CITATION

Milliff, R.F., J. Fiechter, W.B. Leeds, R. Herbei, C.K. Wikle, M.B. Hooten, A.M. Moore, T.M. Powell, and J. Brown. 2013. Uncertainty management in coupled physical-biological lower trophic level ocean ecosystem models. *Oceanography* 26(4):98–115, http://dx.doi.org/10.5670/ oceanog.2013.78.

DOI

http://dx.doi.org/10.5670/oceanog.2013.78

Uncertainty Management

in Coupled Physical-Biological Lower Trophic Level Ocean Ecosystem Models

A large phytoplankton bloom colored the surface waters green on May 20, 2002. *Image courtesy the SeaWiFS Project, NASA GSFC, and ORBIMAGE*
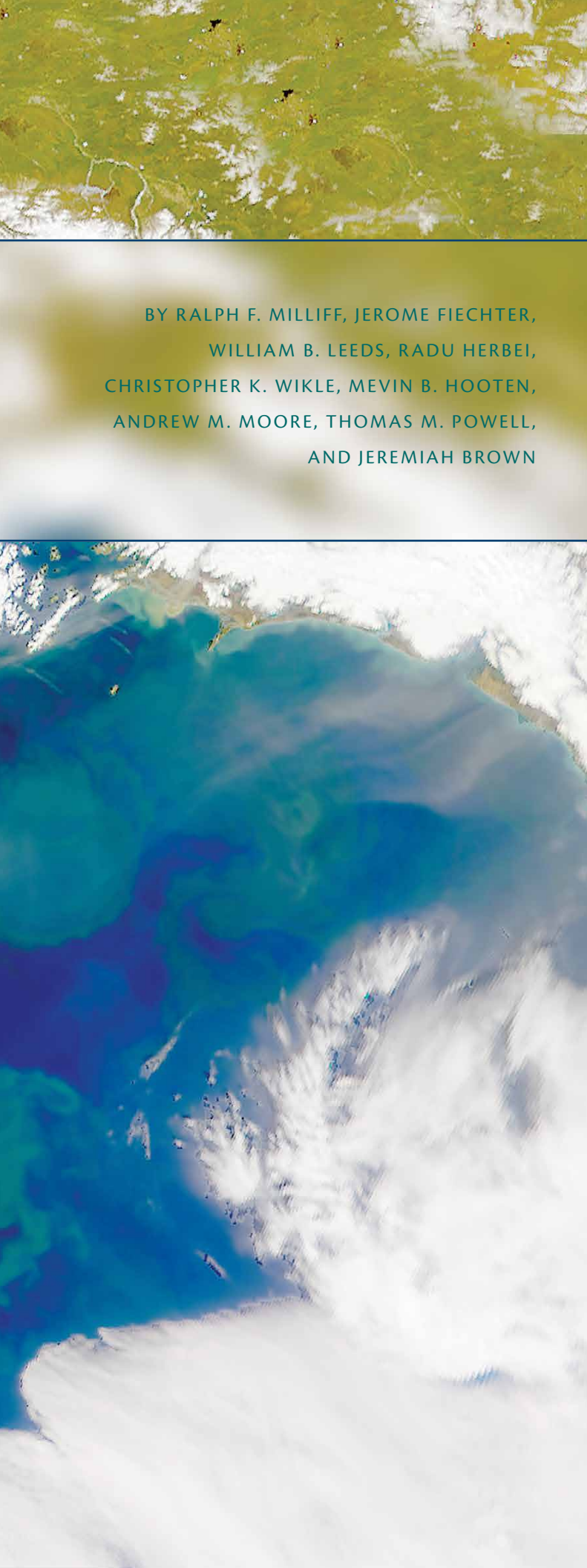
BY RALPH F. MILLIFF, JEROME FIECHTER,

WILLIAM B. LEEDS, RADU HERBEI,

CHRISTOPHER K. WIKLE, MEVIN B. HOOTEN,

ANDREW M. MOORE, THOMAS M. POWELL,

AND JEREMIAH BROWN

ABSTRACT. Lower trophic level (LTL) ocean ecosystem models are important tools for understanding ocean biogeochemical variability and its role in Earth's climate system. These models are often replete with parameters that cannot be well constrained by the sparse observational data available. LTL ocean ecosystem model parameter estimation is examined from a probabilistic perspective, using a Bayesian hierarchical model (BHM), in the coastal Gulf of Alaska (CGOA) domain that benefits from ocean station observations obtained in repeated US GLOBEC cruises. Data entering the BHM include daily average SeaWiFS satellite estimates of surface chlorophyll and GLOBEC observations of nutrient and phytoplankton profiles at inner and outer shelf stations on the Seward Line. The final form of the BHM process model component is comprised of a discrete version of the Nutrient-Phytoplankton-Zooplankton-Detritus LTL ecosystem model equations augmented to address iron limitation in the CGOA (i.e., NPZDFe), and including a vertical diffusion term to constrain the timing of the phytoplankton bloom in spring.

Even in the relatively data-rich GLOBEC context, parameter estimation in the BHM requires guidance from a suite of calculations in a coupled physical-biological deterministic model—the Regional Ocean Model System coupled to an NPZDFe component (ROMS-NPZDFe). ROMS-NPZDFe simulations are used to: (1) validate the BHM formulation, (2) separate BHM limitations due to sampling from those due to LTL model approximations, and (3) obtain output distributions for zooplankton grazing rate and phytoplankton nutrient uptake rate using GLOBEC and SeaWiFS data for 2001. Uncertainty is evident from the spreads in output distributions for model parameters in the BHM. Experiments driven by simulated data from ROMS-NPZDFe helped to optimize the utility of GLOBEC observations for LTL ocean ecosystem model parameter estimation, given ever-present uncertainty issues.

The ROMS-NPZDFe simulations are also used to build Bayesian statistical models as surrogates for the deterministic model. Two applications are briefly described. One estimates output distributions for selected ocean ecosystem parameters while accounting for spatial variability across the GLOBEC stations in the CGOA. A second application assimilates SeaWiFS data and simulated data from a ROMS-NPZDFe control run for 2002 to estimate complete fields of surface phytoplankton concentration, with associated spatial and temporal uncertainties.

INTRODUCTION

Sources of uncertainty in ocean ecosystem models arise from approximations necessary to reduce complexity and represent essential ecosystem processes in an aggregate sense.

Often, the ocean ecosystem model parameters are poorly known or are obtained from experiments, times, and locations that are not specific to the ecosystem under study. Moreover, the model parameters are correlated in some instances, and in almost all cases, ocean ecosystem model parameters do not benefit from an abundance of data. Uncertainty also arises in the sparse observations of ecosystem variables, both in terms of measurement error and errors of representativeness. The underdetermination problem[1] in parameter specification (e.g., Ward et al., 2010) is an inherent issue in ocean ecosystem model development and interpretation.

The mismatch between sparse information from imperfect observations versus many correlated and imperfectly defined parameters is well known to ocean ecosystem modelers. It is the topic of recurring workshops in marine biogeochemical modeling (e.g., see special issue of *Journal of Marine Systems*, vol. 80, March 2010), and it has been qualified and quantified in a pioneering feasibility demonstration by Harmon and Challenor (1997), and more recently in papers by Friedrichs and co-workers (e.g., 2007, 2009; Ward et al., 2010), Dowd (e.g., 2007, 2011), and others (e.g., Malve et al., 2007, Margvelashvili and Campbell, 2012; Parslow et al., 2013).

In the coastal Gulf of Alaska (CGOA), US GLOBEC observations provide a

focus on the lower trophic level (LTL) ecosystem response to environmental forcings. Underdetermination issues complicate LTL ocean ecosystem model parameter estimation efforts in the CGOA. Nonetheless, qualification and quantification of underdetermination and ecosystem model parameter estimations have been achieved. This paper recounts the limitations and success in the CGOA ocean ecosystem parameter estimation process. The story involves interplay between deterministic and probabilistic approaches that provides an update on methodological tools for ocean ecosystem model parameter estimation, given underdetermination.

Physical-Biological Setting in the CGOA

General circulation features of the CGOA (Figure 1) include the Alaska Current, entering the domain from the northeast along a narrow shelf offshore of Sitka and Yakutat, and the Alaska Stream, exiting the domain from northeast to southwest along the broad shelf offshore of the Kenai Peninsula, Kodiak Island, and the Shumagin Islands. Important synoptic-scale circulation features include Yakutat eddies that propagate along the shelf break in the

direction of the Alaska Current and Alaska Stream for several weeks at a time, driving exchanges of shelf and basin waters with important biological implications (Brown and Fiechter, 2012; Fiechter and Moore, 2012).

The LTL ecosystem characteristics in the CGOA consist of shelf, shelf break, and ocean basin regimes. The shelf regime is iron-rich due to river sources and resuspension of bottom sediments. Spring bloom dynamics comprise the principal driver for primary production on the shelf. A weaker bloom also occurs in fall in most years (see Fiechter, 2012, and references therein). With increasing distance off the shelf, phytoplankton abundance is increasingly limited by iron availability such that the basin waters are a high-nutrient low-chlorophyll (HNLC) regime. The shelf and basin regimes are affected by Yakutat eddies that propagate slowly along the shelf break and transport iron-rich shelf waters offshore and nutrient-rich basin waters onshore.

Bayesian Hierarchical Modeling

We adopt the Bayesian hierarchical model (BHM) approach to parameter estimation for the LTL ocean ecosystem model in the CGOA for at least three reasons. First, BHM is a probabilistic

---

[1] The Bayesian Hierarchical Model approach adopts a specialized language to convey specific meanings. Some of these terms are defined in Box 1 for readers unfamiliar with this approach.

Ralph F. Milliff (*milliff@colorado.edu*) is Senior Research Associate, Cooperative Institute for Research in Environmental Science, University of Colorado, Boulder, CO, USA. **Jerome Fiechter** is Assistant Researcher, Institute of Marine Sciences, University of California, Santa Cruz, CA, USA. **William B. Leeds** is Postdoctoral Fellow, Department of Statistics and Department of Geophysical Sciences, University of Chicago, Chicago, IL, USA. **Radu Herbei** is Associate Professor, Department of Statistics, Ohio State University, Columbus, OH, USA. **Christopher K. Wikle** is Professor, Department of Statistics, University of Missouri, Columbia, MO, USA. **Mevin B. Hooten** is Associate Professor, Department of Fish, Wildlife and Conservation Biology and Department of Statistics, Colorado State University, Fort Collins, CO, USA. **Andrew M. Moore** is Professor, Ocean Sciences Department, University of California, Santa Cruz, CA, USA. **Thomas M. Powell** is Professor Emeritus, Department of Integrative Biology, University of California, Berkeley, CA, USA. **Jeremiah Brown** is Research Scientist, Principal Scientific Group, Fishers, IN, USA.

BOX 1. GLOSSARY AND BAYESIAN HIERARCHICAL MODEL DEFINITIONS

STATE VARIABLES AND PARAMETERS: Components of the process model (see below) for the system of interest. State, or dependent, variables are the unknowns of interest and parameters are terms necessary in defining the evolution of the state variables. The NPZDFe equations in Box 2 express the state variable evolutions, and the parameters are identified in Box 3. We are extending the BHM methodology here to see what *parameters* can be estimated in this system, given sparse data for only a few *state variables*.

UNDERDETERMINATION: A mismatch in the complexity of a model for a given system versus the volume, variety, and quality of data available to inform parameters of the model. A model with a large number and wide variety of parameters but sparse data is typically underdetermined as available observations will inform only a few parameters.

RANDOM VARIABLE: A state variable or a parameter endowed with a probability distribution. The prescribed probability distribution need not be uniform or "white." The distribution of a random variable A is denoted by [A]. The joint distribution of two random variables A,B is [A,B], and the conditional distribution of A given a specific value of B is [A|B].

INDENTIFIABILITY: A random variable is said to be identifiable if its posterior distribution is informed by the available data—that is, it is concentrated around a certain value. Identifiability and the *Bayesian learning* property described in the Introduction are closely related. Identifiability issues arise in cases of underdetermination (i.e., when the data are insufficient to inform some parameters and no learning occurs).

ENSEMBLE CALCULATIONS: A related collection of deterministic model runs wherein uncertain parameters are perturbed to create a unique set for each deterministic model run. The outcome of each run is referred to as an ensemble member. Depending on the efficacy of the perturbations, it is hoped that the ensemble response is a measure of the sensitivity to the perturbed parameters.

ERRORS OF REPRESENTATIVENESS: Errors and/or uncertainty stemming from different properties and information content inherent in point observations *vs.* discrete model grid points and/or time steps.

EMULATOR: A simple and computationally efficient model that replaces the deterministic model step in quantifying state variable dependences on parameter values (e.g., as in *Ensemble calculations*, above). We demonstrate a Bayesian emulator wherein the P field is emulated using a singular-value decomposition (SVD) with the right singular vectors treated as random variables. A 50-member ensemble calculation using the ROMS-NPZDFe deterministic model was used in the SVD to build the Bayesian emulator.

DISTRIBUTION HIERARCHY: A set of probability distributions wherein random parameters in one distribution depend on specific realizations of parameters from another distribution (i.e., forming linkages or a "chain" of distributions to be sampled at each level of the hierarchy). The hierarchy is designed to isolate prescriptions for variables and parameters at the lowest levels, leading to distributions for the process and parameters of interest, at the highest levels, all according to Bayes Theorem:

$$[X, \theta_d, \theta_p \,|\, Y] = \frac{[Y|X, \theta_d][X|\theta_p][\theta_d][\theta_p]}{\iint [Y|X, \theta_d][X|\theta_p][\theta_{d,p}] \, dX \, d\theta_{d,p}}$$ (B1.1)

COMPONENTS OF A BHM

DATA STAGE DISTRIBUTION: The first conditional distribution in the numerator (B1.1; i.e., $[Y\,|\,X, \theta_d]$). The data stage distribution quantifies uncertainty in observations Y, or *data stage inputs*. For example, data stage uncertainties arise from measurement error, measurement resolutions in space and time, and/or representativeness. The parameters describing these uncertainties are denoted by θ_d in (B1.1). We use surface chlorophyll retrievals from SeaWiFS and station N and P profiles from the GLOBEC inner and outer shelf stations on the Seward Line (Figure 1) as data stage inputs in the BHM described in this paper.

PROCESS MODEL DISTRIBUTION: The second conditional distribution in the numerator (B1.1; i.e., $[X\,|\,\theta_p]$). The process model distribution codifies the ecosystem model and allows for uncertainty in the model approximation. The equations in Box 2 have been discretized and specific parameters are treated as random variables (i.e., at a lower level of the BHM hierarchy, see *Distribution Hierarchy* above).

PARAMETER DISTRIBUTIONS: The distributions for parameters included in the data stage and process model distributions in the numerator of (B1.1; i.e., $[\theta_d]$ and $[\theta_p]$, respectively). The $[\theta_d]$ include error model parameters for SeaWiFS observations and profile data from GLOBEC stations. The $[\theta_p]$ are indicated in Box 3. We treated six and then two parameters of the NPZDFe system as random in the BHM implementation described here.

POSTERIOR DISTRIBUTION: The output distribution of interest for state variables (N, P, Z, D and Fe) and parameters $[\theta_d]$ and $[\theta_p]$, i.e., the left-hand side of (B1.1). In the CGOA ecosystem BHM implementation, the focus is on posterior distributions for parameters $[\theta_p]$. If the uniform distributions with which random parameters are initialized evolve over the course of the MCMC solution procedure to exhibit modal structures, then Bayesian learning is said to have occurred and the parameters are informed, at least in part, by the *convolution* of the data stage and process model distributions.

NORMALIZER: The denominator in (B1.1) that is an integral over all possible states and parameter values for the process of interest. The normalizer insures that the posterior distribution is a proper probability distribution (i.e., integrates to 1). However, it is not tractable in large state-space systems like our lower trophic level ocean ecosystem (Boxes 2 and 3). This gives rise to the Markov Chain Monte Carlo (MCMC) methods described below from which estimates of the posterior distribution are obtained, realization by realization.

MARKOV CHAIN MONTE CARLO: The generic name for the algorithm that simulates samples from the posterior distribution. The simplest and most common specific algorithm in the MCMC family is the *Gibbs Sampler*. In this case, the full conditional distributions are sampled sequentially, with updates from previous samples inserted as they are available. In instances where a full conditional distribution is difficult (i.e., because a normalizer is not tractable), a more expensive (less efficient) simulation of the "draw-accept-reject" type is the Metropolis-Hastings algorithm (see Box 4).
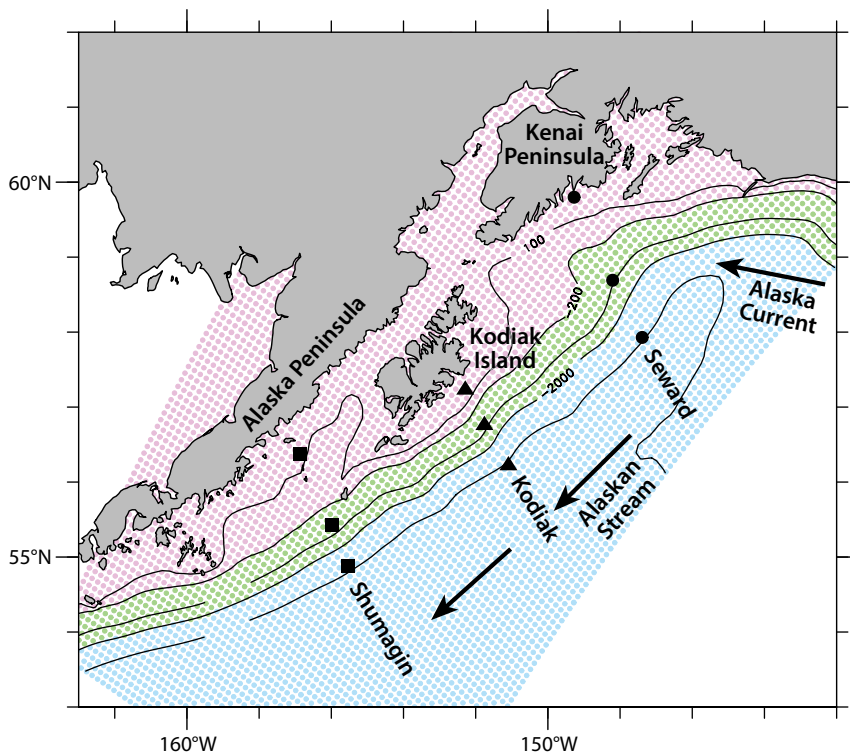
Figure 1. Coastal Gulf of Alaska domain (adapted from Leeds et al., 2012). Circles indicate inner shelf, outer shelf, and offshore stations along the GLOBEC Seward Line. For the purpose of generating a "forest" of 1-D BHM and comparing ecosystem properties, fictitious lines off Kodiak and Shumagin Islands replicating the Seward Line were also considered (squares and triangles). The shelf break is indicated by the 200 m, 1,000 m, and 2,000 m isobaths. Directions for the general circulation features, Alaska Current and Alaskan Stream, are indicated. The Shumagin Islands are located south of the Alaska Peninsula at about 160°W, 55°N. Yakutat, AK, is just out of the figure to the west.

modeling approach wherein the inputs and outputs are probability distributions (Box 1). Parameter estimation is inherently uncertain, and the shapes and spreads in the posterior distributions we obtain from BHM will characterize and quantify this uncertainty. Second, in estimating posterior distributions of interest, the BHM methodology combines probability distributions for (see Box 1): data conditioned on parameters (e.g., as arising from measurement error models), processes conditioned on parameters (e.g., as arising from ocean ecosystem model formulations), and parameters. Parameter estimation in the CGOA will make use of GLOBEC observations that provide far better than typical coverage of biological variables. As such, the ocean

ecosystem parameter estimation in the CGOA quantifies the case for reducing uncertainty when ocean ecosystem variables can be observed. Third, we are interested in extending our experience to adapting and applying BHM methodology in realistic geophysical fluids applications (e.g., Royle et al., 1998; Wikle et al., 2001; Berliner et al., 2000, 2003; Cripps et al., 2005; Fox and Wikle, 2005; Song et al., 2007; Milliff et al., 2011) and terrestrial ecosystem applications (Wikle, 2003a; Wikle and Hooten, 2006; Hooten and Wikle, 2007, 2008; Hooten et al., 2007). BHM references specific to the GLOBEC project in the CGOA include Wikle and Hooten (2010) Hooten et al. (2011), Leeds et al. (2012, 2013), and Fiechter et al.(2013). See also review papers by Wikle (2003b)

and Wikle et al. (2013) and the text by Cressie and Wikle (2011).

The LTL ocean ecosystem parameter estimation, in the face of underdetermination, provides a stringent test of the methodology and builds experience in adapting BHM to difficult problems in Earth science. There are at least two useful properties of BHM relevant to the estimation of LTL ocean ecosystem parameters in the CGOA. First, the property of "Bayesian learning," wherein random variables estimated in the posterior distribution exhibit non-uniform (e.g., modal or multimodal) distributions after having started from uniform specifications over a "reasonable" range of values to initiate the Bayesian estimation procedure (Boxes 1 and 3). Again, the shape and spread in the posterior distribution (i.e., the extent of the departure from uniform) is a measure of the uncertainty. Second, the BHM solution procedure exploits information content in data and process models such that unmeasured quantities can learn from measurements of other quantities. This is the property of "borrowing strength" across processes and parameters in the BHM posterior distribution. To the extent that the posterior distributions for parameters of the LTL ocean ecosystem in the CGOA can borrow strength from GLOBEC and other observations of nutrients and phytoplankton, the parameter posterior distributions will depart from initial uniform distributions to more "shaped" posterior distributions, the spread of which will be a measure of parameter uncertainty.

Fiechter et al. (2009, 2013) describe a deterministic version of the LTL ocean ecosystem model adapted to the BHM in the CGOA. The differential equations and parameter definitions for the LTL model are reproduced from

Fiechter et al. (2013) in Boxes 2 and 3. The model includes state variables for dissolved nitrogen, phytoplankton biomass, zooplankton biomass, and detritus (i.e., Nutrient-Phytoplankton-Zooplankton-Detritus [NPZD], with time-dependent abundances given in units of nitrogen concentration), as well as dissolved iron and phytoplankton-associated iron given in terms of iron concentration. Extensions of the traditional NPZD model are incorporated to: (1) address iron limitation effects on primary production in offshore waters of the CGOA (so, an NPZDFe LTL ocean ecosystem model), and (2) include a vertical mixing term parameterized by mixed-layer depth in each of the state variable equations.

Boxes 1 and 4 provide background and definitions sufficient to put BHM in context for the LTL ocean

ecosystem parameter estimation process described here. More formal and complete descriptions are widely available (e.g., see Cressie and Wikle, 2011). In our CGOA application, the process X (Box 1) represents the time-dependent abundances of N, P, Z, D, and two iron (Fe) state variables. Process model parameters θ_p (Box 1) are given initially by a subset of the parameters as identified in Box 3. Data entering the data stage distribution Y (Box 1) are taken from: (a) time-averaged surface chlorophyll retrievals from SeaWiFS data in the CGOA, and (b) nutrient (N) and phytoplankton (P) concentrations from vertical profiles collected onshore and offshore of the shelf break at GLOBEC stations in the CGOA (Figure 1). The data stage parameters θ_d (Box 1) include measurement error estimates for SeaWiFS and GLOBEC station data.

Estimating the Posterior Distribution

According to Bayes Theorem (Box 1), there is at least the hope that, given sufficient data Y, distributions for the NPZDFe model parameters could be updated and available in the posterior distribution. But how much data is sufficient? What kinds of data are optimal? And what are the underlying issues that complicate parameter estimation in the NPZDFe BHM? To put these issues in context, we review the procedures for estimating the posterior distribution in the BHM in Boxes 1 and 4.

The BHM solution procedure doesn't "solve" a system of equations as in Box 2, but rather estimates the posterior distribution (Box 1) for processes and parameters, given the data, via Monte Carlo methods in Markov Chain Monte Carlo (MCMC) algorithms (Boxes 1 and 4).

BOX 2: LOWER TROPHIC LEVEL OCEAN ECOSYSTEM MODEL FOR GLOBEC STATIONS IN THE CGOA

Dissolved nitrogen: $\dfrac{\partial N}{\partial t} = \delta D + \gamma_n GZ - UP + \kappa \dfrac{\partial^2 N}{\partial z^2}$

Phytoplankton: $\dfrac{\partial P}{\partial t} = UP - GZ - \sigma_d P + \kappa \dfrac{\partial^2 P}{\partial z^2}$

Zooplankton: $\dfrac{\partial Z}{\partial t} = (1 - \gamma_n)GZ - \zeta_d Z + \kappa \dfrac{\partial^2 Z}{\partial z^2}$

Detritus: $\dfrac{\partial D}{\partial t} = \sigma_d P + \zeta_d Z - \delta D + w_d \dfrac{\partial D}{\partial z} + \kappa \dfrac{\partial^2 D}{\partial z^2}$

P-associated iron: $\dfrac{\partial F_p}{\partial t} = F_p \left(U - \dfrac{GZ}{P} - p_m \right) + L_{Fe} + \kappa \dfrac{\partial^2 F_p}{\partial z^2}$

Dissolved iron: $\dfrac{\partial F_d}{\partial t} = F_p \left(f_{rem} \left(\dfrac{GZ}{P} + p_m \right) - U \right) - L_{Fe} + \kappa \dfrac{\partial^2 F_d}{\partial z^2}$

Phytoplankton growth rate: $U = \dfrac{R^2}{R^2 + k_{Fe}^2} \dfrac{V_m N}{k_N + N} \dfrac{\alpha I}{\sqrt{V_m^2 + \alpha^2 I^2}}$

Phytoplankton iron uptake: $L_{Fe} = \dfrac{R_0 - R}{t_{Fe}} P[C:N]$

Empirical and realized [Fe:C] ratios: $R_0 = b F_d^a, \quad R = \dfrac{F_p}{P[C:N]}$

Light availability at depth: $I = I_0 \exp \left(k_z z + k_p \int_0^z P(z')dz' \right)$

Zooplankton growth rate: $G = R_m \left(1 - e^{-\Lambda P} \right)$

Vertical mixing term: $\kappa \sim MLD_{ROMS}$ (MLD = Mixed Layer Depth)

The LTL ecosystem model equations express abundance evolutions for nitrogen, phytoplankton, zooplankton, detritus, and iron associated with phytoplankton and dissolved iron (i.e., NPZDFe). The NPZDFe model is adapted from more common NPZD models. Adaptations include terms for iron abundances and diffusion terms associated with seasonal variability at upper ocean mixed-layer depth.

MCMC sampling can be thought of as analogous to descent algorithms that are perhaps more familiar to deterministic modelers (i.e., in the solution of elliptic operators that arise in data assimilation and primitive equation solvers). The dimension of the state and parameter space in the BHM is tied to the number of random variables. The ranges over which perturbations from the initial values are selected bound the iteration space for MCMC sampling. In many instances, the component distributions comprising the BHM are not known in so-called full conditional form and a Metropolis-Hastings (M-H) sampling step (Box 4) enters the MCMC sampling procedure. Efficient sampling occurs when the M-H algorithm explores local and global extrema in the fewest feasible number of iterations, analogous to finding gradients in steepest descent algorithms. Conversely, if random variables are correlated and conditional distributions for the data do not project upon distributions for specific parameters or state variables, the solution surface is smooth and local extrema are very hard to identify and explore efficiently by iteration.

Initial BHM Experiments

The NPZDFe equations in Box 2 include O(20) parameters, far too many to be identified by the relatively sparse data sets from SeaWiFS retrievals and the GLOBEC stations. We started with six random parameters for the NPZDFe process model (i.e., the $\boldsymbol{\theta}_p$). They were: the phytoplankton maximum growth rate (VmNO3), the half-saturation constant for iron (KFeC), the initial slope of the phytoplankton-light utilization curve (PhyIS), the maximum grazing rate for zooplankton consumption of phytoplankton (ZooGR), the remineralization rate for detritus (DetRR), and the fraction of the available iron that is remineralized (FeRR). Initial values for these parameters in the CGOA were taken from the deterministic coupled physical-biological model calculations described by Fiechter et al. (2009). Ranges over which random perturbations to the initial values could be selected in the M-H algorithm were the subjects of experimentation, but sensible estimates were provided by expert opinion and published values.

Boxes 2 and 3 (reproduced from Fiechter et al., 2013) are included to emphasize the departures from classical NPZD formulations. The boxes demonstrate that the parameters treated as random variables in our BHM enter the NPZDFe LTL ocean ecosystem model

BOX 3: PARAMETERS OF THE LTL NPZDFe MODEL FOR COASTAL GULF OF ALASKA

PARAMETER NAME	SYMBOL	VALUE	UNITS
LIGHT			
Light extinction coefficient	k_z	0.067	m^{-1}
Self-shading coefficient	k_p	0.040	m^2 mmolN^{-1}
PHYTOPLANKTON			
Initial slope of P-I curve (PhyIS)	α	0.020	m^2 W^{-1}
Maximum uptake rate (VmNO3)	V_m	0.800	day^{-1}
Nitrogen half-saturation constant	k_N	1.000	mmolN m^{-3}
Half-saturation for [Fe:C] (KFeC)	k_{Fe}	16.900	mmolFe (molC)$^{-1}$
Empirical [Fe:C] power	a	0.600	nondimensional
Empirical [Fe:C] coefficient	b	64.000	(mmolC m^{-3})$^{-1}$
Iron uptake time scale	t_{Fe}	1.000	day
Mortality	σ_d	0.100	day^{-1}
ZOOPLANKTON			
Maximum grazing rate (ZooGR)	R_m	0.400	day^{-1}
Ivlev constant	Λ	0.840	nondimensional
Excretion efficiency	γ_n	0.300	nondimensional
Mortality	ζ_d	0.145	day^{-1}
REMINERALIZATION			
Detritus remin. rate (DetRR)	δ	0.200	day^{-1}
Detritus sinking	w_d	8.000	m day^{-1}
Iron remin. fraction (FeRR)	f_{rem}	0.500	nondimensional

Parameter names, symbols, values, and units for the NPZDFe model. Parameters treated as random variables (see Box 1) in the BHM framework are indicated (i.e., PhyIS, VmNO3, ZooGR, DetRR, KFeC, FeRR).

in a manner that affects phytoplankton abundance. These are sensible choices for the random parameters of the BHM since most of our data stage information will also be related to phytoplankton abundance. However, the correlations among some of these parameters, and the number of parameters relative to the sparse data, result in M-H acceptance rates well below 25% and parameters that are not identifiable by the data. This latter point is crucial in the interpretations of the posterior distribution from the BHM to be described.

ENSEMBLE CALCULATIONS IN A COUPLED PHYSICAL-BIOLOGICAL EXTENSION OF THE REGIONAL OCEAN MODELING SYSTEM IN THE CGOA
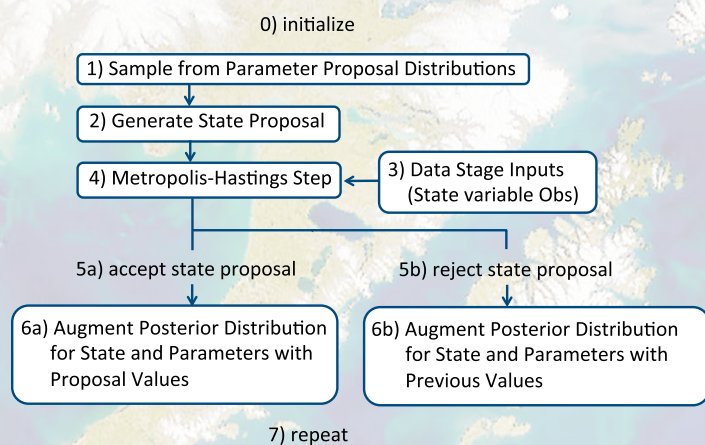
To learn more about the parameter underdetermination in the NPZDFe BHM application in the CGOA, we turned to deterministic tools. An ensemble of forward-model calculations (Fiechter, 2012) was run in the coupled physical-biological model for the CGOA developed by Fiechter and co-workers (Fiechter et al., 2009, 2011; Fiechter and Moore, 2009). The coupled model system for the CGOA is comprised of a physical model component that is the Regional Ocean Modeling System (ROMS; Haidvogel et al., 2008; Moore et al., 2011) and an LTL ecosystem model component that is a six-compartment augmentation of an NPZD model with two additional compartments for iron remineralization and phytoplankton associated iron concentration (i.e., NPZDFe). As noted above, this biological model component is also the basis of the process model in the BHM. Coupled physical-biological forward model calculations in the CGOA with the deterministic system use best estimates for the 19 parameters of the biological model component. Some of these parameters are well known and/or are independent of regional specifics. Others are relatively unknown in the CGOA and based on estimates from other regions (e.g., a California Current System study by Powell et al., 2006). Coupled model simulations successfully reproduce seasonal variability in LTL ecosystem response (Fiechter et al., 2009) as well as signals associated with synoptic eddies in the CGOA (Brown and Fiechter, 2012; Fiechter and Moore, 2012).

Ensemble calculations with the coupled biological-physical model are designed to provide a set of state variable responses to parameter variations. Fiechter (2012) reports ensemble calculations with the coupled biological-physical model for the CGOA for 2001—a year that included strong spring and weaker fall phytoplankton blooms on the shelf and an interaction with a Yakutat eddy off the shelf break in

BOX 4: METROPOLIS-HASTINGS (M-H) ALGORITHM TO SAMPLE FROM "DIFFICULT" DISTRIBUTIONS IN THE GIBBS SAMPLER (MCMC ALGORITHM)

0) initialize

1) Sample from Parameter Proposal Distributions

2) Generate State Proposal

4) Metropolis-Hastings Step ← 3) Data Stage Inputs (State variable Obs)

5a) accept state proposal   5b) reject state proposal

6a) Augment Posterior Distribution for State and Parameters with Proposal Values

6b) Augment Posterior Distribution for State and Parameters with Previous Values

7) repeat

We resort to the M-H algorithm to sample difficult distributions in the Markov Chain of distributions leading to estimates of the posterior distribution. An LTL ocean ecosystem initial condition is given by the set of state variables and parameters in the NPZDFe process model in Boxes 2 and 3. A set of *proposal parameters* is drawn for those parameters treated as random variables in the BHM. The selection of each new set of proposal parameters depends in a tunable way upon the values selected in the previous draw. The proposal parameters and fixed parameters are used to generate a *proposal state*. The proposal state, the state from the previous draw, and observations of the state (identified as data stage inputs in the flowchart) are compared. Conditional distributions (i.e., data stage distributions and parameter distributions) for the proposed state and previous state form a Metropolis ratio. The proposed state is accepted or rejected with a probability depending on the size of this ratio. Acceptance rates of about 25% are necessary for the M-H step implementation to be useful in the MCMC algorithm for the NPZDFe BHM.

summer. Seven biological parameters were varied for each forward model ensemble member according to randomizations, given reasonable ranges around the parameter values from control runs (e.g., Fiechter et al., 2009). The perturbed parameters included: (1) the initial slope of the phytoplankton light utilization curve (PhyIS), (2) the maximum growth rate for phytoplankton (VmNO3), (3) the half-saturation constant for nitrogen (KNO3), (4) the half-saturation constant for iron (KFeC), (5) the zooplankton grazing rate (ZooGR), and (6) remineralization rates for detritus (DetRR) and iron (FeRR). These parameters directly affect the equation for phytoplankton abundance.

The state variable simulations for phytoplankton in each ensemble member are compared with eight-day average surface phytoplankton retrievals from SeaWiFS to identify key biological scenarios for LTL ecosystem evolution

on synoptic and seasonal time scales in the CGOA. The ensemble approach is limited by the costs of the coupled physical-biological calculation for each ensemble member and by the extent to which the biological model parameter space is explored by the random perturbations from control values in a few parameters. Also, the randomized perturbation method does not guarantee that perturbed parameter sets will be biologically or physically self-consistent.

Nonetheless, to identify parameter impacts in subdomains of the CGOA, Fiechter (2012) performed multivariate linear regressions for surface chlorophyll concentration, given terms representing each of the seven biological parameters that were randomly perturbed for each ensemble member. Each regression coefficient (one for each parameter) identifies the relative impact (in a normalized least squares sense) of its associated biological parameter. The parameters

explaining the largest fractions of variance in surface phytoplankton concentration are identified by month and subregion of the CGOA in Figures 2 and 3.

To isolate temporal variability in the leading parameters, the linear regression analysis was spatially averaged over shelf and basin subregions. Figure 2 depicts the normalized monthly average regression coefficient for each parameter for the period March through October 2001 from Fiechter (2012). The left panel shows the correlations on the shelf and the right panel the relative impacts in the CGOA basin. On the shelf, primary production in the spring bloom is consistent with efficient utilization of sunlight and rapid phytoplankton growth (i.e., large amplitude normalized regression coefficients for PhyIS and VmNO3). PhyIS is again important leading into the fall bloom. Zooplankton grazing in summer and fall moderates phytoplankton abundance, as noted by large negative normalized regression coefficients in ZooGR for May to October. While bloom signals for the basin regime are similar in PhyIS, VmNO3, and ZooGR, there are additional important terms having to do with uptake of dissolved iron by phytoplankton cells (KFeC) and iron remineralization (FeRR).

When the spatial averaging is relaxed, point-by-point linear regressions can be used to identify spatial distributions of the dominant parameters in the LTL ecosystem of the CGOA as normalized regression coefficients. Figure 3 depicts maps of dominant biological parameters contributing to phytoplankton abundance for three seasons of 2001: May, July, and September, representing the spring bloom, the summer synoptic eddy season, and the fall bloom.

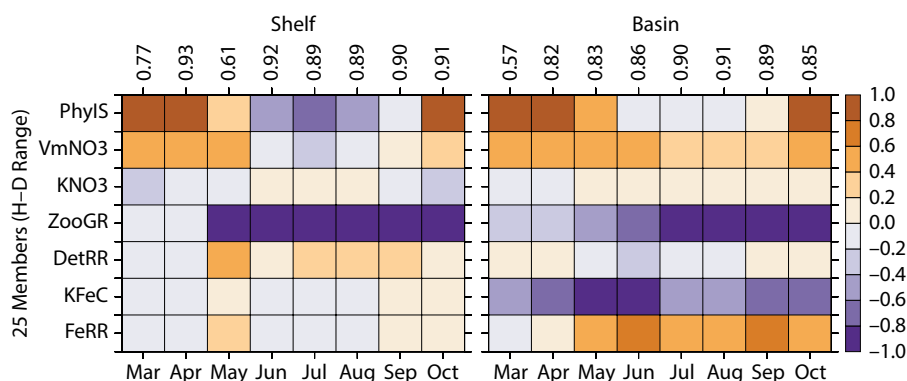The dominant ecosystem process on the shelf in April is the spring bloom,

Figure 2. Normalized regression coefficients for lower trophic level (LTL) ecosystem model parameters (vertical axes) vs. month in 2001 (horizontal axes) for shelf (left) and basin (right) subregions of the CGOA domain. A multivariate linear regression for surface phytoplankton concentration was spatially averaged (shelf and basin) for all ensemble members wherein the LTL ecosystem model parameters were randomly perturbed over the range of half to double of the default values. Regression coefficients are normalized by the largest value regression coefficient for each parameter across all ensemble members. To account for differing units, parameter values were normalized by a control value before fitting the regression. As noted in Fiechter et al. (2013), normalized regression coefficients near 1 mean that the associated parameters have large impacts on phytoplankton concentration and that an increase in the value of the parameter leads to an increase in phytoplankton. A normalized regression coefficient with amplitude near −1 means that the associated parameter is important to phytoplankton concentration, but that an increase in the parameter value leads to a decrease in concentration. Regression fits are quantified by R^2 values for each month atop each column. See Fiechter et al. (2013) for a more detailed discussion.

and the controlling parameter is PhyIS over most of the shelf. Offshore, the ecosystem is iron-limited, and this is reflected in the importance of the half saturation constant for dissolved iron uptake, KeFC. By summer, ZooGR controls the phytoplankton abundance on the shelf in the middle of the domain and across the shelf break into the basin over most of the domain. The light-limitation parameter, PhyIS, is negatively correlated with sustaining primary productivity into summer in nearshore regions (i.e., less-efficient light utilization preserves some nutrients for later in the year). Remineralization of iron (FeRR) plays a dominant role offshore and in the south. There is a hint of nutrients being drawn off the shelf in the north where a spatial patch on the scale of a Yakutat eddy is dominated by the VmNO3 parameter. This corresponds to an eddy location noted in Brown and Fiecther (2012) and Fiecther and Moore (2012). By fall, ZooGR controls phytoplankton abundance from the coast across the

shelf break in the south, while signals of the fall bloom due to DetRR and PhyIS are evident onshore in the north. Dominant parameters in the basin are similar to the summer distributions with FeRR in the south, KeFC in the north, and evidence of onshore-offshore transports along the shelf break where PhyIS and ZooGR are most important.

The ensemble calculations have identified seasonal, synoptic, and subdomain variabilities in dominant parameters of LTL ecosystem dynamics for the CGOA, up to the limits explored in terms of ensemble size and ranges in the subset of biological parameters (i.e., 7 of 19) varied in the experiments. Randomly selected parameter values led to surface phytoplankton estimates that could be compared with eight-day and monthly averages from SeaWiFS retrievals. For the purposes of parameter probability distribution estimation, the ensemble experiments have identified a few key parameters and suggested reasonable ranges to pose as priors for $[\theta_p]$.

THE PARAMETER-ESTIMATION BHM IN THE CGOA

The NPZDFe BHM was implemented at inner and outer shelf locations on the GLOBEC Seward Line off the Kenai Peninsula (Figure 1) during 2001. The NPZDFe process model is unchanged from the earlier implementation and the data stage inputs are taken separately and in combinations from: (1) surface phytoplankton retrievals from daily SeaWiFS data, (2) temporally intermittent in situ observations of nitrate and chlorophyll at the GLOBEC stations, and (3) from coupled physical-biological model output to be used for sensitivity tests and validation experiments. We focus on 2001 because concurrent measurements of nitrate and chlorophyll are available at both the inner and outer shelf GLOBEC stations that year in April, May, and July (Strom et al., 2006). The inner shelf location is representative of nitrate-limited primary production with strong spring and weaker fall phytoplankton blooms. The outer shelf location is offshore of
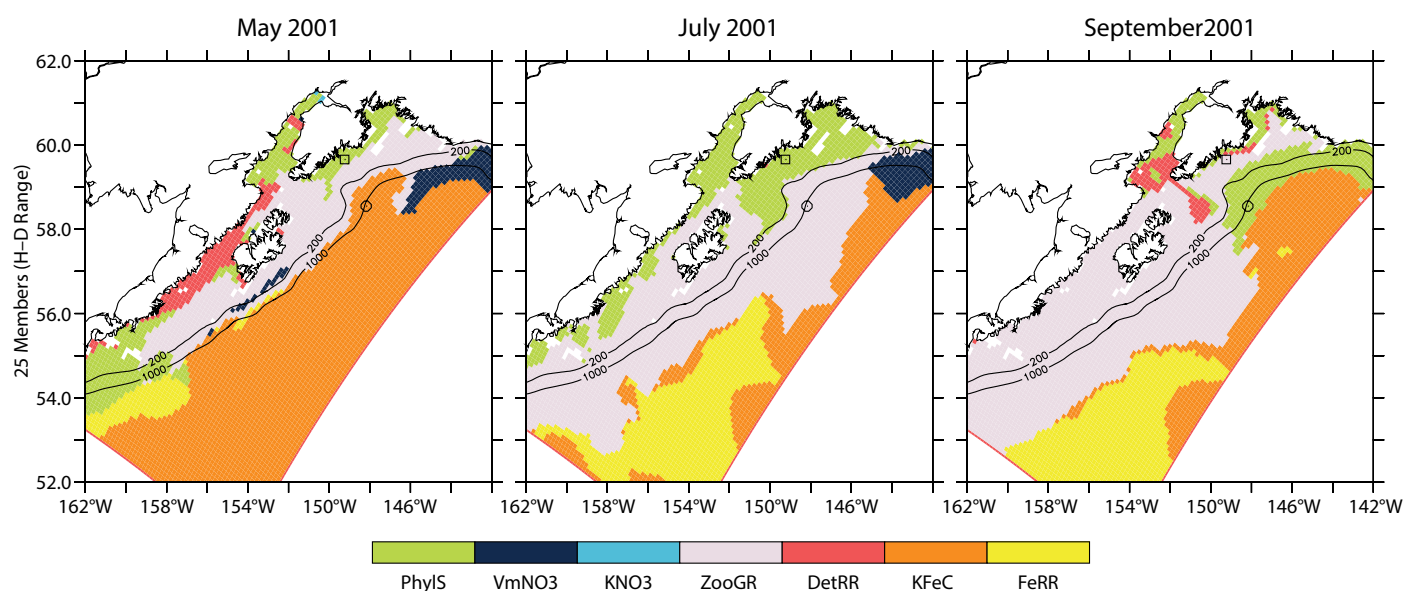

Figure 3. Largest-amplitude normalized regression coefficients for LTL ecosystem model for three seasons in 2001. A multivariate linear regression for surface phytoplankton concentration was computed for each surface location in the CGOA domain, for all ensemble members wherein the LTL ecosystem model parameters were randomly perturbed over the range of half to double the default values. Regression coefficients are normalized by the largest value for each parameter across all ensemble members. *Adapted from Fiechter et al. (2013)*

the shelf break where iron limitation is important and phytoplankton abundance does not exhibit a strong seasonal signal. To demonstrate parameter estimation in the NPZDFe BHM, we highlight a few interpretations and implications from the posterior distribution estimates obtained, with a focus on the inner shelf location. For a more complete analysis, including equal emphasis on posterior distribution estimates for the outer shelf location, see Fiechter et al. (2013).

The parameters VmNO3 and ZooGR were shown to be important in controlling phytoplankton abundance in the ensemble calculations. We treat them as random variables (Box 1) in the refocused BHM experiments. VmNO3 is the most flexible parameter to reflect the full range of potential phytoplankton growth rates. PhyIS was important in setting the onset of the spring bloom, but its functional dependence (i.e., as the coefficient for the light limitation term) is limited to the range [0,1] as are other terms (e.g., Michaelis-Menten nutrient limitation) in the equation governing phytoplankton growth. Also, recall that in the NPZDFe BHM, a vertical mixing term, based on estimates of mixed-layer depth, refines the timing of the spring bloom onset. So, we revisit the NPZDFe BHM with VmNO3 and ZooGR as random parameters, while fixing the other parameters at their default values.

The NPZDFe BHM can be validated in part by replacing in situ and remotely sensed data stage inputs with simulated observations from the control run of the coupled physical-biological model. These are so-called "nearly perfect data experiments." The goal is to reproduce in the posterior distribution of the NPZDFe BHM the default values used in the forward model for the random parameters VmNO3 and ZooGR. The simulated data

in these experiments are "nearly perfect" because there is no explicit account in the one-dimensional BHM for the time-dependent three-dimensional effects of ocean circulation that affects phytoplankton in the forward model in and around the GLOBEC station locations. The nearly perfect data experiments reproduce, with negligible spread, posterior mean values for VmNO3 and ZooGR, thereby validating the implementation of the NPZDFe BHM and indicating that three-dimensional circulation effects on phytoplankton were secondary (Fiechter et al., 2013).

The nearly perfect data sampling was degraded in time and content to study the effects of more realistic data stage inputs and to help interpret posterior distributions from NPZDFe BHM experiments using GLOBEC station data and SeaWiFS data stage inputs. Inferences regarding sampling that emerge from these sensitivity studies (Fiechter et al., 2013) include: (a) it is important to capture data for both the onset (i.e., dominated by phytoplankton growth) and the decay (i.e., controlled by zooplankton grazing) phases of the spring bloom on the shelf, (b) in situ samples of more than one state variable (e.g., chlorophyll and nitrate) usefully constrain posterior distributions of interest, and (c) data stage inputs with widely different space-time properties (e.g., resolution, seasonality, vertical vs. surface biases in coverage) might not be additive in their contributions to refining posterior distribution estimates.

Figure 4 shows estimates of the posterior distributions for VmNO3 (top row) and ZooGR (bottom row) for the inner shelf location when degraded forward model outputs (denoted ROMS-NPZDFe in Fiechter et al., 2013) are used as data stage inputs in the BHM to mimic the

temporal and vertical sampling and data types collected by GLOBEC (left column), SeaWiFS (middle column), and GLOBEC and SeaWiFS combined (right column). Fiechter et al. (2013) show that the lack of a well-defined mode (i.e., large uncertainty) and unrealistic positive bias in VmNO3 for GLOBEC sampling of forward model output is due to the absence of data during the onset of the spring bloom when VmNO3 controls phytoplankton growth. As noted in Fiechter et al. (2013), the spring bloom in the ROMS-NPZDFe calculations occurs slightly earlier than in the real system as measured by GLOBEC.

The uncertainty in VmNO3 is greatly reduced and more realistic values are recovered when using forward model output with SeaWiFS sampling intervals. The same is true for combined GLOBEC and SeaWiFS (right column). SeaWiFS sampling covers the initial phases of the spring bloom in ROMS-NPZDFe, and thus provides sufficient information to reliably estimate VmNO3 in the BHM. In contrast, ZooGR is well estimated, given forward model data stage inputs for all sampling intervals, as zooplankton grazing does not control phytoplankton abundance until after the peak in the spring bloom on the shelf (e.g., see Figure 3).

When the real in situ GLOBEC station data and SeaWiFS remote-sensing data are used in isolation and in combination (Figure 5), the posterior distribution estimates for VmNO3 and ZooGR at the inner shelf location are more complicated. Vertical profiles of nitrate and chlorophyll from the GLOBEC station data are sufficient to estimate posterior mean values for VmNO3 and ZooGR that are close to default values, with very little uncertainty. This suggests that the GLOBEC observations captured both
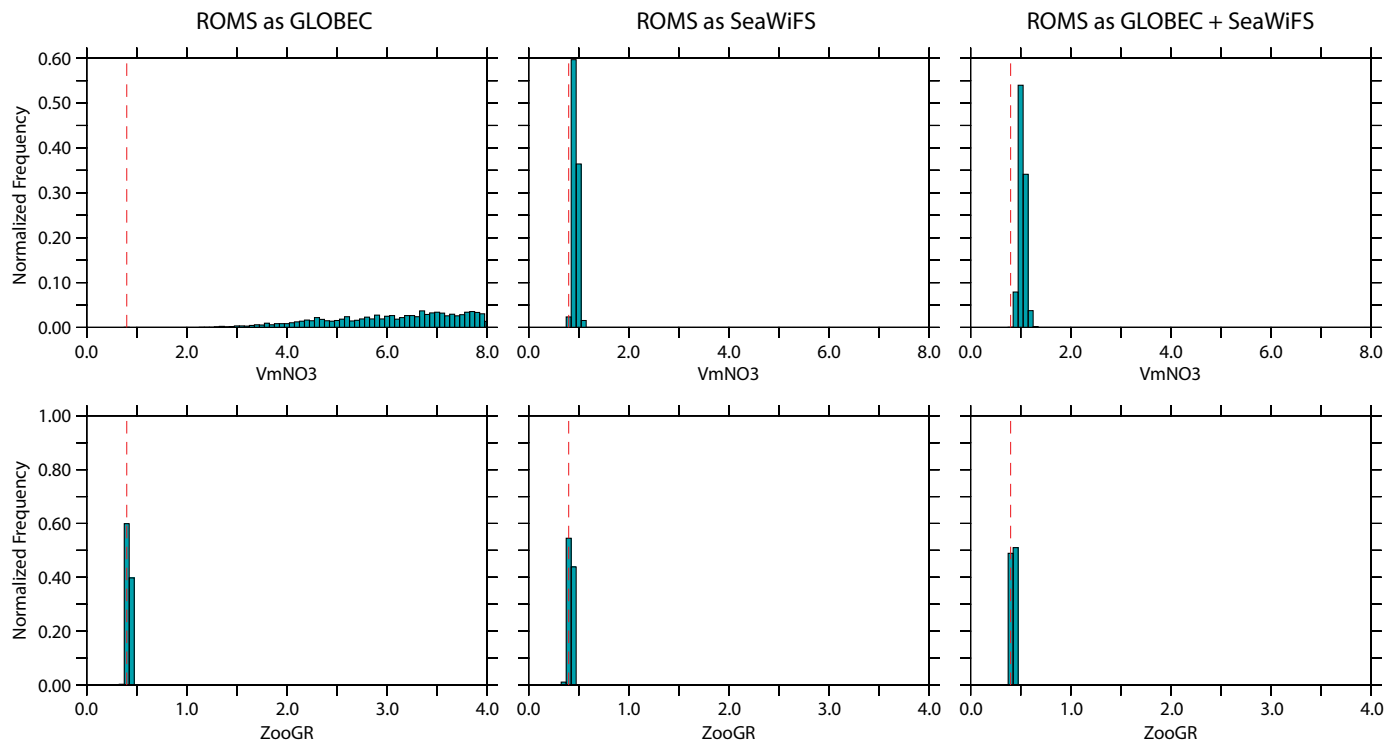
Figure 4. CGOA inner-shelf posterior distributions for phytoplankton growth rate (VmNO3, top) and zooplankton grazing rate (ZooGR, bottom) using ROMS-NPZDFe subsampled as GLOBEC (left), SeaWiFS (center), and both (right) as data stage. Dashed vertical red lines indicate default parameter values (see Box 3). Two thousand realizations of the posterior distribution estimate were used to build these histograms. The frequencies in each bin are divided by the total number of realizations (i.e., 2,000) so that the normalized frequency is the fraction of total realizations (note that this is a different normalization than was used in Fiechter et al., 2013).
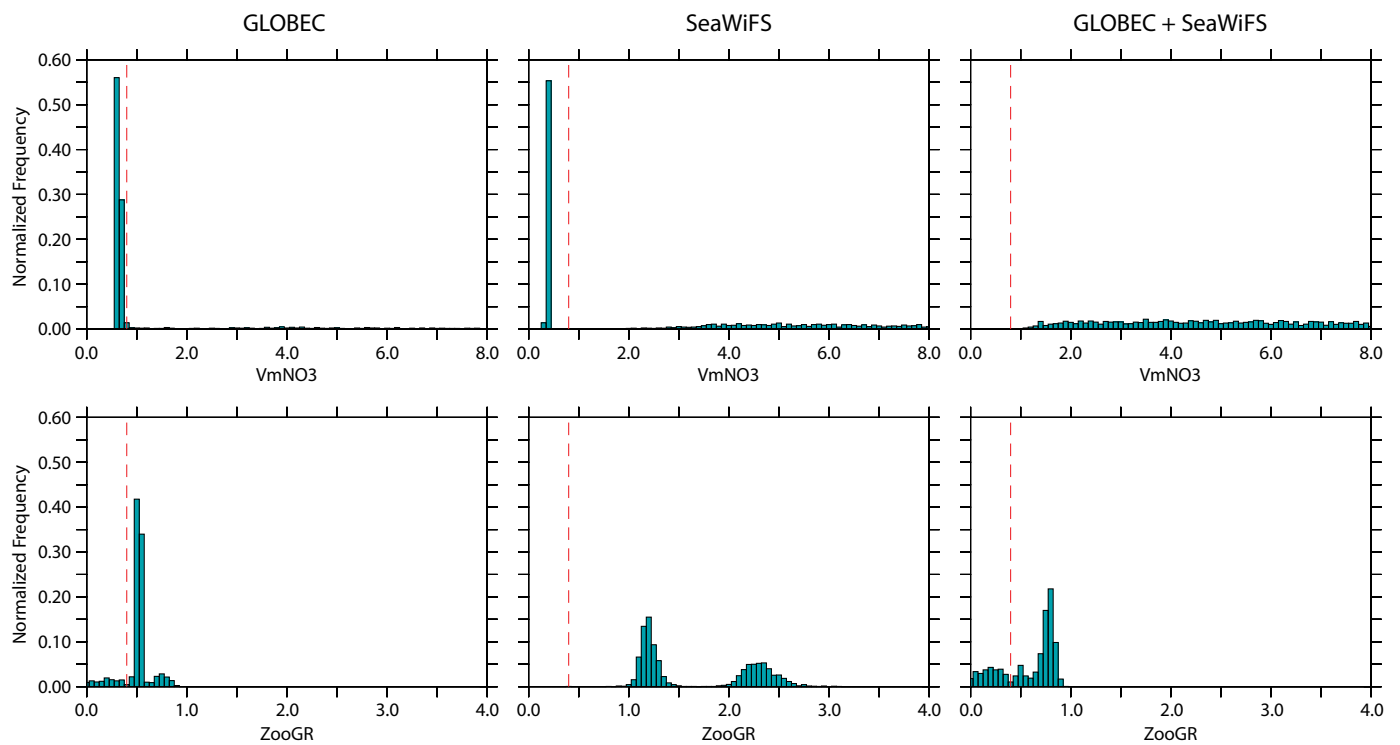

Figure 5. CGOA inner-shelf posterior distributions for phytoplankton growth rate (VmNO3, top) and zooplankton grazing rate (ZooGR, bottom) using observations from GLOBEC (left), SeaWiFS (center), and both (right) as data stage. Dashed vertical red lines indicate default parameter values (see Box 3). Frequency of occurrence is normalized as in Figure 4.

the onset and decay phases of the spring bloom in the CGOA in 2001. However, the posterior distribution estimates using SeaWiFS-only data stage inputs are farther from the default values. There is evidence of noise in the posterior for VmNO3 that Fiechter et al. (2013) attribute to high-frequency variability in the SeaWiFS data that is missing in smoother forward model output (see Figure 4). Posterior distribution estimates for ZooGR using SeaWiFS-only data stage inputs exhibit an as yet unexplained bi-modality with modal values much larger than the default values. The combined GLOBEC and SeaWiFS data lead to a posterior distribution estimate for VmNO3 that is highly uncertain, emphasizing the influence of noisy and more abundant SeaWiFS data. The bi-modality in ZooGR disappears and the posterior mean value is closer to the default. Recall that SeaWiFS provides an estimate of chlorophyll only at the surface, averaged over a 10 km area. The GLOBEC station data include profiles of chlorophyll and nitrate at 10 m intervals

in the vertical. Apparently, these data sets are detecting different processes affecting phytoplankton abundance at the GLOBEC inner shelf station.

Identifiability (Box 1) issues begin to arise when the number of random parameters is expanded to six. Figure 6 shows the posterior distribution estimates for (from left) PhyIS, VmNO3, ZooGR, DetRR, KFeC, and FeRR at the inner shelf location, given GLOBEC station data (top row), SeaWiFS surface chlorophyll retrievals (middle row), and the combined GLOBEC and SeaWiFS data (bottom row). These data stage inputs to the NPZDFe BHM now identify six random parameters controlling phytoplankton abundance limitations due to light, nitrogen, iron, and remineralization of detritus and iron. Individual data sets (either GLOBEC or SeaWiFS) lead to posterior distributions exhibiting significant uncertainties for almost all parameters. In the GLOBEC data case, the uncertainty in VmNO3 noted in the two random parameter BHM has been compensated in some sense by a low, but

relatively certain, distribution for PhyIS. Low values of the light limitation parameter are offsetting large but uncertain values in growth rate. The compensation appears to go the other way (i.e., large and uncertain PhyIS and lower, but relatively certain, VmNO3) in the SeaWiFS only data stage case that focuses on surface chlorophyll only. This is an example of parameter correlation making the interpretation of ecosystem dynamics from BHM output more challenging. Note that in the six random parameter BHM, combining data sets reduces uncertainty in the posterior distributions for all parameters, with many parameter posterior mean values near default values. The exception is FeRR, which was not shown to be important on the shelf in the ensemble experiments (Fiechter, 2012; see also Figures 2 and 3).

Careful consideration of the posterior distribution estimates for the parameters of the NPZDFe BHM can be used to quantify parameter identifiability, evaluate differing information content in differently sampled data stage inputs, and
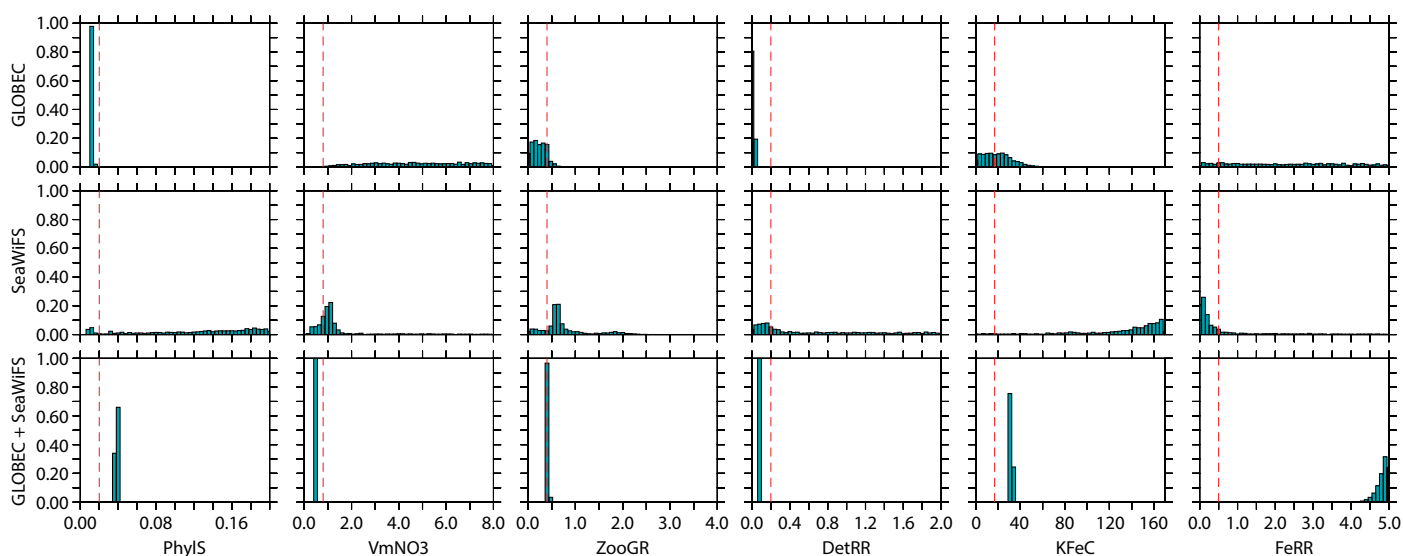

Figure 6. CGOA inner-shelf posterior distributions for PhyIS, VmNO3, ZooGR, DetRR, KFeC, and FeRR (from left to right) using observations from GLOBEC (top), SeaWiFS (middle), and both (bottom). Dashed vertical red lines indicate default parameter values (see Box 3). Frequency of occurrence is normalized as in Figure 4.

qualify ecosystem dynamical interpretations (i.e., in terms of uncertainties). Validation and sensitivity experiments with simulated data stage inputs from skillful forward model integrations are essential to diagnosing these issues as well (i.e., the nearly perfect data experiments). However, computational costs associated with many thousands of iterations through the MCMC algorithm constrain the number of experiments that can be run. Limitations in the abundance and precision of data stage inputs constrain the number of parameters that can be treated as random and identified in the posterior. These constraints preclude BHM experiments wherein random parameters are identified in space (i.e., at each grid point or in subregions on the shelf, at shelf break, or in the basin) and time (i.e., for different phases of the spring bloom). In the next section, we describe developments to circumvent computational costs so that many more degrees of freedom enter the Bayesian analysis.

BAYESIAN STATISTICAL EMULATORS FOR ESTIMATING PARAMETERS AND STATE VARIABLES

The ROMS-NPZDFe ensemble calculations were put to a second use after serving to validate and recast the CGOA ocean ecosystem model parameter estimation in probability distribution terms. Statistical "emulators" or "surrogates" are increasingly used to approximate complex forward models, often for the purposes of inferring model parameters (Box 1; e.g., see also Kennedy and O'Hagan, 2001; Higdon et al., 2008; Rougier, 2008; Conti et al., 2009). The emulator approach is implemented in two stages. First, emulator training data sets are constructed by running

the forward model under multiple sets of parameters as we have done with ROMS-NPZDFe via the ensemble calculations described in the previous section. In the second stage, a summary of the forward model output is devised to serve as a response relative to the input parameters. A statistical model that fits the covariance properties of the forward model responses is constructed so that forward model response can be predicted for parameter settings that have not been used in the training calculations. If observations are available that correspond to the response summary, then the emulator can take a BHM form where the observations inform the data stage distribution and the formulation of the statistical model is used for the process model stage (Box 1). Parameter uncertainty and state variable field estimations were achieved using emulators in the CGOA.

We modified the standard emulator practice described in the references above by fitting the statistical model (i.e., stage 2) to the mean response itself rather than the covariance (see Box 5). This follows the methods developed by Hooten et al. (2011), where the dimensions of the forward model are reduced by combining a few leading spatial patterns that characterize most of the spatial variability evident in the forward model ensemble calculations. The spatial patterns are the summary output that is modeled statistically as functions of the forward model parameters.

The Bayesian emulator applications in the CGOA used surface chlorophyll retrievals from SeaWiFS for data stage inputs. Posterior distributions for parameters of the NPZDFe ocean ecosystem model were estimated in a Bayesian emulator that linked all the GLOBEC and other stations in Figure 1 via a spatial covariance model (Leeds et al., 2013). In

a second application, a Bayesian emulator was used to obtain posterior distributions for complete fields of surface phytoplankton in the CGOA domain, with space-time variable uncertainty estimates. Examples from these Bayesian emulator applications are presented below.

In both cases, given the observations and the statistical model for the summary of forward model outputs, the posterior distribution on parameters can be estimated. Box 5 describes the specifics of the procedure.
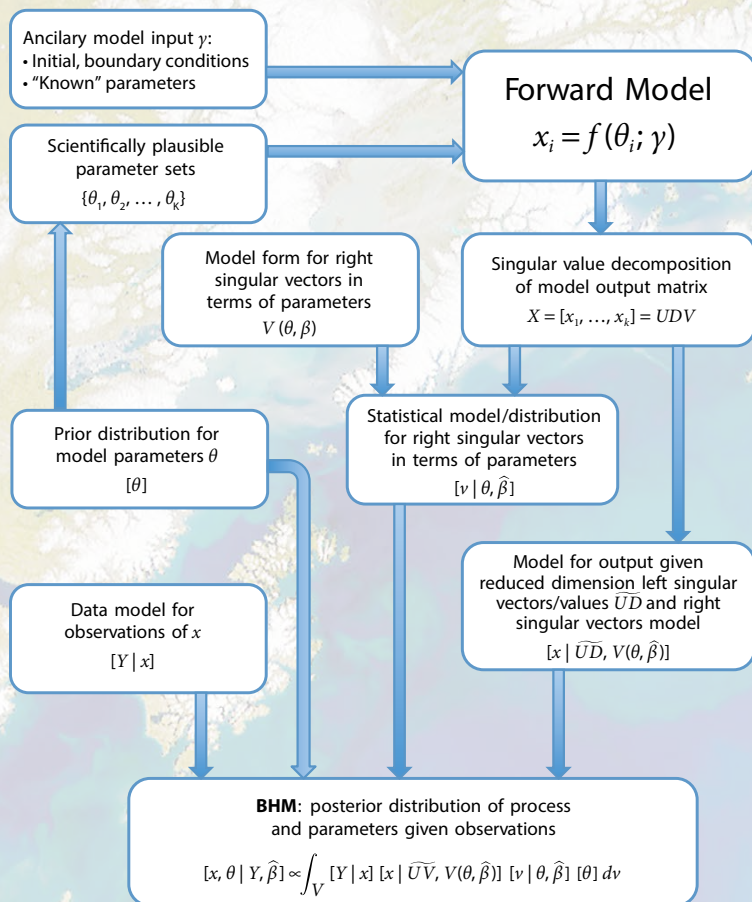
First-Order Emulator-Assisted Parameter Estimation: Modeling 3-D Processes with a Forest of 1-D Emulators

Figures 2 and 3 demonstrate the widely held contention that biogeochemical model parameters vary in space (e.g., Friedrichs et al., 2007). We sought to link, in the horizontal spatial dimension, a collection of 1-D vertical NPZDFe models through a hierarchical statistical model for the spatially distributed model parameters. As we have seen, the dynamics and ecosystem processes of the inner shelf and outer shelf are distinct. The Bayesian emulator built from the ensemble of ROMS-NPZDFe calculations was recast to emulate vertical NPZDFe ecosystem variability at a "forest" of 1-D locations corresponding to the stations along the GLOBEC Seward Line and two additional "fictitious" lines off Kodiak and Shumagin Islands in Figure 1. The model input parameters were assigned spatially dependent prior distributions according to the noted along-shelf and across-shelf dependencies. Posterior distributions for ZooGR and the half-saturation constant for iron, KFeC, were obtained using SeaWiFS chlorophyll retrievals as the data stage input. Both parameters exhibited Bayesian

BOX 5: BAYESIAN EMULATOR

A BHM is developed to estimate state variables and their uncertainties based on training from an ensemble of forward (deterministic) model calculations. The goal is to *emulate* the forward model in an efficient way. In the coastal Gulf of Alaska context, the forward model ensemble is from ROMS-NPZDFe, given perturbations in seven LTL ocean ecosystem model parameters (see text). We suggest a simple model for a few right singular vectors obtained from an SVD of the ROMS-NPZDFe ensemble. The BHM data stage distribution is based on SeaWiFS surface chlorophyll retrievals.

A generic flowchart and algorithm for a Bayesian emulator are as follows:

Ancillary model input γ:
• Initial, boundary conditions
• "Known" parameters

Scientifically plausible parameter sets
$\{\theta_1, \theta_2, \ldots, \theta_k\}$

Forward Model
$x_i = f(\theta_i; \gamma)$

Model form for right singular vectors in terms of parameters
$V(\theta, \beta)$

Singular value decomposition of model output matrix
$X = [x_1, \ldots, x_k] = UDV$

Prior distribution for model parameters θ
$[\theta]$

Statistical model/distribution for right singular vectors in terms of parameters
$[v \mid \theta, \hat{\beta}]$

Model for output given reduced dimension left singular vectors/values $\widetilde{UD}$ and right singular vectors model
$[x \mid \widetilde{UD}, V(\theta, \hat{\beta})]$

Data model for observations of x
$[Y \mid x]$

BHM: posterior distribution of process and parameters given observations

$$[x, \theta \mid Y, \hat{\beta}] \propto \int_V [Y \mid x] [x \mid \widetilde{UV}, V(\theta, \hat{\beta})] [v \mid \theta, \hat{\beta}] [\theta] \, dv$$

1. Select sets of 7 randomized input parameters, $\theta_{p1}, \ldots, \theta_{pK}$ (i.e., from the prior distribution $[\theta_p]$).
2. For θ_{pi} run the forward model θ_{pi}, γ (where γ corresponds to ancillary model input such as boundary and initial conditions and "fixed" parameters), obtaining the desired output vector x_i. These constitute ensemble calculations.
3. Collect output into matrix $X = (x_1, \ldots, x_K)$, i.e., the ensemble response.
4. Perform singular value decomposition of the matrix in step 3 to get $X = UDV'$ and approximate it by a subset of left and right singular vectors, i.e., $X \approx \widetilde{U}\widetilde{D}\widetilde{V}'$.
5. Develop a model for each right singular vector with θ_{pi} as a predictor variable corresponding to response variable v_i in the ith column of $\widetilde{V}$. The resulting model is $v(\theta, \hat{\beta})$ where $\hat{\beta}$ are the estimated model parameters.
6. Perform model calibration, using $\widetilde{U}\widetilde{V}v(\theta^*, \hat{\beta})$ in place of $f(\theta^*, \gamma)$.

learning, with spatial linkages more important for KFeC than for ZooGR (Leeds et al., 2013).

Emulator-Assisted Biogeochemical Data Assimilation

In a final BHM application in the CGOA presented here, the focus shifts from parameter estimation in the LTL ocean ecosystem model to an estimate of the surface phytoplankton field, with associated uncertainties. This is the BHM version of the data assimilation procedure that is very much a topic of research in deterministic forecast systems for coupled physical-biogeochemical models.

The Bayesian emulator was constructed from the forward model ensemble calculations for the years 1998–2001. The CGOA surface phytoplankton field for 2002 will be estimated using data stage inputs for sea surface height (SSH) and sea surface temperature (SST) sampled from the ROMS-NPZDFe simulation for that year and using surface chlorophyll concentration retrievals from real SeaWiFS observations for 2002. The data stage inputs from ROMS-NPZDFe were degraded to eight-day averages to match the coverage issues from SeaWiFS. The statistical process model for BHM biogeochemical data assimilation demonstrates a new statistical technology in that an explicitly nonlinear model for the spatial structure function is implemented (Leeds et al., 2012).

As Figure 7 shows, the BHM data assimilation model blends the ROMS-NPZDFe dynamics with observations from SeaWiFS in a framework that provides plausible surface phytoplankton fields and measures spatially and temporally varying uncertainty. In validation experiments (not shown), the posterior mean fields from the Bayesian emulator

compared well with "truth" fields from ROMS-NPZDFe with data assimilation (Leeds et al., 2012).

SUMMARY

The interplay between deterministic and probabilistic methods leads to clearer understanding of LTL ecosystem dynamics in the CGOA and the extent to which those dynamics are conditioned upon key parameters of the ecosystem model.

The quantification of uncertainty in the posterior distribution of the BHM is an important advance in understanding at the abstracted level of the NPZDFe LTL ecosystem model approximation. In addition to these specific results, much of the work here can be considered a "proof of methodology" as well. The refinement of the NPZDFe BHM and the specific focus on key parameters of the LTL ecosystem model in the BHM

(Fiechter et al., 2013) depend upon intuition gained in ensemble experiments in the coupled physical-biological forward model (Fiechter, 2012). Limitations in the state-space dimension tractable in the BHM are overcome by constructing Bayesian emulators based on leading space-time patterns deduced from, again, ensemble forward model calculations (Hooten et al., 2011). In the larger state-space Bayesian emulator
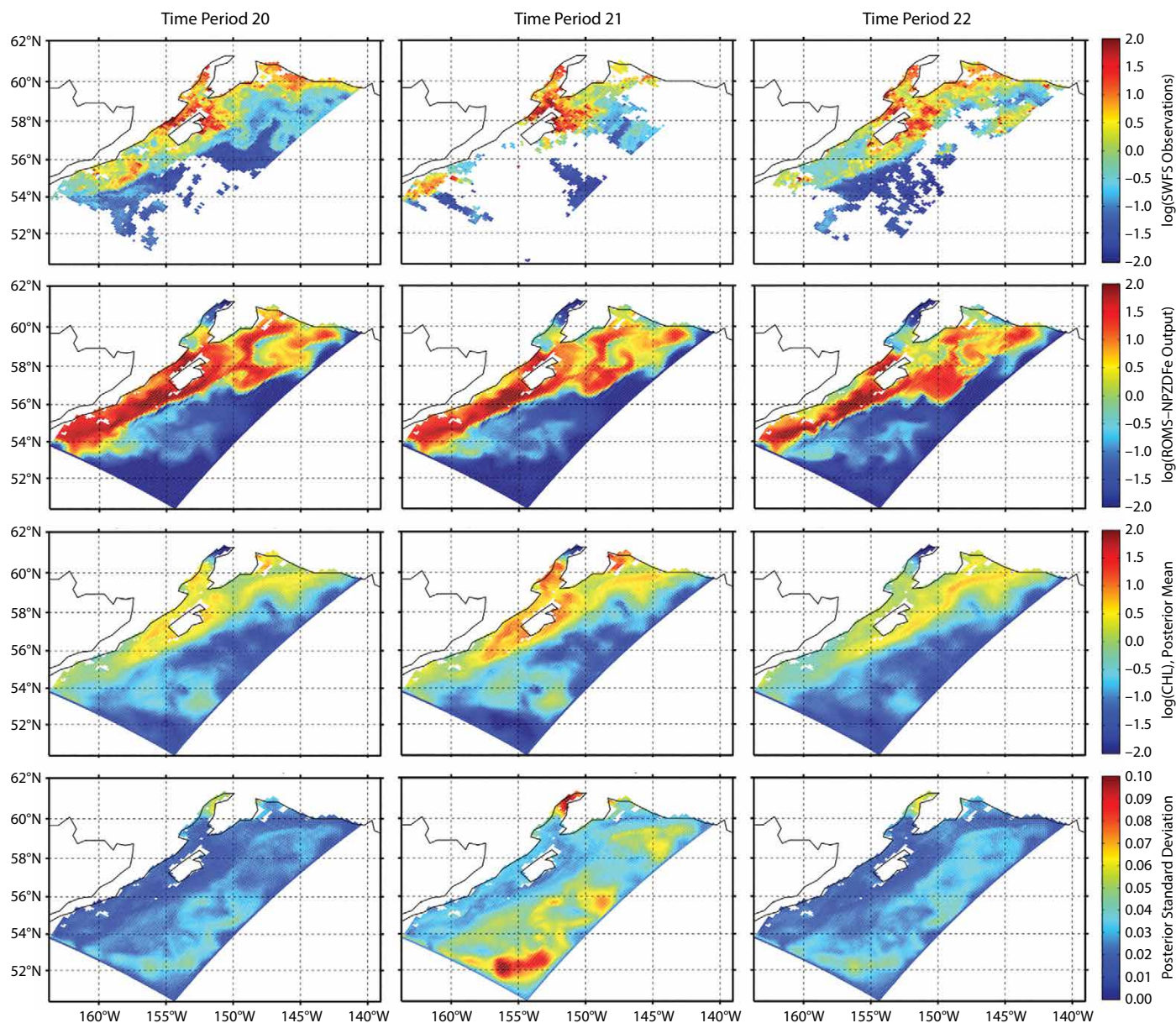
Figure 7. Plots of log-transformed SeaWiFS ocean color observations (top row), ROMS-NPZDFe phytoplankton output (second row), posterior mean (third row), and posterior standard deviation (fourth row), for three eight-day time periods: June 2, 2002, to June 9, 2002 (left column), June 10, 2002, to June 17, 2002 (center column), and June 18, 2002, to June 25, 2002 (right column). *Adapted from Leeds et al. (2012)*

applications, parameters borrow strength in horizontal spatial dimensions such that estimates for parameters are obtained for shelf, shelf break, and basin subregions of the CGOA domain (Leeds et al., 2013). Bayesian emulators are also used to provide estimates, with space and time-variable uncertainties, for surface phytoplankton fields from sparse and imperfect SeaWiFS observations (Leeds et al., 2012).

The underdetermination problem is not going away. Although we have relatively large amounts of satellite-derived estimates of near-surface phytoplankton abundance from the ocean color proxy, these observations are incomplete and fairly uncertain. The NPZDFe models considered here are abstractions of more complicated multicomponent LTL ocean ecosystem models (e.g., NEMURO; see Kishi et al., 2007). The identifiability issues discussed here are only going to be amplified in these more complicated models. This suggests that uncertainty quantification in biogeochemical models will be focused on the relatively few identifiable parameters, or the focus will change to one of state prediction rather than parameter inference. In this case, there is a great need to account for the uncertainties in these predictions and to use these predictive distributions to link to other higher trophic levels of the ocean ecosystem. A major use of these linkages will be to study both the consequences of management decisions and global climate change. The major components of these models are likely to include stochastic parameterizations or emulator-based processes. Both approaches will require significant contributions from statistical scientists in collaboration with physical and biological oceanographers. We believe

that the results presented here provide a template for how such collaborations can be achieved.

REFERENCES

Berliner, L.M., R.F. Milliff, and C.K. Wikle. 2003. Bayesian hierarchical modeling of air-sea interaction. *Journal of Geophysical Research* 108, C43104, http://dx.doi.org/10.1029/2002JC001413.

Berliner, L.M., C.K. Wikle, and N. Cressie. 2000. Long-lead prediction of Pacific SSTs via Bayesian Dynamic Modeling. *Journal of Climate* 13:3,953–3,968, http://dx.doi.org/10.1175/1520-0442(2001)013<3953:LLPOPS>2.0.CO;2.

Brown, J., and J. Fiechter. 2012. Quantifying eddy-chlorophyll covariability in the coastal Gulf of Alaska. *Dynamics of Atmospheres and Oceans* 55–56:1–21, http://dx.doi.org/10.1016/j.dynatmoce.2012.04.001.

Conti, S., J. Gosling, J. Oakley, and A. O'Hagan. 2009. Gaussian process emulation for dynamic computer codes. *Biometrika* 96:663–676, http://dx.doi.org/10.1093/biomet/asp028.

Cressie, N., and C.K. Wikle. 2011. *Statistics for Spatio-Temporal Data*. Wiley Series in Probability and Statistics, John Wiley and Sons Inc., 588 pp.

Cripps, E., D. Nott, W.T.M. Dunsmuir, and C.K. Wikle. 2005. Space-time modelling of Sydney Harbour Wind. *Australian and New Zealand Journal of Statistics* 47:3–17, http://dx.doi.org/10.1111/j.1467-842X.2005.00368.x.

Dowd, M. 2007. Bayesian statistical data assimilation for ecosystem models using Markov Chain Monte Carlo. *Journal of Marine Systems* 68:439–456, http://dx.doi.org/10.1016/j.jmarsys.2007.01.007.

Dowd, M. 2011. Estimating parameters for a stochastic dynamic marine ecological system. *Environmetrics* 22:501–515, http://dx.doi.org/10.1002/env.1083.

Fiechter, J. 2012. Assessing marine ecosystem model properties from ensemble calculations. *Ecological Modelling* 242:164–179, http://dx.doi.org/10.1016/j.ecolmodel.2012.05.016.

Fiechter, J., G. Broquet, A.M. Moore, and H.G. Arango. 2011. A data-assimilative, coupled physical-biological model for the coastal Gulf of Alaska. *Dynamics of Atmospheres and Oceans* 52:95–118, http://dx.doi.org/10.1016/j.dynatmoce.2011.01.002.

Fiechter, J., R. Herbei, W. Leeds, J. Brown, R. Milliff, C. Wikle, A. Moore, and T. Powell. 2013. A Bayesian parameter estimation method applied to a marine ecosystem model for the coastal Gulf of Alaska. *Ecological Modelling* 258:122–133, http://dx.doi.org/10.1016/j.ecolmodel.2013.03.003.

Fiechter, J., and A.M. Moore. 2009. Interannual spring bloom variability and Ekman pumping in the coastal Gulf of Alaska. *Journal of Geophysical Research* 114, C06004, http://dx.doi.org/10.1029/2008JC005140.

Fiechter, J., and A.M. Moore. 2012. Iron limitation impact on eddy-induced ecosystem variability in the coastal Gulf of Alaska. *Journal of Marine Systems* 92:1–15, http://dx.doi.org/10.1016/j.jmarsys.2011.09.012.

Fiechter, J., A.M. Moore, C.A. Edwards, K.W. Bruland, E. Di Lorenzo, C.V.W. Lewis, T.M. Powell, E.N. Curchitser, and K. Hedstrom. 2009. Modeling iron limitation of primary

production in the coastal Gulf of Alaska. *Deep Sea Research Part II* 56:2,503–2,519, http://dx.doi.org/10.1016/j.dsr2.2009.02.010.

Fox, N.I., and C.K. Wikle. 2005. A Bayesian quantitative precipitation nowcast scheme. *Weather and Forecasting* 2:264–275, http://dx.doi.org/10.1175/WAF845.1.

Friedrichs, M.A.M, M.-E. Carr, R.T. Barber, M. Scardi, D. Antoine, R.A. Armstrong, I. Asanuma, M.J. Behrenfeld, E.T. Buitenhuis, F. Chai, and others. 2009. Assessing the uncertainties in model estimates of primary productivity in the tropical Pacific Ocean. *Journal of Marine Systems* 76:113–133, http://dx.doi.org/10.1016/j.jmarsys.2008.05.010.

Friedrichs, M.A.M., J.A. Dusenberry, L.A. Anderson, R.A. Armstrong, F. Chai, J.R. Christian, S.C. Doney, J. Dunne, M. Fujii, R. Hood, and others. 2007. Assessment of skill and portability in regional marine biogeochemical models: Role of multiple planktonic groups. *Journal of Geophysical Research* 112, C08001, http://dx.doi.org/10.1029/2006JC003852.

Haidvogel, D.B., H. Arango, W.P. Budgell, B.D. Cornuelle, E. Curchitser, E. Di Lorenzo, K. Fennel, W.R. Geyer, A.J. Hermann, L. Lanerolle, and others. 2008. Ocean forecasting in terrain-following coordinates: Formulation and skill assessment of the Regional Ocean Modeling System. *Journal of Computational Physics* 227:3,595–3,624, http://dx.doi.org/10.1016/j.jcp.2007.06.016.

Harmon, R., and P. Challenor. 1997. A Markov chain Monte Carlo method for estimation and assimilation into models. *Ecological Modelling* 101:41–59, http://dx.doi.org/10.1016/S0304-3800(97)01947-9.

Higdon, D., J. Gattiker, B. Williams, and M. Rightly. 2008. Computer model calibration using high-dimensional output. *Journal of the American Statistical Association* 103:570–583, http://dx.doi.org/10.1198/016214507000000888.

Hooten, M.B., and C.K. Wikle. 2007. Shifts in the spatio-temporal growth dynamics of shortleaf pine. *Environmental and Ecological Statistics* 14:207–227, http://dx.doi.org/10.1007/s10651-007-0016-1.

Hooten, M.B., and C.K. Wikle. 2008. A hierarchical Bayesian non-linear spatio-temporal model for the spread of invasive species with application to the Eurasian Collared-Dove. *Environmental and Ecological Statistics* 15:59–70, http://dx.doi.org/10.1007/s10651-007-0040-1.

Hooten, M.B., W.B. Leeds, J. Fiechter, and C.K. Wikle. 2011. Assessing first-order emulator inference for physical parameters in nonlinear mechanistic models. *Journal of Agricultural, Biological and Environmental Statistics* 16:475–494, http://dx.doi.org/10.1007/s13253-011-0073-7.

Hooten, M.B., C.K. Wikle, R.M. Dorazio, and J.A. Royle. 2007. Hierarchical spatio-temporal matrix models for characterizing invasions. *Biometrics* 63:558–567, http://dx.doi.org/10.1111/j.1541-0420.2006.00725.x.

Kennedy, M., and A. O'Hagan. 2001. Bayesian calibration of computer models. *Journal of the Royal Statistical Society Series B* 63:425–464, http://dx.doi.org/10.1111/1467-9868.00294.

Kishi, M., M. Kashiwai, D.M. Ware, B.A. Megrey, D.L. Eslinger, F.E. Werner, M.N. Aita, T. Azumaya, M. Fujii, S. Hashimoto, and others. 2007. NEMURO: A lower trophic level model for the North Pacific marine ecosystem. *Ecological Modelling* 202:12–25, http://dx.doi.org/10.1016/j.ecolmodel.2006.08.021.

Leeds, W.B, C.K. Wikle, and J. Fiechter. 2012. Emulator-assisted reduced-rank ecological data assimilation for nonlinear multivariate dynamical spatio-temporal processes. *Statistical Methodology* 17:126–138, http://dx.doi.org/10.1016/j.stamet.2012.11.004.

Leeds, W.B., C.K. Wikle, J. Fiechter, J.L. Brown, and R.F. Milliff. 2013. Modeling 3-D spatio-temporal biogeochemical processes with a forest of 1-D statistical emulators. *Environmetrics* 24:1–12, http://dx.doi.org/10.1002/env.2187.

Malve, O., M. Laine, H. Haario, T. Kirkkala, and J. Sarvala. 2007. Bayesian modeling of algal mass occurrences—using adaptive MCMC methods with a lake water quality model. *Environmental Modelling & Software* 22:966–977, http://dx.doi.org/10.1016/j.envsoft.2006.06.016.

Margvelashvili, N., and E.P. Campbell. 2012. Sequential data assimilation in fine-resolution models using error-subspace emulators: Theory and preliminary evaluation. *Journal of Marine Systems* 90:13–22, http://dx.doi.org/10.1016/j.jmarsys.2011.08.004.

Milliff, R.F., A. Bonazzi, C.K. Wikle, N. Pinardi, and L.M. Berliner. 2011. Ocean ensemble forecasting. Part 1: Ensemble Mediterranean winds from a Bayesian hierarchical model. *Quarterly Journal of the Royal Meteorological Society* 137:858–878, http://dx.doi.org/10.1002/qj.767.

Moore, A.M., H.G. Arango, G. Broquet, B.S. Powell, J. Zavala-Garay, and A.T. Weaver. 2011. The Regional Ocean Modeling System (ROMS) 4-dimensional variational data assimilation systems: Part I – System overview and formulation. *Progress in Oceanography* 91:34–49, http://dx.doi.org/10.1016/j.pocean.2011.05.004.

Parslow, J., N. Cressie, E. Campbell, E. Jones, and L. Murray. 2013. Bayesian learning and predictability in a stochastic nonlinear dynamical model. *Ecological Applications* 26:679–698, http://dx.doi.org/10.1890/12-0312.1.

Powell, T.M., C.V.W. Lewis, E.N. Curchitser, D.B. Haidvogel, A.J. Hermann, and E.L. Dobbins. 2006. Results from a three-dimensional nested, biological-physical model of the California Current System and comparisons with statistics from satellite imagery. *Journal of Geophysical Research* 111, C07018, http://dx.doi.org/10.1029/2004JC002506.

Rougier, J. 2008. Efficient emulators for multivariate deterministic functions. *Journal of Computational and Graphical Statistics* 17:827–843, http://dx.doi.org/10.1198/106186008X384032.

Royle, J.A., L.M. Berliner, C.K. Wikle, and R.F. Milliff. 1998. A hierarchical spatial model for constructing wind fields from scatterometer data in the Labrador Sea. Pp. 367–381 in *Case Studies in Bayesian Statistics IV*. C. Gatsonis, R.E. Kass, B. Carlin, A. Cariquiry, A. Gelman, I. Verdinelli, and M. West, eds, Springer-Verlag.

Song, Y., C.K. Wikle, C.J. Anderson, and S.A. Lack. 2007. Bayesian estimation of stochastic parameterizations in a numerical weather forecasting model. *Monthly Weather Review* 135:4,045–4,059, http://dx.doi.org/10.1175/2007MWR1928.1.

Strom, S.L., M. Brady Olson, E.L. Macri, and C.W. Mordy. 2006. Cross-shelf gradient in phytoplankton community structure, nutrient utilization and growth rate in the coastal Gulf of Alaska. *Marine Ecology Progress Series* 328:75–92, http://dx.doi.org/10.3354/meps328075.

Ward, B.A., M.A.M. Friedrichs, T.R. Anderson, and A. Oschlies. 2010. Parameter optimization techniques and the problem of underdetermination in marine biogeochemical models. *Journal of Marine Systems* 81:34–43, http://dx.doi.org/10.1016/j.jmarsys.2009.12.005.

Wikle, C.K. 2003a. Hierarchical Bayesian models for predicting the spread of ecological processes. *Ecology* 84:1,382–1,394, http://dx.doi.org/10.1890/0012-9658(2003)084[1382:HBMFPT]2.0.CO;2.

Wikle, C.K. 2003b. Hierarchical models in environmental science. *International Statistical Review* 71:181–199, http://dx.doi.org/10.1111/j.1751-5823.2003.tb00192.x.

Wikle, C.K., R.F. Milliff, D. Nychka, and L.M. Berliner. 2001. Spatiotemporal hierarchical Bayesian modeling: Tropical ocean surface winds. *Journal of the American Statistical Association* 96:383–397, http://dx.doi.org/10.1198/016214501753168109.

Wikle, C.K., and M.B. Hooten. 2006. Hierarchical Bayesian spatio-temporal models for population spread. Pp. 145–169 in *Applications of Computational Statistics in the Environmental Sciences: Hierarchical Bayes and MCMC Methods*. J.S. Clark and A. Gelfand, eds, Oxford University Press.

Wikle, C.K., and M.B. Hooten. 2010. A general science-based framework for spatio-temporal dynamical models. *TEST* 19:417–451, http://dx.doi.org/10.1007/s11749-010-0209-z.

Wikle, C.K., R.F. Milliff, R. Herbei, and W.B. Leeds. 2013. Modern statistical methods in oceanography: A hierarchical perspective. *Statistical Science* 28:466–486, http://dx.doi.org/10.1214/13-STS436.