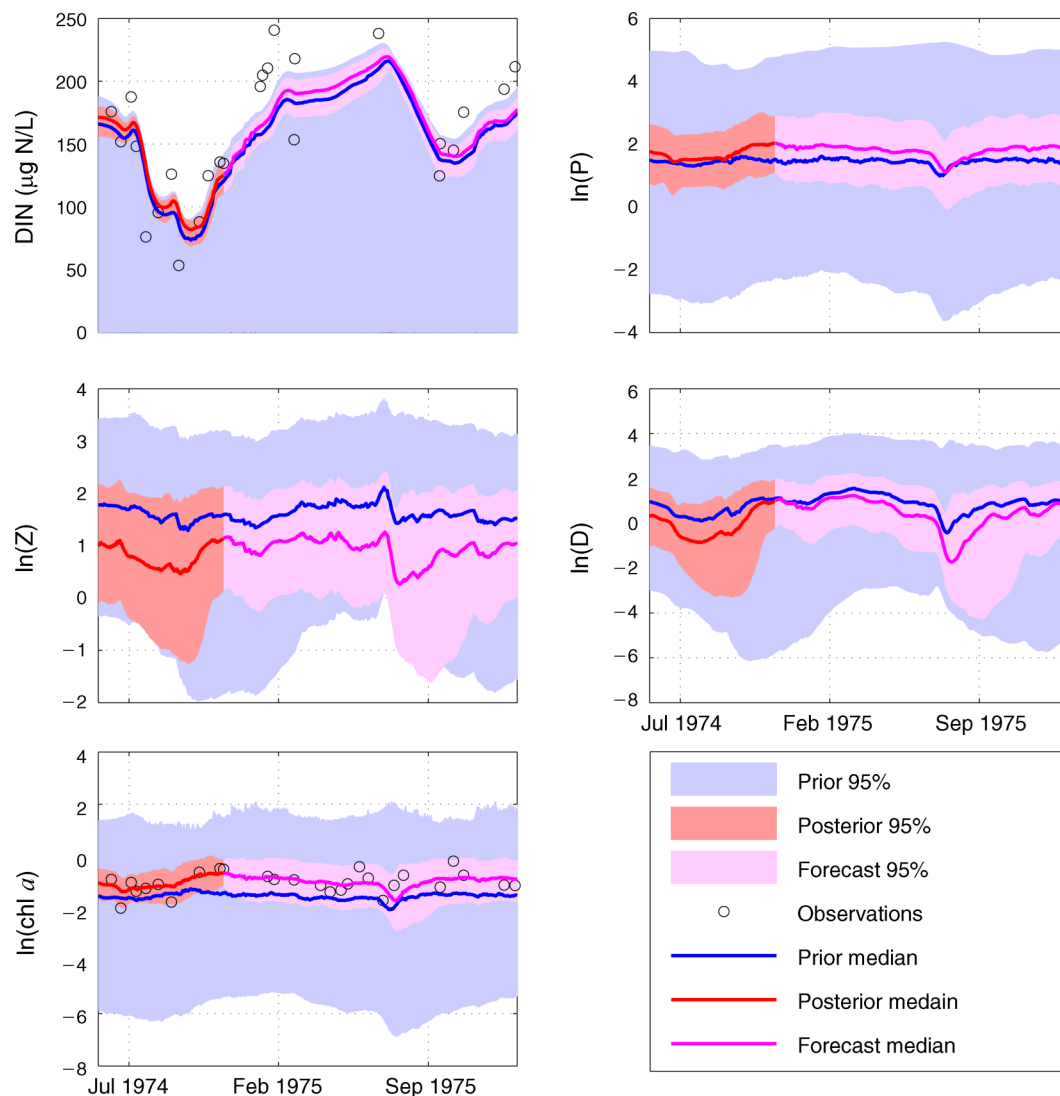


FIG. 10. A forecast (magenta shading) for the model's state variables of phytoplankton ( $P$ ), zooplankton ( $Z$ ), and detritus ( $D$ ) in the common currency of nitrogen (DIN) and the instantaneous concentration of chl  $a$  pigment for the period December 1974–December 1975. The forecast was initialized from the posterior values shown in Figs 7, 8, and 9. The posterior and prior of the state are given by the red and blue shading, respectively. Observations are denoted by the open black circles; during the forecast period, the observations were not assimilated.

Using the limited set of historical observations available for OSP, the inference procedure is able to extract information about a few key parameters only (Fig. 9). For the most part, these parameters directly control the key processes involved in zooplankton grazing control of phytoplankton. The posterior distributions for the parameters controlling the variance in model parameters (PDF and ZDF) are shifted toward lower values, compared with the priors. At the inferred lower levels of stochastic noise, trajectories are less likely to escape the local stability domain corresponding to grazing control.

The Bayesian inference procedure provides posterior distributions for the trajectories of the stochastic BGC

processes,  $\mathbf{B}(t)$ . Reliable information on changes in these processes would be of particular interest to plankton ecologists. However, even in a twin experiment with daily data on all state variables, we were only able to obtain suggestive (but not confirmatory) information about temporal variation in two parameters. This limited success is understandable, given that we are effectively trying to extract information about changes in unobserved variables on relatively short time scales, when the evidence of these changes is available only indirectly through changes in the time derivatives of the observed state variables. Even modest levels of observation noise are sufficient to confound this attempt. We conclude that higher frequency observations, and/or

lower observation noise, would be required to learn about temporal variation in these community properties from observations of state variables alone.

A twin experiment with similar forcing, sampling pattern and observation noise to the historical OSP observations yielded qualitatively equivalent results to those obtained using the real data. While not conclusive, this does suggest that the limited information about state and parameters obtained using the historical observations may be attributed to their sparseness and high observation noise, rather than an inconsistency with structural model assumptions.

It is common practice to distinguish short-term forecasts, in which uncertainty is dominated by the error in estimates of the current system state, from longer-term forecasts or projections, in which uncertainty may be dominated by errors in model structure, errors in parameter estimates, and the underlying stochastic process error. The methods used here allow us to move seamlessly from short-term to long-term forecasts. The forecast results are encouraging (Fig. 10), especially given that the inference procedure and observations have provided information about a small subset only of model parameters. This limited information, combined with prior information on other parameters, is sufficient to produce a long-term forecast that agrees both qualitatively and quantitatively with observations.

Given the limited identifiability of both parameters and the related stochastic biogeochemical processes, one could reasonably ask whether the model is over-parameterized. This would be the case if we were building a model specifically for the purpose of explanation or prediction at OSP that ignored prior information on model structure and parameters. However, we are engaged in developing a generic model, based on well accepted principles and strong prior information. The model is applied at OSP, but we envisage the same model (or a similar model) being applied at many other locations, and in the long run used as a basis for basin-scale or global BGC models spanning many different environmental conditions. Under these circumstances, it would be inappropriate to eliminate processes from the model on the grounds that they are not important at OSP, or that they are not identifiable from a particular set of historical observations from OSP. We are interested rather in the question of what such a model allows us to infer and predict about OSP and other regions, given generic objective prior information and the limited available observations.

Models with many, poorly identified parameters can be subject to over-tuning and poor predictive performance, especially if parameter estimation procedures are heuristic, and/or are designed to produce a single "optimal" parameter set. The BHM framework and inference procedures used here provide protection against over-tuning. The posterior distribution yields

samples from the full range of possible parameters and states, conditional on priors and observations, and it therefore provides a realistic picture of the effects of equifinality (Von Bertalanffy 1969, Beven and Binley 1992) on model hindcasts and predictions. The performance of the posterior for the long-term forecast for OSP (Fig. 10) supports this conclusion.

Emerging observing systems promise much richer data sets than in the past. New automated in situ and remote sensors can provide data for more variables with much higher temporal resolution and/or spatial coverage. The twin experiment with daily observations presented here provides a hint of what we might expect from such improved observing systems. Data assimilating models are increasingly being used to assess the information value of alternative observing system designs, as part of so-called observing system simulation experiments (OSSEs; e.g., Masutani et al. 2010). The twin experiments presented were intended primarily as a check on the consistency and performance of the inference methods; an OSSE would require careful attention to observing system elements and costs, and the use of replicate experiments. We anticipate using the BHM framework to build OSSEs. Oceanographic field studies often include local in situ or ship-board experiments that effectively measure the instantaneous values of community ecophysiological properties. The model formulation proposed here offers the opportunity to integrate these measurements with standard observations of state variables (biomass) within a consistent and rigorous inference framework. We see this as an interesting direction for further research using both OSSEs and real observations.

We recognize that, in order to fully exploit the potential for OSSEs, and for hindcasting and forecasting more generally, it will be necessary to extend our approach from the zero-dimensional box model considered here to spatially resolved models, including both one-dimensional vertical mixing models (e.g., Mattern et al. 2010) and three-dimensional circulation models (cf. Gregg 2008). The adoption of spatially resolved models would avoid the ambiguity about spatial scales inherent in the box model and allow a more rigorous treatment of spatial sampling errors. We do not foresee major conceptual problems in extending the formulation to spatially resolved models, but Bayesian inference in these models will involve formidable computational challenges, and may require the development of effective approximate techniques.

We believe that the example presented here delivers at least in part on the promise described by Berliner (2003) and Cressie et al. (2009) of BHM as a self-consistent probabilistic framework that integrates statistical and mechanistic process models. The specific process model developed here shows promise as a basis for applications for many local and regional aquatic BGC applications. We hope that some of the methods developed here,

including the use of stochastic processes for aggregate community properties, will find broader application.

#### ACKNOWLEDGMENTS

We are indebted to two anonymous referees for their careful and constructive reviews of an earlier version of this paper, which led to significant improvements both in clarity and presentation. We are grateful for helpful comments and constructive criticism at various stages of this research from John Taylor, Richard Matear, Yong Song, Nugzar Margvelashvili, and David Clifford. We are also grateful for the input of many colleagues in helping us formulate our thinking, and in particular we thank Karen Wild-Allen for her time and patience in helping the more statistically inclined members of the project team understand more of BGC modeling and its applications. We have benefited from Peter Oke's many insights from a data assimilation perspective. We thank Arnaud Doucet for his insights on sequential Monte Carlo, which led to a much-improved inference algorithm. Bronwyn Harch was instrumental in starting the project leading to the work reported here, and we thank her for her enthusiasm and support. Finally, E. P. Campbell is particularly grateful to Mavis Dias for her help, enthusiastically given, at a number stages in the preparation of this paper.

#### LITERATURE CITED

- Allen, J. I., J. T. Holt, J. Blackford, and R. Proctor. 2007. Error quantification of a high-resolution coupled hydrodynamic-ecosystem coastal-ocean model: part 2. Chlorophyll-a, nutrients and SPM. *Journal of Marine Systems* 68:381–404.
- Andrieu, C., A. Doucet, and R. Holenstein. 2010. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B* 72:269–302.
- Arhonditsis, G. B., G. Perhar, W. Zhang, E. Massos, M. Shi, and A. Das. 2008. Addressing equifinality and uncertainty in eutrophication models. *Water Resources Research* 44:W01420.
- Arhonditsis, G. B., S. S. Qian, C. A. Stow, E. Lamon, and K. H. Reckhow. 2007. Eutrophication risk assessment using Bayesian calibration of process-based models: application to a mesotrophic lake. *Ecological Modelling* 208:215–229.
- Baker, D. F., S. C. Doney, and D. S. Schimel. 2006. Variational data assimilation for atmospheric CO<sub>2</sub>. *Tellus, Series B* 58:359–365.
- Barber, J. J., and A. E. Gelfand. 2007. Hierarchical spatial modeling for estimation of population size. *Environmental and Ecological Statistics* 14:193–205.
- Berliner, L. M. 2003. Physical-statistical modeling in geophysics. *Journal of Geophysical Research* 108. dx.doi.org/10.1029/2002JD002865
- Berliner, L. M., C. K. Wikle, and N. Cressie. 2000. Long-lead prediction of Pacific SSTs via Bayesian dynamic modeling. *Journal of Climate* 13:3953–3968.
- Beven, K., and A. Binley. 1992. The future of distributed models: model calibration and uncertainty prediction. *Hydrological Processes* 6:279–298.
- Bishop, J., S. Calvert, and M. Soon. 1999. Spatial and temporal variability of POC in the northeast Subarctic Pacific. *Deep-Sea Research Part II—Topical Studies In Oceanography* 46:2699–2733.
- Briers, M., A. Doucet, and S. Maskell. 2010. Smoothing algorithms for state-space models. *Annals of the Institute of Statistical Mathematics* 62:61–89.
- Calder, C. A., M. Lavine, P. Mueller, and J. S. Clark. 2003. Incorporating multiple sources of stochasticity in population dynamic models. *Ecology* 84:1395–1402.
- Campbell, J. W. 1995. The lognormal-distribution as a model for biooptical variability in the sea. *Journal of Geophysical Research-Oceans* 100:13237–13254.
- Clark, J. S., and O. N. Bjornstaad. 2004. Population time series: process variability, observation errors, missing values, lags, and hidden states. *Ecology* 85:3140–3150.
- Clark, J. S., and A. E. Gelfand. 2006. A future for models and data in ecology. *Trends in Ecology and Evolution* 21:375–380.
- Cox, D., and D. Hinkley. 1986. *Theoretical statistics*. Chapman and Hall, London UK.
- Cressie, N., C. A. Calder, J. S. Clark, J. M. Ver Hoef, and C. K. Wikle. 2009. Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. *Ecological Applications* 19:553–570.
- Cressie, N., and C. K. Wikle. 2011. *Statistics for spatio-temporal data*. Wiley, Hoboken, New Jersey, USA.
- Denman, K. 2003. Modelling planktonic ecosystems: parameterizing complexity. *Progress in Oceanography* 57:429–452.
- Denman, K., and M. Pena. 1999. A coupled 1-D biological/physical model of the Northeast SubArctic Pacific Ocean with iron limitation. *Deep Sea Research Part II: Topical Studies In Oceanography* 46:2877–2908.
- Dowd, M. 2006. A sequential Monte Carlo approach for marine ecological prediction. *Environmetrics* 17:435–455.
- Dowd, M. 2007. Bayesian statistical data assimilation for ecosystem models using Markov Chain Monte Carlo. *Journal of Marine Systems* 68:439–456.
- Dowd, M. 2011. Estimating parameters for a stochastic dynamic marine ecological system. *Environmetrics* 22:501–515.
- Dowd, M., and R. Meyer. 2003. A Bayesian approach to the ecosystem inverse problem. *Ecological Modelling* 168:39–55.
- Fearnhead, P., D. Wyncoll, and J. Tawn. 2010. A sequential smoothing algorithm with linear computational cost. *Biometrika* 97:447–464.
- Geider, R., H. MacIntyre, and T. Kana. 1998. A dynamic regulatory model of phytoplanktonic acclimation to light, nutrients, and temperature. *Limnology and Oceanography* 43:679–694.
- Gregg, W. W. 2008. Assimilation of SeaWiFS ocean chlorophyll data into a three-dimensional global ocean model. *Journal of Marine Systems* 69:205–225.
- Harrison, P. 2002. Station Papa time series: Insights into ecosystem dynamics. *Journal of Oceanography* 58:259–264.
- Holling, C. S. 1965. The functional response of predators to prey density and its role in mimicry and population regulation. *Memoirs of the Entomology Society Of Canada* 45:60.
- Hood, R. R., et al. 2006. Pelagic functional group modeling: progress, challenges and prospects. *Deep Sea Research Part II: Topical Studies in Oceanography* 53:459–512.
- Hooten, M. B., C. K. Wikle, R. M. Dorazio, and J. A. Royle. 2007. Hierarchical spatiotemporal matrix models for characterizing invasions. *Biometrics* 63:558–567.
- Jones, E., J. Parslow, and L. Murray. 2010. A Bayesian approach to state and parameter estimation in a phytoplankton-zooplankton model. *Australian Meteorological and Oceanographic Journal* 59:7–16.
- Link, W. A., E. Cam, J. D. Nichols, and E. G. Cooch. 2001. Of BUGS and birds: Markov chain Monte Carlo for hierarchical modelling in wildlife research. *Journal of Wildlife Management* 66:277–291.
- Luo, Y., K. Ogle, C. Tucker, S. Fei, C. Gao, S. LaDeau, J. S. Clark, and D. S. Schimel. 2011. Ecological forecasting and data assimilation in a data-rich era. *Ecological Applications* 21:1429–1442.
- Masutani, M., J. S. Woollen, S. J. Lord, G. D. Emmitt, T. J. Kleespies, S. A. Wood, S. Greco, H. Sun, J. Terry, V. Kapoor, R. Treadon, and K. A. Campana. 2010. Observing system simulation experiments at the National Centers for Environmental Prediction. *Journal of Geophysical Research-Atmospheres* 115:D07101.

- Matear, R. 1995. Parameter optimization and analysis of ecosystem models using simulated annealing: a case study at Station P. *Journal of Marine Research* 53:571–607.
- Mattern, J., M. Dowd, and K. Fennel. 2010. Sequential data assimilation applied to a physical-biological model for the Bermuda Atlantic Time Series station. *Journal of Marine Systems* 79:144–156.
- Maunder, M. N. 2004. Population viability analysis based on combining integrated, Bayesian, and hierarchical analysis. *Acta Oecologica* 26:85–94.
- Ogle, K., R. L. Wolpert, and J. F. Reynolds. 2004. Reconstructing plant root area and water uptake profiles. *Ecology* 85:1967–1978.
- Pena, M., and S. Bograd. 2007. Time series of the northeast Pacific. *Progress in Oceanography* 75:115–119.
- Sacks, W. J., D. S. Schimel, R. K. Monson, and B. H. Braswell. 2006. Model-data synthesis of diurnal and seasonal CO<sub>2</sub> fluxes at Niwot Ridge, Colorado. *Global Change Biology* 12:240–259.
- Steele, J. 1976. Role of predation in ecosystem models. *Marine Biology* 35:9–11.
- Steele, J., and E. Henderson. 1992. The role of predation in plankton models. *Journal of Plankton Research* 14:157–172.
- Tierney, L. 1994. Markov chains for exploring posterior distributions. *Annals of Statistics* 22:1701–1762.
- Ver Hoef, J. M. 1996. Parametric empirical Bayes methods for ecological applications. *Ecological Applications* 6:1047–1055.
- Von Bertalanffy, L. 1969. *General system theory: foundations, development, applications*. George Braziller, New York, New York, USA.
- Whitney, F., and H. Freeland. 1999. Variability in upper-ocean water properties in the NE Pacific Ocean. *Deep Sea Research Part II: Topical Studies in Oceanography* 46:2351–2370.
- Wikle, C. K. 2003. Hierarchical Bayesian models for predicting the spread of ecological processes. *Ecology* 84:1382–1394.
- Wild-Allen, K., M. Herzfeld, P. A. Thompson, U. Rosebrock, J. Parslow, and J. K. Volkman. 2010. Applied coastal biogeochemical modelling to quantify the environmental impact of fish farm nutrients and inform managers. *Journal of Marine Systems* 81:134–147.
- Xu, T., L. White, D. Hui, and Y. Luo. 2006. Probabilistic inversion of a terrestrial ecosystem model: analysis of uncertainty in parameter estimation and model prediction. *Global Biogeochemical Cycles*:GB2007.
- Zhang, W., and G. B. Arhonditsis. 2009. A Bayesian hierarchical framework for calibrating aquatic biogeochemical models. *Ecological Modelling* 220:2142–2161.
- Zobitz, J. M., S. P. Burns, J. Ogee, M. Reichstein, and R. Bowling. 2007. Partitioning net ecosystem exchange of CO<sub>2</sub>: a comparison of a Bayesian/isotope approach to environmental regression methods. *Journal of Geophysical Research—Biogeosciences* 112:G03013.
- Zobitz, J. M., A. R. Desai, D. J. P. Moore, and M. A. Chadwick. 2011. A primer for data assimilation with ecological models using Markov Chain Monte Carlo (MCMC). *Oecologia* 167:599–611.
- Zobitz, J. M., D. J. P. Moore, W. J. Sacks, R. K. Monson, D. R. Bowling, and D. S. Schimel. 2008. Integration of process-based soil respiration models with whole-ecosystem CO<sub>2</sub> measurements. *Ecosystems* 11:250–269.

## SUPPLEMENTAL MATERIAL

### Appendix A

Detailed information on the process model developed and used in this paper, including derivation of the prior distributions for the process model parameters ([Ecological Archives A023-033-A1](#)).

### Appendix B

Derivation of the autoregressive processes used in the paper ([Ecological Archives A023-033-A2](#)).

### Appendix C

Some technical details on the particle Markov chain Monte Carlo (PMCMC) sampling approach used to conduct statistical inference ([Ecological Archives A023-033-A3](#)).

### Appendix D

Supplementary material describing the forcing data used in the case study ([Ecological Archives A023-033-A4](#)).