Research Article Environmetrics

Received: 25 February 2012, Revised: 12 September 2012, Accepted: 25 October 2012, Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/env.2187

Modeling 3-D spatio-temporal biogeochemical processes with a forest of 1-D statistical emulators

W. B. Leeds[a,b]*, C. K. Wikle[b], J. Fiechter[c], J. Brown[d] and R. F. Milliff[e]

This paper focuses on the spatio-temporal dynamical processes in lower trophic level marine ecosystems, where various sources of uncertainty make statistical modeling difficult. Such dynamical processes exhibit nonlinearity in time and potential nonstationarity in space. Planktonic organisms are microscopic, making it difficult to measure their abundance and resulting in limited data. Further, deterministic, component-based ecosystem models contain a large number of parameters, some of which can be difficult to estimate. We consider a Bayesian hierarchical framework for parameter estimation that uses an approximation to the dynamical models for computational feasibility. Specifically, we develop a computationally inexpensive first-order statistical emulator for a one-dimensional NPZD model with iron limitation. Then, we introduce a novel approach to the modeling of three-dimensional lower trophic level marine ecosystem processes, linking the one-dimensional emulators via a two-dimensional spatial field on the parameters. This methodology is used to estimate important biological parameters on the coastal Gulf of Alaska, leading to a reduction in Bayesian credible interval width compared with a nonspatial model. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: Bayesian; emulator; hierarchical model; nonlinear model; marine ecosystem; lower trophic level

1. INTRODUCTION

Many scientific processes can be represented by spatial fields that evolve in time. For such processes, spatio-temporal statistical models are useful for assessing the various sources of uncertainty. A common approach to developing these models is through descriptive statistical models that characterize the process through first-order and second-order moment (i.e., mean and covariance) specifications. However, it is also sometimes useful to consider a dynamical model where the current value of the underlying scientific process is modeled as an evolution from the process in the previous time step(s), and possibly related to spatial locations of nearby proximity (e.g., see Cressie and Wikle, 2011, for an overview).

A common pitfall with dynamical spatio-temporal models (DSTMs) is the so-called curse of dimensionality. Even relatively simple, linear DSTMs are often overparameterized in higher-dimensional settings. Bayesian hierarchical models (BHMs) provide a partial solution to this problem by allowing dependence among parameters through a series of conditional models, but even this is often not enough in the case of more complicated nonlinear processes, and the specification of nonlinear DSTMs. However, BHMs also allow for the inclusion of scientific information into the model. We not only can include scientific information about the parameters but also can use scientific knowledge about the process dynamics. In other words, we can develop a science-based parameterization for the statistical model (e.g., Leeds and Wikle, 2012). Such a model can allow for dependencies in the parameters and in the underlying scientific process. Incorporating such information can be carried out in multiple ways.

1.1. Incorporation of scientific information

One way of incorporating scientific information about the process into a BHM would be through a "physical–statistical" or "mechanistically motivated statistical" model (Royle *et al.*, 1999; Wikle *et al.*, 2001; Berliner, 2003; Wikle, 2003; Wikle and Hooten, 2010; Milliff *et al.*,

* *Correspondence to: W. B. Leeds, Department of Statistics and Department of the Geophysical Sciences, University of Chicago, Chicago, IL, U.S.A. E-mail: leedsw@uchicago.edu*

a *Department of Statistics and Department of the Geophysical Sciences, University of Chicago, IL, U.S.A.*

b *Department of Statistics, University of Missouri, Columbia, MO, U.S.A.*

c *Institute of Marine Science, University of California-Santa Cruz, Santa Cruz, CA, U.S.A.*

d *Principal Scientific Group, Fishers, IN, U.S.A.*

e *CIRES, University of Colorado, Boulder, CO, U.S.A.*

2012), where scientific information about a physical (or biological, ecological, etc.) process is used to motivate a sensible parameterization of the DSTM.

It is also common to use the mechanistic model (or, more specifically, numerical solutions to a mechanistic model) directly in the BHM, rather than using it simply to motivate the parameterization of the BHM. This is related to the Bayesian melding approach (Poole and Raftery, 2000; Fuentes and Raftery, 2005; Finley *et al.*, 2011), which includes the use of implicit and explicit priors on the input and output of a mechanistic model. These four priors are then pooled together in order to perform Bayesian inference. Mechanistic models are also used in a framework that does not involve the pooling of implicit and explicit priors. For example, Royle *et al.* (1999) used spatial output from mechanistic models as data, along with real-world observations. Wikle *et al.* (2001) used spatial output as data along with a reduced-rank mechanistically motivated statistical model. This framework has been shown to accommodate both upscaling and downscaling (Fuentes and Raftery, 2005; Wikle and Berliner, 2005; Berrocal *et al.*, 2010).

1.2. Statistical emulators

In many instances, the mechanistic model is often computationally expensive, and so running it iteratively in a Markov chain Monte Carlo algorithm poses a problem. We follow the foundational work of Sacks *et al.* (1989), Currin *et al.* (1991), and Kennedy and O'Hagan (2001) and use a statistical approximation to the computer code as a surrogate model. This surrogate model acts as a fast approximation to the mechanistic model and allows one to perform uncertainty analysis, sensitivity analysis (O'Hagan, 2006), and model calibration and prediction (Higdon *et al.*, 2004, 2008) with a high degree of accuracy while requiring relatively few runs of the mechanistic model. Recent research has extended the methodology to incorporate dimension reduction (Higdon *et al.*, 2008), dynamical models (Drignei, 2008; Conti *et al.*, 2009; Liu and West, 2009), and multivariate output (Rougier, 2008; Conti and OHagan, 2010).

Although most statistical emulation has been carried out through the use of second-order (covariance) model specification (Sacks *et al.*, 1989; Kennedy and O'Hagan, 2001; O'Hagan, 2006), it may be desirable to model the input–output relationship for a mechanistic model by using first-order characteristics (e.g., van der Merwe *et al.*, 2007; Frolov *et al.*, 2009). For example, Hooten *et al.* (2011) used random forests to model the inputs to the mechanistic model to the right singular vectors of a singular value decomposition of an ensemble of realizations from a mechanistic model. Their goal was to facilitate parameter estimation for a lower trophic level marine ecosystem model. These emulators can then be used in the place of the deterministic model inside the BHM.

We are interested in modeling lower trophic level marine ecosystem processes. Because the associated mechanistic models were originally developed to investigate phytoplankton response to oceanic mixed layer dynamics, their formulation is inherently one-dimensional (1-D) in the vertical direction. However, important spatial variability also occurs in the horizontal directions. Computer models for lower trophic level marine ecosystems generally take this horizontal spatial variability into account by coupling a vertical marine ecosystem model to a physical ocean circulation model. As such, the spatial variability cannot all be attributed to differences in parameters at different spatial locations—a notion that is scientifically plausible (Friedrichs *et al.*, 2007).

In this paper, we construct a first-order emulator to act as a surrogate for a 1-D (vertical) mechanistic model representing lower trophic level marine ecosystem dynamics. Because the 1-D mechanistic model includes vertical dynamics, we think of a single 1-D emulator as a "tree." Then, we link together these 1-D emulators through 2-D spatial random fields on the parameters, creating a "forest" of 1-D models that allows for the modeling of 3-D spatio-temporal processes. Kennedy *et al.* (2008) proposed a similar concept using Gaussian process emulation of dynamic vegetation models but did not consider a spatial Gaussian process on the input parameters themselves, as is carried out in this paper.

An outline of the paper is given as follows. Section 2 discusses more fully our motivating problem, including the area of study, the data that were used, and the mechanistic model that was emulated. Section 3 provides a brief background on the two-stage approach to using a first-order emulator in a BHM. Section 4 describes a proof-of-concept example, based on simulated data, and an application related to lower trophic level marine ecosystem processes in the coastal Gulf of Alaska (CGOA). Finally, Section 5 provides our discussion of the results and related future work.

2. MOTIVATING PROBLEM

The methodology presented here is motivated by lower trophic level marine ecosystem dynamics. Because lower trophic level marine ecosystem processes are tied to so-called primary and secondary productions, it is critical to be able to monitor the abundance of phytoplankton and zooplankton. This involves accounting for the interactions between biological and physical processes that are involved. These relationships impact the food chain at all trophic levels, so it is important to have an appropriate understanding of the scientific process. In particular, having a proper understanding of the dynamics of phytoplankton throughout the year is important to biological oceanographers. "Biogeochemical models" that take into account important bio-physical interactions range in complexity from the simpler NPZ (nutrient, phytoplankton, zooplankton) model (Franks, 2002) to the more complicated North Pacific Ecosystem Model for Understanding Regional Oceanography model (Kishi *et al.*, 2007), which includes extensions that take into account 1-D (Fujii *et al.*, 2007), 2-D (Wainwright *et al.*, 2007), and 3-D (Aita *et al.*, 2007) spatial structure. These models account for important interactions and are a useful way to predict a process for which fewer data are available.

Typically, model calibration and prediction have been considered from a deterministic perspective. Parameters are "estimated" in the sense that they are constrained so that model output closely resembles the data. Then, the state is predicted by using these constrained values and running the model for future time points. However, a Bayesian framework is a useful way to perform state and parameter estimation from a statistical perspective. This framework has been used for state estimation (Harmon and Challenor, 1997; Evensen, 2003; Dowd, 2006, 2007; Jones *et al.*, 2010) as well as parameter estimation (Harmon and Challenor, 1997; Malve *et al.*, 2007; Jones *et al.*, 2010; Dowd, 2011).

The use of emulators is becoming more common in this area, with Mattern *et al.* (2012) and Margvelashvili and Campbell (2012) using emulators for such models in a data assimilation framework. This is in contrast to Hooten *et al.* (2011), who use the emulator in order to assist in parameter inference.

We consider an application of our methodology to the U.S. Global Ocean Ecosystem Dynamics study area along the CGOA. We develop a BHM that includes remotely sensed ocean color observations (that serve as a proxy variable for phytoplankton biomass), as well as a first-order emulator for the 1-D (in the vertical) NZPDFe model proposed by Fiechter *et al.* (2009). The 1-D NPZDFe model is an NZPD (nutrient, phytoplankton, zooplankton, detritus) lower trophic level ecosystem model with iron (Fe) limitation and a detritus sinking term (see Appendix A for further details).

Our data are based on ocean color observations from the sea-viewing wide field-of-view sensor (SeaWiFS). These remotely sensed observations are effectively chlorophyll measurements, which have been transformed so that they correspond to the surface "P" output from the 1-D NPZDFe model. Although we emulate the 1-D NPZDFe mechanistic model, we only considered the surface levels for phytoplankton, because this corresponded to our only available data. However, the other components (for which we have no observations) are a critical part of the dynamics, meaning that an emulator for the 1-D NPZDFe model should be superior to a 0-D (i.e., no spatial dimension) NPZ (or, even simpler, a predator–prey model) in its place. Because of extensive missing observations on daily and 8-day composites, because of extensive cloud coverage in the CGOA, we consider monthly averages of the SeaWiFS data (Brown and Fiechter, 2012). Then, we construct an emulator by using monthly averages for 5 years for the 1-D NPZDFe model at the study area. For the purpose of this study, nine locations are selected to represent alongshelf and cross-shelf variability (Figure 1). At each alongshelf location (off the Kenai Peninsula, Kodiak Island, and Shumagin Islands), three cross-shelf locations are considered: the inner shelf (closest to shore), the outer shelf (near the shelf break), and the eddy corridor (where oceanic mesoscale processes dominate). Satellite observations were then sampled at those nine locations.

An obvious issue relating to the use of mechanistic models in this framework is the inability of the mechanistic model to generate as much variability as the observations. This is typically the case even when all parameters are allowed to vary and the mechanistic model parameters are not constrained (or, at the very least, have wider constraints than those we have used in our example). Because of a limited number of observations and the noise in those observations (for a discussion of measurement error of SeaWiFS observations, see Hooker and McClain, 2000), it would be difficult to adequately estimate all the model parameters. In similar situations, common approaches include using a simpler mechanistic model, limiting the number of parameters allowed to vary or further restricting the range on certain parameters. We feel that this is a balancing act and compromises must be made. This issue of not being able to adequately estimate all parameters is well known in the biogeochemical modeling community and is referred to as the underdetermination problem (Ward *et al.*, 2010). We note also that in most nonlinear multiparameter models (such as those of interest here), it is possible that multiple parameter sets can give very similar output states. Thus, these two issues (insufficient data and multiple parameter states) require that in traditional optimization approaches and modern Bayesian approaches that one informatively constrain parameters (e.g., Ward *et al.*, 2010).

When deciding to reduce the number of parameters that are allowed to vary, it is important to choose parameters that the model is sensitive to and that have important biological meaning. It would be the hope that only varying a few parameters can adequately capture most of the variation in the output (when all parameters vary randomly). Fiechter (2012) provides a more detailed discussion of this issue, but we note that in our case, the choice of which parameters to vary is only of secondary importance, as compared with demonstration of the methodology that considers a Gaussian process on the parameters to allow us to model 3-D spatio-temporal processes. For this paper,

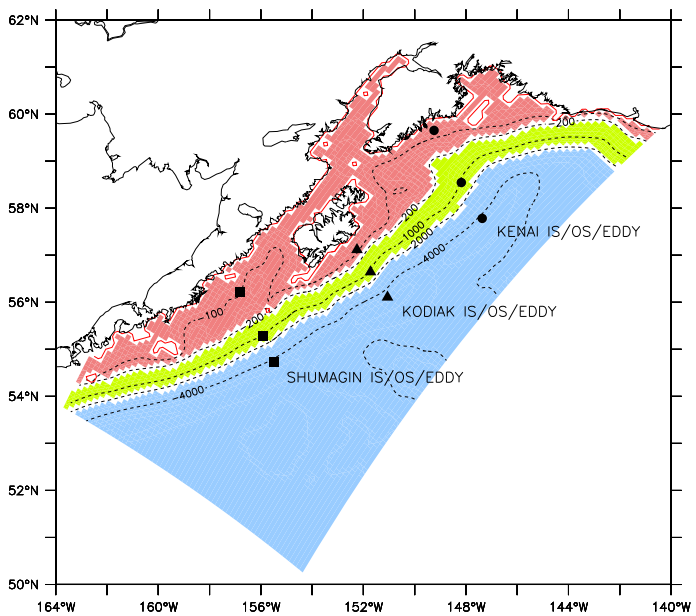

Figure 1. A plot of the study area in the coastal Gulf of Alaska. The three locations on each line are the inner shelf (location nearest the shore), the outer shelf (the middle location on the line) and the eddy corridor (the location furthest off shore)

we allow two critical 1-D NPZDFe model parameters (inputs) to vary: the half-saturation constant for iron (KFeC) and zooplankton maximum grazing rate (ZooGR). On the basis of a priori expert opinion, KFeC varies between 8.45 and 33.8 day^{-1} and ZooGR varies between 0.20 and 0.80 $\mu\text{molFe(molC)}^{-1}$. Other inputs (e.g., biological parameters, initial conditions, etc.) are fixed.

Although the methodology presented herein is motivated by the modeling of lower trophic level marine ecosystem dynamics, it extends to other problems as well. In particular, the methodology is generically applicable to modeling multivariate spatio-temporal processes exhibiting nonlinear behaviors.

3. METHODS

We use a two-stage approach for fitting the model, similar to the one outlined in Hooten *et al.* (2011). First, we develop a first-order emulator, modeling the input–output relationship via random forests. Then, we embed the emulator into the BHM. A detailed explanation of the first-order emulator approach is provided in Appendix B. We split the BHM into a series of conditional models: the data, process, and parameter models, as outlined in Berliner (1996) and Cressie and Wikle (2011). Each model is described in the following subsections. Note that $[X]$ denotes the marginal probability distribution of X, $[X, Y]$ denotes the joint probability distribution of X and Y, and $[X|Y]$ denotes the conditional probability distribution of X given Y.

3.1. Data model

Let $z_{k,t}(\boldsymbol{s}_i)$ represent a realization of the kth variable ($k = 1, \ldots, K$) at time t ($t = 1, \ldots, T$) and location $\boldsymbol{s}_i = (x_i, y_i)$ ($i = 1, \ldots, n$). For our application, we require that the data model be bounded above and below. Thus, we use the following truncated normal distribution for our data model:

$$z_{k,t}(\boldsymbol{s}_i)|y_{k,t}(\boldsymbol{s}_i), \sigma_k^2(\boldsymbol{s}_i) \sim TN\left(y_{k,t}(\boldsymbol{s}_i), \sigma_k^2(\boldsymbol{s}_i)\right)_{\ell_k}^{u_k}$$

where $y_{k,t}(\boldsymbol{s}_i)$ is the latent kth process at time t and location $\boldsymbol{s}_i$. We also assume a process-specific and location-specific measurement error variance. Then, ℓ_k and u_k are the lower and upper bounds for the kth variable. As in all statistical analyses, there is some subjectivity on the choice of likelihood (or data) distributions. We decided upon a truncated normal data model for two reasons. First, the data for our application (phytoplankton concentrations) are required to be positive. Second, although other distributions (e.g., the gamma distribution) could account for the first issue, it is also necessary in our first-order emulator formulation that the mean be allowed to be less than or equal to zero (because our random forest emulator predictions need not be positive). The truncated normal distribution seemed a natural fit in this case. As a result, the joint probability distribution for our data model is

$$\left[z|\boldsymbol{y}, \boldsymbol{\sigma}^2\right] = \prod_{i=1}^{n}\prod_{k=1}^{K}\prod_{t=1}^{T}\left[z_{k,t}(\boldsymbol{s}_i)|y_{k,t}(\boldsymbol{s}_i), \sigma_k^2(\boldsymbol{s}_i)\right]$$

where $\boldsymbol{z} = \left(z_{1,1}(\boldsymbol{s}_1), \ldots, z_{K,T}(\boldsymbol{s}_n)\right)'$, $\boldsymbol{y} = \left(y_{1,1}(\boldsymbol{s}_1), \ldots, y_{K,T}(\boldsymbol{s}_n)\right)'$, and $\boldsymbol{\sigma}^2 = \left(\sigma_1^2(\boldsymbol{s}_1), \ldots, \sigma_K^2(\boldsymbol{s}_n)\right)'$. We also note that this model could easily accommodate multiple sources of data, as well as missing observations, but we retain this specific format for notational convenience.

3.2. Process model

In our model formulation, output from the mechanistic model is included in the process stage of the BHM (as opposed to its use as another source of "observations"). As mentioned in Section 2, we use the 1-D NPZDFe model to motivate the parameterization for the process model:

$$\boldsymbol{y}_t = \mathcal{M}(\boldsymbol{y}_{t-1}; \boldsymbol{\theta_m}; \boldsymbol{\epsilon}_t)$$

where $\boldsymbol{y}_t = (y_{1,t}(\boldsymbol{s}_1), \ldots, y_{K,t}(\boldsymbol{s}_n))'$ and $\mathcal{M}$ is a function that describes the evolution of the process from $t-1$ to t, with parameters $\boldsymbol{\theta_m}$. This mapping could also accommodate a random error process, $\boldsymbol{\epsilon}_t$ (possibly spatially colored).

Often, the process evolution represented by $\mathcal{M}$ is nonlinear, and a general statistical model for the process would likely be overparameterized. As previously mentioned, we use a first-order emulator of the 1-D NPZDFe model, allowing us to include important process dynamics into our statistical model but with computational improvements over using the mechanistic model itself. The first-order emulator is fit offline (i.e., the parameters were estimated and then the emulator was used to represent the process in the BHM) in a two-stage approach. Further details are provided in Appendix B, but we include a brief description of the emulator here.

For $\boldsymbol{y}(\boldsymbol{s}_i) = \left(y_{1,1}(\boldsymbol{s}_i), \ldots, y_{1,T}(\boldsymbol{s}_i), \ldots, y_{K,1}(\boldsymbol{s}_i), \ldots, y_{K,T}(\boldsymbol{s}_i)\right)'$, we have the following process model:

$$\begin{aligned}\boldsymbol{y}(\boldsymbol{s}_i) &= \widetilde{\boldsymbol{\Phi}}_i\boldsymbol{\alpha}_i\\ \boldsymbol{\alpha}_i|\boldsymbol{\theta}_i, \hat{\boldsymbol{\beta}}_i &\sim \left[\boldsymbol{\alpha}_i|\boldsymbol{\theta}_i, \hat{\boldsymbol{\beta}}_i\right]\end{aligned}$$

where $\widetilde{\boldsymbol{\Phi}}_i\boldsymbol{\alpha}_i$ is a truncated spectral representation of the underlying process. The matrix $\widetilde{\boldsymbol{\Phi}}_i$ is a $TK \times q_i$ matrix of basis functions that are the retained left singular vectors (and singular values) from a truncated singular value decomposition (svd) of the ensemble of output from the

1-D NPZDFe model. Here, the size of the truncation (i.e., the choice of q_i) is problem specific, but we typically choose q_i so that we retain almost all (e.g, greater than 95%) of the variation in the mechanistic model output. We caution that the right singular vectors explaining the most variation may not necessarily be the best predictors in certain instances, and so some methods that automatically select which singular vectors to keep (e.g., stochastic search variable selection) may be appropriate in specific applications. In the present application, the 95% truncation works quite well.

The vector $\boldsymbol{\alpha}_i$ is a q_i-dimensional vector with elements corresponding to the retained right singular vectors from the svd, and $\boldsymbol{\theta}_i$ is a p-dimensional vector with elements representing the biological parameters in the mechanistic model for location $\boldsymbol{s}_i$. We note that although the svd is a linear decomposition of the model output, the emulator was not entirely a "linear approximation," as $\boldsymbol{\alpha}_i$ is predicted on the basis of $\boldsymbol{\theta}_i$ (the mechanistic model parameters for location $\boldsymbol{s}_i$) using a nonlinear statistical model to account for uncertainty. Lastly, $\boldsymbol{\beta}_i$ is a vector whose elements are statistical parameters used in this nonlinear statistical model.

Because we follow a two-stage approach, we use the matrices $\widetilde{\boldsymbol{\Phi}}_i$ derived from the left singular vectors and singular values of an svd from the output from the 1-D NPZDFe model. Then, we use a nonlinear statistical model (random forests), to estimate the statistical parameters $\boldsymbol{\beta}_i$, giving us the distribution $[\boldsymbol{\alpha}_i|\boldsymbol{\theta}_i, \hat{\boldsymbol{\beta}}_i]$. This distribution allowed us to predict right singular vectors (corresponding to $\boldsymbol{\alpha}_i$) from biological parameters, $\boldsymbol{\theta}_i$. The construction of an approximate predictive distribution for $\boldsymbol{\alpha}_i$ given $\boldsymbol{\theta}_i$ using random forests is given in Appendix B, but it is possible to use another nonlinear statistical model (e.g., neural networks and radial basis functions) in place of random forests. Thus, we are left with the joint distribution for the process model:

$$\left[\boldsymbol{y}|\widetilde{\boldsymbol{\Phi}}_{\text{all}}, \boldsymbol{\alpha}_{\text{all}}\right] = \prod_{i=1}^{n} \left[\boldsymbol{y}(\boldsymbol{s}_i)|\widetilde{\boldsymbol{\Phi}}_i, \boldsymbol{\alpha}_i\right]$$

$$\left[\boldsymbol{\alpha}_{\text{all}}|\boldsymbol{\theta}_{\text{all}}, \hat{\boldsymbol{\beta}}_{\text{all}}\right] = \prod_{i=1}^{n} \left[\boldsymbol{\alpha}_i|\boldsymbol{\theta}_i, \hat{\boldsymbol{\beta}}_i\right]$$

where the "all" subscript refers to all n spatial locations.

3.3. Parameter model

As mentioned previously, vertical spatial variability is accounted for within the 1-D NPZDFe model itself. Rather than accounting for horizontal spatial variability through the coupling of the 1-D NPZDFe model to an ocean circulation model, we choose to allow the selected biological parameters to vary in space. Because for our example we have an a priori prescribed range for the parameters in our mechanistic model, we consider a transformation of the parameters to facilitate the use of a spatial Gaussian process model. That is, for a parameter vector $\boldsymbol{\gamma}_i = (\gamma_{1,i}, \ldots, \gamma_{p,i})'$, we provide lower and upper bounds for $\gamma_{j,i}$, a_j, and b_j, respectively. Then, we consider the transformation $\theta_{j,i} = \log((\gamma_{j,i} - a_j)/(b_j - \gamma_{i,i}))$, so that $\theta_{j,i} \in \mathbb{R}$. Next, we assign $\boldsymbol{\theta}_i$ a Gaussian prior with mean $\mathbf{0}_p$ and, for locations (x_i, y_i) and $(x_{i'}, y_{i'})$, $\text{cov}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{i'}) = \text{diag}\left(\tau_1^2 \rho_1^{||\boldsymbol{A}\boldsymbol{h}||}, \ldots, \tau_p^2 \rho_p^{||\boldsymbol{A}\boldsymbol{h}||}\right)$, where $\boldsymbol{h} = (x_i - x_{i'}, y_i - y_{i'})'$ and $\boldsymbol{A}$ is the matrix:

$$\boldsymbol{A} = \begin{bmatrix} 1/a_{\max} & 0 \\ 0 & 1/a_{\min} \end{bmatrix} \begin{bmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{bmatrix}$$

Here, $a_{\max}$ is the range of the spatial dependence in direction ϕ, and $a_{\min}$ is the range in the direction $\phi \pm 90°$ (see, e.g., Cressie, 1993; Cressie and Wikle, 2011). The angle ϕ is fixed to allow maximum range in the alongshelf direction (and the minimum range is in the cross-shelf direction). The decision to use an anisotropic correlation matrix with this property is based on published research concerning alongshelf dependence relative to cross-shelf dependence (e.g., see Brown and Fiechter, 2012).

We use a Uniform(0, 1) prior for each of the spatial correlation parameters, ρ_j, $j = 1, \ldots, p$, and we use a log-normal prior for $\sigma^2(\boldsymbol{s}_i)$, $\log(\sigma^2(\boldsymbol{s}_i)) \sim N\left(0, \sigma_\sigma^2\right)$, for $i = 1, \ldots, n$. For the NPZDFe model parameters, $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$, we want the prior to be relatively noninformative regarding central tendencies of the parameter. If one was to place a Uniform(a_j, b_j) prior on $\gamma_{j,i}$, this would be similar to a $N(0, 2^2)$ prior on $\theta_{j,i}$. Thus, we fix $\tau_j^2 = 4$ for $j = 1, \ldots, p$.

Posterior distribution/sampling. The aforementioned hierarchical model formulation implies the following posterior distribution:

$$\left[\boldsymbol{\alpha}_{\text{all}}, \boldsymbol{\theta}_{\text{all}}, \boldsymbol{\sigma}^{\boldsymbol{2}}, \boldsymbol{\rho}|z\right] \propto \left[z|\boldsymbol{y}, \boldsymbol{\sigma}^{\boldsymbol{2}}\right] \left[\boldsymbol{y}|\widetilde{\boldsymbol{\Phi}}_{\text{all}}, \boldsymbol{\alpha}_{\text{all}}\right] \left[\boldsymbol{\alpha}_{\text{all}}|\boldsymbol{\theta}_{\text{all}}, \hat{\boldsymbol{\beta}}_{\text{all}}\right] [\boldsymbol{\theta}_{\text{all}}|\boldsymbol{\rho}]\, [\boldsymbol{\sigma}^{\boldsymbol{2}}][\boldsymbol{\rho}]$$

where $[\boldsymbol{\theta}_{\text{all}}|\boldsymbol{\rho}]$ represents the prior distribution for $\boldsymbol{\theta}_{\text{all}} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$, where $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_p)'$. For this application, our primary interest is in the estimation of the 1-D NPZDFe model parameters, rather than both the estimation of the parameters and the process. We note that for our process distribution, $\left[\boldsymbol{y}|\widetilde{\boldsymbol{\Phi}}_{\text{all}}, \boldsymbol{\alpha}_{\text{all}}\right] = \prod_{i=1}^{n} \left[\boldsymbol{y}(\boldsymbol{s}_i)|\widetilde{\boldsymbol{\Phi}}_i, \boldsymbol{\alpha}_i\right]$, $\left[\boldsymbol{y}(\boldsymbol{s}_i)|\widetilde{\boldsymbol{\Phi}}_i, \boldsymbol{\alpha}_i\right]$ is a degenerate distribution with all of the probability mass at $\widetilde{\boldsymbol{\Phi}}_i \boldsymbol{\alpha}_i$. Our process model only accounts for uncertainty in the use of the emulator with respect to the mechanistic model, rather than uncertainty in the underlying marine ecosystem process. So we consider the integrated posterior

$$\left[\boldsymbol{\theta}_{\text{all}}, \boldsymbol{\sigma}^{\boldsymbol{2}}, \boldsymbol{\rho}|z\right] \propto \int \left[z|\boldsymbol{y}, \boldsymbol{\sigma}^{\boldsymbol{2}}\right] \left[\boldsymbol{y}|\widetilde{\boldsymbol{\Phi}}_{\text{all}}, \boldsymbol{\alpha}_{\text{all}}\right] \left[\boldsymbol{\alpha}_{\text{all}}|\boldsymbol{\theta}_{\text{all}}, \hat{\boldsymbol{\beta}}_{\text{all}}\right] [\boldsymbol{\theta}_{\text{all}}|\boldsymbol{\rho}] \left[\boldsymbol{\sigma}^{\boldsymbol{2}}\right] [\boldsymbol{\rho}]\, \mathrm{d}\boldsymbol{\alpha}_{\text{all}}$$

This integration is carried out within the context of an Markov chain Monte Carlo algorithm. Specifically, during a Metropolis–Hastings update for the parameter $\boldsymbol{\theta}_i$, for a given proposal $\boldsymbol{\theta}_i^*$, we generate a realization $\boldsymbol{\alpha}_i^*$ from $\left[\boldsymbol{\alpha}_i | \boldsymbol{\theta}_i^*, \hat{\boldsymbol{\beta}}_i\right]$ as mentioned in Section 3.2. Then, $\boldsymbol{\alpha}_i^*$ is used in the likelihood portion of the Metropolis–Hastings acceptance probability.

4. APPLICATION

4.1. Proof-of-concept example

We briefly discuss a proof-of-concept example to illustrate the ability of the emulator to act as a surrogate for the 1-D NPZDFe model. A first-order emulator is constructed (and used inside a Gibbs sampler) with simulated observations from the 1-D NPZDFe model. For consistency with the application, we use locations corresponding to the sites of the application (i.e., the coordinates were set to be the same as the locations on the CGOA). Parameter sets $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ are simulated from a Gaussian distribution with mean $\mathbf{0}_p$ and spatial covariance function provided in Section 3.3. For $\boldsymbol{\theta}_i = (\theta_{1,i}, \theta_{2,i})'$, $\theta_{1,i}$ represents the `KFeC` parameter and $\theta_{2,i}$ represents the `ZooGR` parameter. Known inputs (e.g., mixed layer depth and short wave radiation data) at each location for the 1-D NPZDFe model were different, so we construct $n = 9$ emulators, one for each location. For the covariance and spatial dependence parameters, we set $\tau_1^2 = 4$, $\tau_2^2 = 4$, $\rho_1 = 0.95$, and $\rho_2 = 0.2$ (i.e., we simulate one parameter to have high spatial dependence and another to have low spatial dependence). The parameters ϕ, $a_{\min}$, and $a_{\max}$ are fixed at their "true" values (i.e., those values used to simulate the data). The values $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ are then transformed back to $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_n$, and the 1-D NPZDFe model is run. At each location, only the output for the surface values of phytoplankton is retained in order to investigate our ability to recover parameter values with only SeaWiFS surface phytoplankton data. Also, at one location (with coordinates corresponding to the outer shelf off Kodiak Island), 80% of the monthly observations were removed at random, in order to illustrate the usefulness of the spatial model at locations where fewer data are available. For construction of the emulator, only two right singular vectors are retained. For each location/emulator, this accounts for over 97% of the variation in the mechanistic model output.

At all locations, the data are generated as realizations from a truncated normal distribution with mean parameter equal to the surface phytoplankton "truth" output and variance parameter equal to 0.5^2. We plot the "true" process at the inner shelf off of each location off Kodiak Island, simulated from the 1-D NPZDFe model, along with the mean of the approximate predictive distribution for the emulator given the "true" parameter values (Figure 2). The emulator at the "true" parameter values is nearly identical to the mechanistic model output at those same locations, but we point out the inability to capture the "valley" between the spring and fall blooms. Further, a Metropolis–Hastings within Gibbs sampler is used to obtain samples from the posterior distribution. This sampler is run for 100,000 iterations with a burn-in period of 50,000 iterations, and thinning every 10th iteration, leaving a posterior sample size of 5000. In addition to the model described in Section 3, we also run a Gibbs sampler where ρ_1 and ρ_2 are fixed at zero, in order to assess any benefit of being able to borrow strength across space. In both situations, the posterior distributions for all parameters contain the "true" values. With regard to the biological parameters, this suggests again that the emulator can act as a surrogate for the 1-D NPZDFe model. Further, the posterior distributions for the `KFeC` parameter in the linked model (i.e., the model where ρ_1 and ρ_2 are not fixed at zero) are significantly smaller in width (less than half the width) at five of the nine locations.

4.2. Application to coastal Gulf of Alaska

We use the model described in Section 3 applied to phytoplankton data for the CGOA. We use monthly averages of SeaWiFS chlorophyll measurements at the nine locations. The SeaWiFS observations are transformed by multiplying by the appropriate carbon-to-chlorophyll and nitrogen-to-carbon contents within phytoplankton cells and are thus matched up with the phytoplankton output of the 1-D NPZDFe model. One extension to the model described in Section 3 is that we allow for a seasonally varying variance parameter.

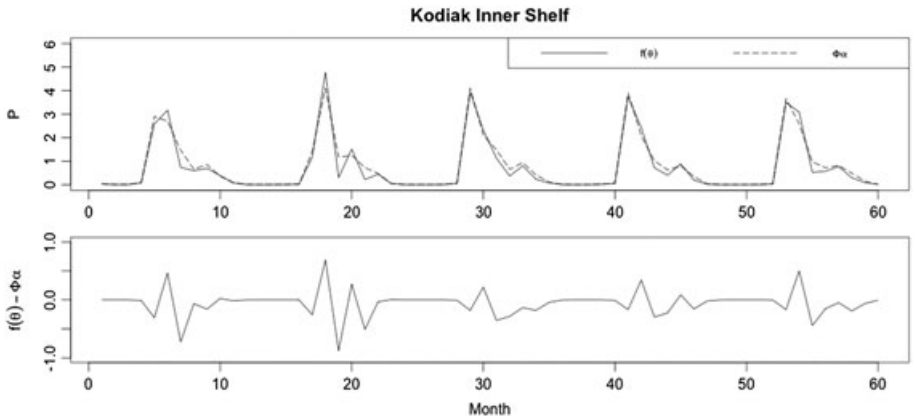

Figure 2. The upper panel is a plot of the true simulated process (solid line) for surface phytoplankton, along with the prediction (dashed line) using the first-order emulator and the true parameter values. The lower panel shows the difference between the simulation with the true model ($f(\boldsymbol{\theta})$) and the emulator ($\boldsymbol{\Phi\alpha}$). Units on the y-axis are mmolN m^{-3}

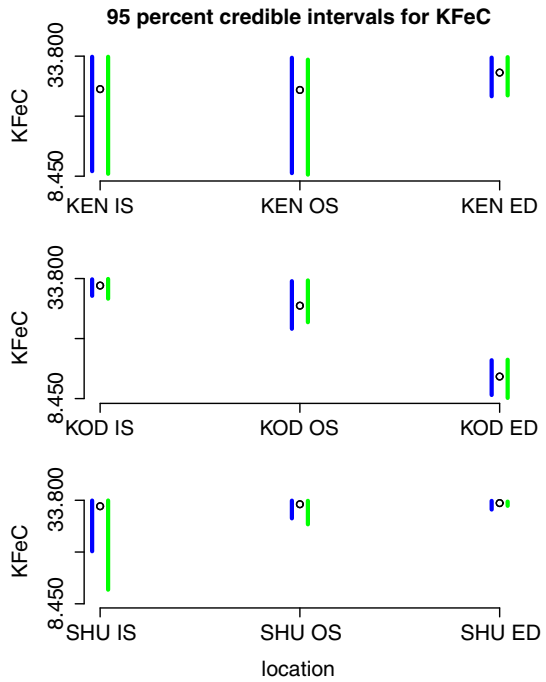

Figure 3. A plot of the 95% credible intervals for the posterior distribution of half-saturation constant for iron, using the SeaWiFS data. The line segment on the left at each location is the credible interval for the model with the spatial prior on the parameters. The line segment on the right at each location is the credible interval for the model without the spatial prior on the parameters. The circle in the middle of the two lines is the parameter value from which the data were simulated at each location. At the location with coordinates corresponding to the outer shelf off Kodiak Island, 80% of the data were removed at random

The 95% credible intervals for ρ_1 and ρ_2 are $(0.030, 0.657)$ and $(0.031, 0.488)$, respectively, suggesting low-to-moderate spatial dependence for both parameters. In Figure 3, the 95% credible intervals for the posterior distribution of `KFeC` are plotted by location. The least amount of uncertainty in the estimates seems to be off Kodiak Island, where estimates decrease moving further offshore. Figure 4 shows the credible intervals for `ZooGR`. Overall, there is less uncertainty in the estimates for `ZooGR` than `KFeC`. The estimates for zooplankton grazing rate are near the upper bound of the range, with the exception of the inner shelf location off Kodiak Island, where `ZooGR` is estimated to be near the lower limit of the range. For the most part, there is little difference between the spatial and nonspatial models with regard to the estimates for `ZooGR`. However, for `KFeC`, we see a reduction in the credible interval width at the inner shelf location off Shumagin island.

Figure 5 shows 95% prediction intervals for phytoplankton off the Kenai Peninsula, Kodiak Island, and Shumagin Islands, along with the data for all locations. Uncertainty in the process is highest during the fall bloom on the inner shelf off Kodiak Island and during the spring bloom on the outer shelf off Shumagin Island.

5. DISCUSSION

Allowing parameters to vary spatially is a scientifically plausible idea for marine ecosystems based on previous research (Friedrichs *et al.*, 2007). By doing so, we reduce the uncertainty in the parameter estimates and for the half-saturation constant for iron on the inner shelf off the Shumagin Islands. Zooplankton grazing rates at all but one location were estimated to be near the upper limit of the a priori prescribed range, which suggests that either the range is too restrictive (i.e., zooplankton grazing pressure is underestimated) or that other processes contributing to phytoplankton loss (e..g, phytoplankton mortality) are underestimated.

Sparse and uncertain data make it difficult to model primary production in marine ecosystems. For that reason, mechanistic models that take into account complex relationships between the different parameters and response variables are a critical part of understanding which processes control lower trophic level marine ecosystem dynamics. Although uncertainty is assessed through uncertainty analysis and sensitivity analysis, and although parameter and process inference can be performed somewhat through model calibration and prediction, the inclusion of these models into a BHM allows us to assess uncertainty in the data, process, and parameters, as well as uncertainty in the mechanistic model.

Sometimes, a mechanistic model may be prohibitively expensive computationally to run iteratively in a Gibbs sampler, making it impossible to sample from the posterior distribution for Bayesian inference. First-order emulators for mechanistic models allow for a simple and straightforward way to include important dynamics from a mechanistic model that may be too computationally expensive to run in an iterative way. Further, using an emulator for a complex mechanistic model, as opposed to a simpler mechanistic model, may include important dynamics from the complex model that the simpler mechanistic model may have excluded.

Emulators are built on a limited number of runs of the true mechanistic model, in order to predict the output of the mechanistic model at untried input settings. As such, the emulator may not be reasonable for predicting outside the range of inputs in the training data set.

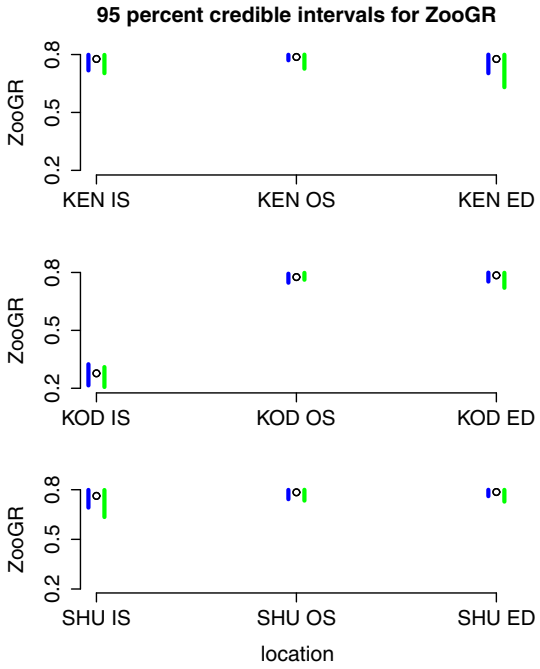

Figure 4. A plot of the 95% credible intervals for the posterior distribution of zooplankton grazing rate, using the SeaWiFS data. The line segment on the left at each location is the credible interval for the model with the spatial prior on the parameters. The line segment on the right at each location is the credible interval for the model without the spatial prior on the parameters. The circle in the middle of the two lines is the posterior mean for the spatial model

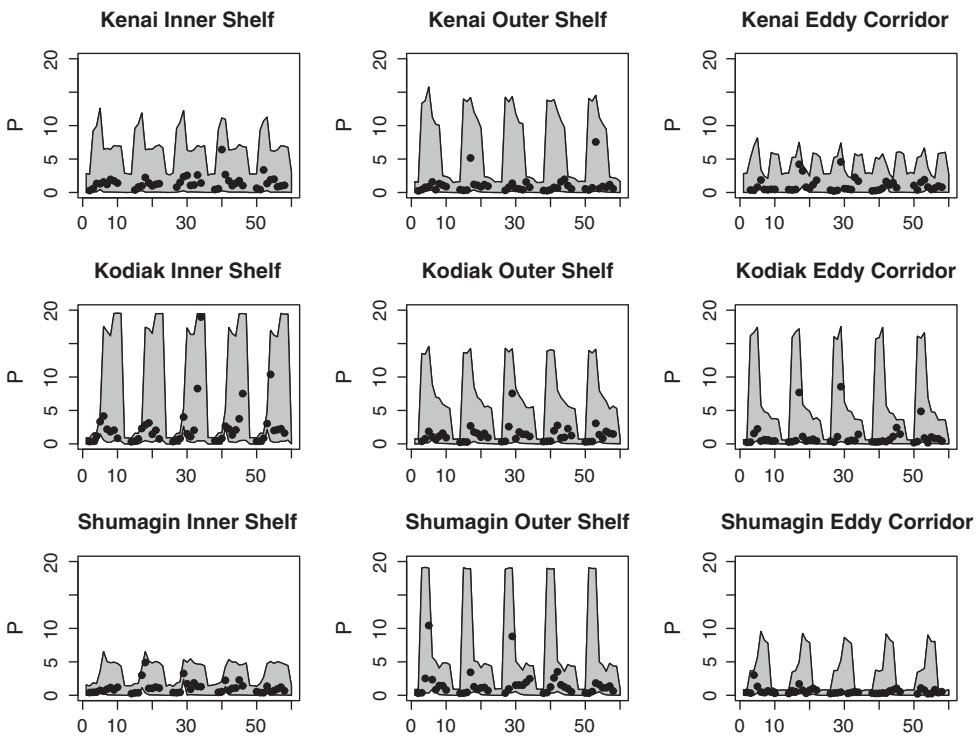

Figure 5. A plot 95% prediction intervals of the state process (shaded region) and the data (circles). Units on the y-axis are mmolN m^{-3}. Units on the x-axis are months. The reference time (e.g., $t = 0$) is January of 1998

Further, in the case of certain models, nonlinearity may cause regime shifts outside the training data set. In these situations, it may be more appropriate to consider the true mechanistic model. If this model is still too computationally expensive to run iteratively in a Gibbs sampler, then one might consider a mechanistic model that includes important dynamics yet still is computationally efficient.

As an extension, we could expand this framework to include more parameters and more locations, as well as to allow for more flexible fitting of the spatial covariance matrix (e.g., allowing $a_{\min}$ and $a_{\max}$ to vary). Further, because the posterior estimates for zooplankton grazing rate pushed the upper bounds of the a priori range, we may consider constructing an emulator on the basis of new ensembles of the 1-D NPZDFe model, where the range is increased for zooplankton grazing rate, or phytoplankton mortality is allowed to vary. We could also create an emulator that considers nutrient output in addition to phytoplankton and could include observations at multiple depths.

Although we did have a process stage in our model, it does not explicitly account for uncertainty in the process or uncertainty in the mechanistic model itself. Rather, it only accounts for uncertainty in the use of an emulator in place of the 1-D NPZDFe model. The former types of uncertainty are accounted for indirectly and lumped together with measurement error and small-scale variability in the data model. In the future, we want to include a process stage that deals with model uncertainty and process uncertainty, also accounting for extra spatial variability in the process not accounted for by the spatially varying parameters. Lastly, the truncated normal distribution is biased, in the sense that the expected value of a truncated normal random variable is not the realization from our emulator. Using a bias correction procedure would correct this problem and potentially reduce uncertainty (e.g., Cangelosi and Hooten, 2009). However, this would be at the cost of increasing computational time in the Gibbs sampler.

Finally, we note that the methodology described here is a special case of data assimilation, where we are seeking to fuse data and scientific knowledge to estimate parameters and predict state processes in the presence of uncertainty. The literature on ocean data assimilation is quite large (e.g., see the overviews of Bennett (2002); Bertino *et al.* (2003); Evensen (2009) as well as the important ensemble example for the 1-D marine ecosystem model by Eknes and Evensen (2002)). To date, the traditional 4-D variational approaches for data assimilation for high-resolution regional ocean models are only now just starting to deal with the assimilation of biological variables (with the obvious limitation of the non-Gaussian nature of the state variables and the constrained support of the state processes and parameters). However, such approaches are more suitable in general than the niche model presented here. However, the approach described here is very cheap computationally and may provide quick solutions in problems for which more traditional data assimilation is currently untenable.

Acknowledgements

The authors thank Mevin Hooten, Andrew Moore, Zach Powell, and Nadia Pinardi for their helpful feedback and interesting discussions at annual meetings. Further, we thank the AE and three anonymous reviewers for helpful feedback that greatly enhanced the clarity of this manuscript. Funding for this project was provided through DMS-1049093 and OCE-0814934.

REFERENCES

Aita M, Yamanaka Y, Kishi M. 2007. Interdecadal variation of the lower trophic ecosystem in the northern Pacific between 1948 and 2002, in a 3-D implementation of the NEMURO model. *Ecological Modelling* **202**(1-2): 81–94.
Bennett A. 2002. *Inverse Modeling of the Ocean and Atmosphere*. Cambridge University Press: Cambridge.
Berliner L. 1996. Hierarchical Bayesian time-series models. *Fundamental Theories of Physics* **79**: 15–22.
Berliner L. 2003. Physical–statistical modeling in geophysics. *Journal of Geophysical Research* **108**(8776): STS 3-1–STS 3-10.
Berrocal V, Gelfand A, Holland D. 2010. A spatio-temporal downscaler for output from numerical models. *Journal of Agricultural, Biological, and Environmental Statistics* **15**(2): 176–197.
Bertino L, Evensen G, Wackernagel H. 2003. Sequential data assimilation techniques in oceanography. *International Statistical Review* **71**(2): 223–241.
Breiman L. 2001. Random forests. *Machine Learning* **45**(1): 5–32.
Brown J, Fiechter J. 2012. Quantifying eddychlorophyll covariability in the coastal Gulf of Alaska. *Dynamics of Atmospheres and Oceans* **5556**(0): 1–21.
Cangelosi A, Hooten M. 2009. Models for bounded systems with continuous dynamics. *Biometrics* **65**(3): 850–856.
Conti S, Gosling J, Oakley J, O'Hagan A. 2009. Gaussian process emulation of dynamic computer codes. *Biometrika* **96**(3): 663–676.
Conti S, OHagan A. 2010. Bayesian emulation of complex multi-output and dynamic computer models. *Journal of Statistical Planning and Inference* **140**(3): 640–651.
Cressie N. 1993. *Statistics for Spatial Data*. Wiley: New York.
Cressie N, Wikle C. 2011. *Statistics for Spatio-Temporal Data*. Wiley: New York.
Currin C, Mitchell T, Morris M, Ylvisaker D. 1991. Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. *Journal of the American Statistical Association* **86**: 953–963.
Dowd M. 2006. A sequential Monte Carlo approach for marine ecological prediction. *Environmetrics* **17**(5): 435–455.
Dowd M. 2007. Bayesian statistical data assimilation for ecosystem models using Markov chain Monte Carlo. *Journal of Marine Systems* **68**(3-4): 439–456.
Dowd M. 2011. Estimating parameters for a stochastic dynamic marine ecological system. *Environmetrics* **22**: 501–515.
Drignei D. 2008. Fast statistical surrogates for dynamical 3D computer models of brain tumors. *Journal of Computational and Graphical Statistics* **17**(4): 844–859.
Eknes M, Evensen G. 2002. An ensemble Kalman filter with a 1-D marine ecosystem model. *Journal of Marine Systems* **36**(1): 75–100.
Evensen G. 2003. The ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean Dynamics* **53**(4): 343–367.
Evensen G. 2009. *Data Assimilation: The Ensemble Kalman Filter*. Springer: New York.
Fiechter J. 2012. Assessing marine ecosystem model properties from ensemble calculations. *Ecological Modelling* **242**: 164–179.
Fiechter J, Broquet G, Moore A, Arango H. 2011. A data assimilative, coupled physical-biological model for the coastal Gulf of Alaska. *Dynamics of Atmospheres and Oceans* **51**(3): 75–98.
Fiechter J, Moore A, Edwards C, Bruland K, Di Lorenzo E, Lewis C, Powell T, Curchitser E, Hedstrom K. 2009. Modeling iron limitation of primary production in the coastal Gulf of Alaska. *Deep Sea Research Part II: Topical Studies in Oceanography* **56**(24): 2503–2519.
Finley A, Banerjee S, Basso B. 2011. Improving crop model inference through Bayesian melding with spatially varying parameters. *Journal of Agricultural, Biological, and Environmental Statistics* **16**(4): 453–474.

Franks P. 2002. NPZ models of plankton dynamics: their construction, coupling to physics, and application. *Journal of Oceanography* **58**(2): 379–387.
Friedrichs M, Dusenberry J, Anderson L, Armstrong R, Chai F, Christian J, Doney S, Dunne J, Fujii M, Hood R, McGillicuddy DJ, Moore JK, Schartau M, Spitz YH, Wiggert JD. 2007. Assessment of skill and portability in regional marine biogeochemical models: role of multiple planktonic groups. *Journal of Geophysical Research* **112**: 1–22.
Frolov S, Baptista A, Leen T, Lu Z, van der Merwe R. 2009. Fast data assimilation using a nonlinear Kalman filter and a model surrogate: an application to the Columbia River Estuary. *Dynamics of Atmospheres and Oceans* **48**(1-3): 16–45.
Fuentes M, Raftery A. 2005. Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. *Biometrics* **61**(1): 36–45.
Fujii M, Yamanaka Y, Nojiri Y, Kishi M, Chai F. 2007. Comparison of seasonal characteristics in biogeochemistry among the subarctic North Pacific stations described with a NEMURO-based marine ecosystem model. *Ecological Modelling* **202**(1-2): 52–67.
Harmon R, Challenor P. 1997. A Markov chain Monte Carlo method for estimation and assimilation into models. *Ecological modelling* **101**(1): 41–59.
Hastie T, Tibshirani R, Friedman J. 2009. *The Elements of Statistical Learning*. Springer: New York.
Higdon D, Gattiker J, Williams B, Rightley M. 2008. Computer model calibration using high-dimensional output. *Journal of the American Statistical Association* **103**(482): 570–583.
Higdon D, Kennedy M, Cavendish J, Cafeo J, Ryne R. 2004. Combining field data and computer simulations for calibration and prediction. *SIAM Journal on Scientific Computing* **26**: 448–466.
Hooker S, McClain C. 2000. The calibration and validation of SeaWIFS data. *Progress in Oceanography* **45**(3-4): 427–465.
Hooten M, Leeds W, Fiechter J, Wikle C. 2011. Assessing first-order emulator inference for physical parameters in nonlinear mechanistic models. *Journal of Agricultural, Biological, and Environmental Statistics* **16**(4): 475–494.
Jones E, Parslow J, Murray L. 2010. A Bayesian approach to state and parameter estimation in a phytoplankton–zooplankton model. *Australian Meteorological and Oceanographic Journal* **59**: 7–16.
Kennedy M, Anderson C, O'Hagan A, Lomas M, Woodward I, Gosling J, Heinemeyer A. 2008. Quantifying uncertainty in the biospheric carbon flux for England and Wales. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **171**(1): 109–135.
Kennedy M, O'Hagan A. 2001. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63**(3): 425–464.
Kishi M, Kashiwai M, Ware D, Megrey B, Eslinger D, Werner F, Noguchi-Aita M, Azumaya T, Fujii M, Hashimoto S, Huang D, Iizumi H, Ishida Y, Kang S, Kantakov G, Kim H, Komatsu K, Navrotsky V, Smtih S, Tadokor K, Tsuda A, Yamamura O, Yamanaka Y, Yokouchi K, Yoshie N, Zhang J, Zuenko Y, Zvalinsky V. 2007. NEMURO–a lower trophic level model for the North Pacific marine ecosystem. *Ecological Modelling* **202**(1-2): 12–25.
Large W, Yeager S. 2009. The global climatology of an interannually varying air–sea flux data set. *Climate Dynamics* **33**(2): 341–364.
Leeds W, Wikle C. 2012. Science-based parameterizations for dynamical spatio-temporal models. *WIREs Computational Statistics* **4**(6): 554–560.
Liu F, West M. 2009. A dynamic modelling strategy for Bayesian computer model emulation. *Bayesian Analysis* **4**(2): 393–412.
Malve O, Laine M, Haario H, Kirkkala T, Sarvala J. 2007. Bayesian modelling of algal mass occurrences–using adaptive MCMC methods with a lake water quality model. *Environmental Modelling & Software* **22**(7): 966–977.
Margvelashvili N, Campbell E. 2012. Sequential data assimilation in fine-resolution models using error-subspace emulators: theory and preliminary evaluation. *Journal of Marine Systems* **90**: 13–22.
Martin J, Gordon R, Fitzwater S, Broenkow W. 1989. Vertex: phytoplankton/iron studies in the Gulf of Alaska. *Deep Sea Research Part A. Oceanographic Research Papers* **36**(5): 649–680.
Mattern J, Fennel K, Dowd M. 2012. Estimating time-dependent parameters for a biological ocean model using an emulator approach. *Journal of Marine Systems* **96-97**: 32–47.
Milliff R, Bonazzi A, Wikle C, Pinardi N, Berliner L. 2012. Ocean ensemble forecasting. Part I: ensemble mediterranean winds from a Bayesian hierarchical model. *Quarterly Journal of the Royal Meteorological Society* **137**: 858–878.
O'Hagan A. 2006. Bayesian analysis of computer code outputs: a tutorial. *Reliability Engineering & System Safety* **91**(10): 1290–1300.
Poole D, Raftery A. 2000. Inference for deterministic simulation models: the Bayesian melding approach. *Journal of the American Statistical Association* **95**(452): 1244–1255.
Rougier J. 2008. Efficient emulators for multivariate deterministic functions. *Journal of Computational and Graphical Statistics* **17**(4): 827–843.
Royle A, Berliner LM, Wikle CK, Milliff R. 1999. 2012. A hierarchical spatial model for constructing wind fields from scatterometer data in the Labrador sea. In *Case Studies in Bayesian Statistics IV*, Gatsonis G, Carlin B, Gelman A, West M, Kass RE, Carriquiry A, Verdinelli I (eds). Springer-Verlag: New York; 367–272.
Sacks J, Welch W, Mitchell T, Wynn H. 1989. Design and analysis of computer experiments. *Statistical Science* **4**(4): 409–423.
van der Merwe R, Leen T, Lu Z, Frolov S, Baptista A. 2007. Fast neural network surrogates for very high dimensional physics-based models in computational oceanography. *Neural Networks* **20**(4): 462–478.
Wainwright TC, Feinberg LR, Hooff RC, Peterson WT. 2007. A comparison of two lower trophic models for the California Current system. *Ecological Modelling* **202**(1-2): 120–131.
Ward B, Friedrichs M, Anderson T, Oschlies A. 2010. Parameter optimisation techniques and the problem of underdetermination in marine biogeochemical models. *Journal of Marine Systems* **81**(1-2): 34–43.
Wikle C. 2003. Hierarchical Bayesian models for predicting the spread of ecological processes. *Ecology* **84**(6): 1382–1394.
Wikle C, Berliner L. 2005. Combining information across spatial scales. *Technometrics* **47**(1): 80–91.
Wikle C, Hooten M. 2010. A general science-based framework for dynamical spatio-temporal models. *Test* **19**: 1–35.
Wikle C, Milliff R, Nychka D, Berliner L. 2001. Spatiotemporal hierarchical Bayesian modeling of tropical ocean surface winds. *Journal of the American Statistical Association* **96**(454): 382–397.

APPENDIX A. THE 1-D NPZDFE MODEL

We provide a brief description here of the 1-D NPZDFe model for which the emulator was constructed. Appropriate units of measurement, along with default values for parameters are provided in parentheses after they are mentioned. Readers interested in further details will find them provided in Fiechter *et al.* (2009). First, the system contains an NPZD model:

$$\frac{\partial N}{\partial t} = \delta D + \gamma_n GZ - UP$$
$$\frac{\partial P}{\partial t} = UP - GZ - \sigma_d P$$

$$\frac{\partial Z}{\partial t} = (1-\gamma_n)GZ - \zeta_d Z$$
$$\frac{\partial D}{\partial t} = \sigma_d P + \zeta_d Z + w_d \frac{\partial D}{\partial z}$$

where N, P, Z, and D refer to concentrations of nitrate, phytoplankton, zooplankton, and detritus, respectively. Note that the system of equations also contains a lowercase "z," which refers to vertical depth and should not be confused with the uppercase "Z," which corresponds to zooplankton concentration. The following model contains linear parameters: the detritus remineralization rate (δ; 1.0 day^{-1}), zooplankton excretion efficiency (γ_n; 0.3), phytoplankton senescence (σ_d; 0.1 day^{-1}), zooplankton mortality (ζ_d; 0.145 day^{-1}), and a detritus sinking term (w_d; 8.0 mday^{-1}), as well as nonlinear functions, such as the zooplankton growth rate:

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

with the following parameters: zooplankton grazing rate (R_m; 0.65 day^{-1}) and the Ivlev constant (Λ; 0.84). The model also includes a nitrate-limited phytoplankton growth rate:

$$U_N = \frac{V_m N}{N + k_N}\frac{\alpha I}{\sqrt{V_m^2 + \alpha^2 I^2}}$$

with additional parameters as follows: phytoplankton nitrate uptake rate (V_m; 1.0 day^{-1}), nitrate half-saturation constant (k_N; 1.0 mmolN m^{-3}), and the initial slope of P–I curve (α; 0.02 m^{-2} W^{-1}). This function also includes a term for light availability at depth (negative z):

$$I = I_0 \exp\left(k_z z + k_p \int_z^0 P(z')\mathrm{d}z'\right)$$

with the following parameters: light extinction coefficient (k_z; 0.067 m^{-1}) and phytoplankton self-shading coefficient (k_p; 0.04 m^2 mmolN $^{-1}$). Surface irradiance (I_0; W m^{-2}) is imposed as daily average short-wave radiation from the data sets for Common Ocean-Ice Reference Experiments (CORE2; Large and Yeager, 2009).

In addition, iron limitation is included in the model via governing equations for P-associated iron (F_p),

$$\frac{\partial F_p}{\partial t} = F_p\left(U - \frac{GZ}{P} - \sigma_d\right) + L_{Fe}$$

and dissolved iron (F_d),

$$\frac{\partial F_d}{\partial t} = F_p\left(f_{\mathrm{rem}}\left(\frac{GZ}{P} + \sigma_d\right) - U\right) - L_{Fe}$$

which includes a parameter for the iron remineralization fraction (f_{rem}; 0.5). Other functions in the model include iron uptake by phytoplankton:

$$L_{\mathrm{Fe}} = \frac{R_0 - R}{t_{Fe}} P[C:N]$$

which depends on the iron uptake time scale (t_{Fe}; 1.0 day), the Redfield carbon-to-nitrogen ratio ($[C:N] = 106:16$ molC/molN), and empirically determined and realized iron-to-carbon [Fe:C] ratios:

$$R_0 = bF_d^a, R = \frac{F_p}{P[C:N]}$$

where a (0.6) and b (64 (molC m^{-3})$^{-1}$) are, respectively, the coefficient and power for estimating the empirical phytoplankton iron-to-carbon ratio based on dissolved iron concentration. Maximum phytoplankton growth under both nitrate and iron limitation is then determined as

$$U = \min\left(R^2/\left(R^2 + k_{Fe}^2, U_N\right)\right.$$

which includes the half-saturation constant for iron (k_{Fe}; 16.9 μmolFe (molC)$^{-1}$). The initial condition for dissolved iron concentrations is based on *in situ* measurements for the Gulf of Alaska (Martin *et al.*, 1989) and defined as

$$F_{d,\mathrm{clim}} = F_{d,\max} + c_{Fe}(F_{d,\min} - F_{d,\max})$$

which includes a minimum for dissolved iron concentrations offshore ($F_{d,\min}$; 0.05 μmolFe m^{-3}) and a maximum for dissolved iron concentrations on the shelf ($F_{d,\max}$; 2.0 μmolFe m^{-3}), and

$$c_{Fe} = \max\left(0, \min\left(1, \frac{h - h_{\min}}{h_{\max} - h_{\min}}\right)\right)$$

sets a linear transition (based on total water depth) between elevated iron concentrations inshore of the 200-m isobath (i.e., $h_{\min} = 200$ m) and depleted iron concentrations offshore of the 1500-m isobath (i.e., $h_{\max} = 1500$ m).

Because phytoplankton growth is also modulated by vertical mixing (a process not included in the 1-D NPZDFe model), we use mixed layer depth information from a data assimilative, regional ocean circulation regional ocean modeling system (ROMS) model for the CGOA (Fiechter *et al.*, 2011) to limit phytoplankton growth rates during periods of intense vertical mixing (i.e., a stratified water column allows phytoplankton to remain near the surface and grow more rapidly in response to increased light availability).

APPENDIX B. FIRST-ORDER EMULATORS

We use a first-order emulator, as outlined in Hooten *et al.* (2011), to construct a computationally affordable statistical surrogate for a deterministic mechanistic model, $f(\boldsymbol{c}, \boldsymbol{\theta})$, where $\boldsymbol{c}$ is a vector representing known forcings and $\boldsymbol{\theta}$ is a vector of unknown parameters. For notational convenience, we will limit ourselves to discussing the development of a first-order emulator for one variable at one spatial location.

First, N parameter vectors $\boldsymbol{\Theta} = \left(\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(N)}\right)$ are selected, where $\boldsymbol{\theta}^{(k)} = \left(\theta_1^{(k)}, \ldots, \theta_p^{(k)}\right)'$. These can be selected randomly from the prior, but some methods such as a latin hypercube design may be preferred, in order to make sure the parameter space is explored as efficiently as possible (Sacks *et al.*, 1989). Using these inputs, we generate outputs $\boldsymbol{y}^{(1)} \equiv f(\boldsymbol{c}, \boldsymbol{\theta}^{(1)}), \ldots, \boldsymbol{y}^{(N)} \equiv f(\boldsymbol{c}, \boldsymbol{\theta}^{(N)})$, placed in the $T \times N$ matrix, $\boldsymbol{Y} \equiv (\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(N)})$. Then, $\boldsymbol{Y}$ is decomposed via the svd,

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}'$$

where $\boldsymbol{U}$ is a $T \times T$ matrix of left singular vectors, $\boldsymbol{D}$ is a $T \times N$ diagonal matrix with the jth singular value of $\boldsymbol{Y}$ in the jth row and jth column (and 0 elsewhere), and $\boldsymbol{V}$ is an $N \times N$ matrix of right singular vectors. We retain the first q columns of $\boldsymbol{U}\boldsymbol{D}$, denoted $\widetilde{\boldsymbol{\Phi}}$.

The first q right singular vectors in $\boldsymbol{V}$, $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_q$, are used as dependent variables to develop q nonlinear statistical models. That is, for each of these q right singular vectors of dimension N, the rows of $\boldsymbol{\Theta}$ are used as predictor variables in a nonlinear statistical model. For each $j = 1, \ldots, q$, we compute an approximate predictive distribution $[\alpha_j | \boldsymbol{\theta}, \boldsymbol{\beta}_j]$ to define the statistical relationship between $\boldsymbol{v}_j$ and $(\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(N)})$. So we estimate the $\boldsymbol{\beta}_j$ parameters, and as a result, we are left with a distribution $[\boldsymbol{\alpha} | \boldsymbol{\theta}, \hat{\boldsymbol{\beta}}] = [\alpha_1 | \boldsymbol{\theta}, \hat{\boldsymbol{\beta}}_1] \cdots [\alpha_q | \boldsymbol{\theta}, \hat{\boldsymbol{\beta}}_q]$. (Hooten *et al.*, 2011) modeled this input–output relationship by using the nonparametric bagged-regression tree approach, random forests (Breiman, 2001). Regression trees are robust and low-bias estimators, but they have high variance. Random forests use regression trees, with two substantial modifications. First, it is by using a bootstrap aggregation approach, whereby bootstrap samples from the complete data set are used to create a regression tree, that the prediction at a point is the average of the predictions for each tree and the variance is reduced. Specifically, for the jth right singular vector, we consider input $\left\{\left(v_j^{(1)}, \boldsymbol{\theta}^{(1)}\right), \ldots, \left(v_j^{(N)}, \boldsymbol{\theta}^{(N)}\right)\right\}$. Then, B bootstrap samples are taken from $\left\{\left(v_j^{(1)}, \boldsymbol{\theta}^{(1)}\right), \ldots, \left(v_j^{(N)}, \boldsymbol{\theta}^{(N)}\right)\right\}$. For each bootstrap sample, a regression tree T_b is constructed. Thus, for a new input $\boldsymbol{\theta}^*$, we predict $\hat{\alpha}_j(\boldsymbol{\theta}^*) = \frac{1}{B}\sum_{b=1}^{B} T_{b,j}(\boldsymbol{\theta}^*)$.

A second benefit of using random forests is that using only a subset of the input variables at each step reduces the correlation between the trees, which reduces the variance further (see Hastie *et al.*, 2009, for further details). As a result, our random forest implementation has both low bias and low variance, and $\boldsymbol{\theta}$ should be a good predictor of $\boldsymbol{\alpha}$.

Importantly, random forests use out-of-bag samples, so that the prediction for a particular response variable is based only on the average of trees constructed from bootstrap samples that did not contain that particular response variable (Hastie *et al.*, 2009). In other words, when the jth right singular vector for a known input $\boldsymbol{\theta}^{(k)}$ is predicted, the random forest prediction $\hat{\alpha}_j(\boldsymbol{\theta}^{(k)})$ is based on only regression trees that are constructed when $\left(v_j^{(k)}, \boldsymbol{\theta}^{(k)}\right)$ was not part of the bootstrap data set. Thus, $\eta_j^{(k)} = \hat{\alpha}_j(\boldsymbol{\theta}^{(k)}) - v_j^{(k)}$, for $k = 1, \ldots, N$, can be thought of as a sample of N true predictive residuals. Then, we use this to construct an estimate of the true predictive distribution for the random forest, and define the process model

$$\boldsymbol{\alpha} | \boldsymbol{\theta} \sim \left[\boldsymbol{\alpha} | \boldsymbol{\theta}, \hat{\boldsymbol{\beta}}\right]$$

To obtain a sample from this distribution, we first compute the random forest prediction for each of the q elements of $\boldsymbol{\alpha}$, denoted $\hat{\boldsymbol{\alpha}} = \left(\hat{\alpha}_1(\boldsymbol{\theta}), \ldots, \hat{\alpha}_q(\boldsymbol{\theta})\right)'$, given the parameter vector $\boldsymbol{\theta}$. Then, for $j = 1, \ldots, q$, we add a bootstrap sample $\hat{\eta}_j$ from $\left\{\eta_j^{(1)}, \ldots, \eta_j^{(N)}\right\}$. The realization from the distribution is then $\boldsymbol{\alpha}^* = \hat{\boldsymbol{\alpha}} + \hat{\boldsymbol{\eta}}$, where $\hat{\boldsymbol{\eta}} = \left(\hat{\eta}_1, \ldots, \hat{\eta}_q\right)'$. We use this to effectively "integrate out" the uncertainty due to use of a statistical surrogate (e.g., see Hooten *et al.*, 2011).

