Journal of Marine Systems 76 (2009) 162–170

Contents lists available at ScienceDirect

Journal of Marine Systems

journal homepage: www.elsevier.com/locate/jmarsys

Parameter estimation using an ensemble smoother: The effect of the circulation in biological estimation

Keston W. Smith [a,*], Dennis J. McGillicuddy Jr. [a], Daniel R. Lynch [b]

[a] Woods Hole Oceanographic Institution, Woods Hole, MA 02543, USA
[b] Dartmouth College, Hanover NH, MA 03755, USA

ARTICLE INFO

Article history:
Received 27 July 2007
Received in revised form 21 February 2008
Accepted 2 May 2008
Available online 29 May 2008

Keywords:
Cross-validation
Ensemble smoother
Data assimilation
Skill assessment

ABSTRACT

An ensemble smoother is used to estimate the initial conditions and mortality rates for a spatially explicit model of *Alexandrium fundyense*. The data assimilation procedure is effective at reducing model-data misfit in this strong constraint problem formulation. The skill of this estimation procedure is assessed through cross-validation. The estimation is carried out with three different representations of circulation: no flow, climatology, and a data assimilative hindcast. Although the misfit to the assimilated data is lowest with no flow, the skill of the biological hindcast is best with the hindcast and climatological velocity fields. Mortality estimates fall within the range of observed values, but the inferred spatial structure is not testable with existing data.

1. Introduction

A common theme in marine coupled physical–biological models is the interaction of ocean currents with biological processes via advection. In problems that assimilate measurements to infer the values of the parameters of the biological model, it is natural to question the impact of the presumed advection on the biological estimation. We seek approaches that are directly applicable to field data and spatially explicit data-assimilative computational models, focusing on methods that facilitate quantitative skill assessment and process-level diagnosis of the underlying solutions.

The data assimilation problem can be formulated in several ways. In filtering applications, state estimates are computed using only data that precede the analysis time. In smoothing applications, all data collected over some time interval are used to estimate the model state at all time points, leading to state estimates conditioned on the full data set. Methods that require strict solution of the underlying model equations are commonly referred to as "strong constraint," whereas "weak constraint" methods allow for departures from the model dynamics.

A variety of mathematical approaches are available for solving both weak and strong constraint problems for filtering and smoothing applications. Monte Carlo ensemble methods, such as the ensemble Kalman filter (EnKF), ensemble Kalman Smoother (EnKS), and ensemble smoother (EnS), utilize statistical approximations to Bayesian state estimation with the assumption that the estimator is a linear function of the data. Variational techniques such as the adjoint method are typically employed in strong-constraint formalism to minimize the misfit between predictions and observations (a "cost function") through adjustment of model inputs such as initial conditions, boundary conditions, and model parameters. Formally, these strong constraint methods are analogous to the Bayesian estimation of the control variables, with the cost function being proportional to the log likelihood over the uncertain parameters and data.

The particular coupled physical–biological problem of interest here is posed as an advection–diffusion–reaction equation for the concentration of a toxic dinoflagellate species, *Alexandrium fundyense* (Anderson, 1997). In this context, the reaction term represents the population dynamics of the organism, some aspects of which are relatively well constrained by laboratory measurements (e.g. growth), and others that are

* Corresponding author.
E-mail address: kwsmith@whoi.edu (K.W. Smith).

0924-7963/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jmarsys.2008.05.017

not (e.g. mortality). Our goal is to test this model with shipboard survey data that are non-uniformly distributed in space and time, inverting for initial conditions and the more poorly constrained aspects of the population dynamics (mortality). Sensitivity to the prescribed velocity field is of key importance, given the strongly advective environment in which this study is carried out.

The nature of this endeavor lends itself to a strong constraint methodology. Although adjoint methods have shown promise in spatially-explicit coupled physical–biological problems (Li et al., 2006; Matear and Holloway, 1995; McGillicuddy and Bucklin, 2002; McGillicuddy et al., 1998), in this instance we have chosen a strong constraint EnS approach. Although filtering formulations of this approach (the EnKF) has been applied extensively to coupled physical–biological models of the ocean (Allen et al., 2002; Eknes and Evensen, 2002; Natvik and Evensen, 2003a; Natvik and Evensen, 2003b), applications to strong constraint smoothing problems in the ocean have thus far been limited to spatially aggregated models (Annan and Hargreaves, 2004).

The general approach to ensemble smoothing is presented in text books such as Tarantola (2005) and Evensen (2006), though the Monte Carlo implementation is not directly advocated therein. Monte Carlo methods of data assimilation hold some practical advantages over the variational approach, namely they are “embarrassingly” parallel (i.e. the method can be implemented in a parallel architecture without the need for inter-processor communication and near perfect scaling) and do not require generating the adjoint model code. The latter is especially relevant for ecological models, many of which include complex nonlinear (and potentially non-differentiable) interactions.

How does one go about testing such models when the amount of available data is limited and the number of degrees of freedom in a spatially and temporally explicit model is large? This question motivates three aspects of the present study. Firstly, a Bayesian approach is taken in the data assimilation to make the most use of *a priori* information about the system. Secondly, the skill of this data assimilative model is assessed using a cross-validation procedure (Efron and Tibshirani, 1993; Friedrichs et al., 2007; Friedrichs et al., 2006). Lastly, the posterior distribution of the inferred parameters is evaluated using independent (unassimilated) observations. For example, laboratory and field estimates of grazing can be used to assess the plausibility of the inferred *A. fundyense* mortality.

Herein we seek to address the following specific questions in the context of the aforementioned *A. fundyense* model and the data from a regional survey:

- What is the impact of the presumed circulation on the estimate of biological parameters in a coupled model?
- Can skill metrics for biological fields distinguish between the circulation choices?
- Are some parameters (mortality) more sensitive to the choice of velocity field than others (initial conditions)?

The remainder of this paper is organized as follows: in Section 2, a dynamic model for *A. fundyense* is introduced and a stochastic model for its initial conditions is developed. In Section 3 experiments in which the initial conditions of the *A. fundyense* model are estimated are presented. The skill of the data assimilative model is assessed through cross-validation experiments. In Section 4 we present experiments in which both the mortality field and the initial conditions of the *A. fundyense* model are estimated. Again cross-validation experiments are carried out. The conclusions are reported in Section 5.

2. Dynamic and stochastic models for temporal evolution of *A. fundyense*

2.1. The coupled physical–biological model

The temporal evolution of the *A. fundyense* field is assumed to be governed by a two dimensional advection–diffusion reaction equation

$$\frac{\partial c}{\partial t} + v\cdot\nabla c - \nabla\cdot(D\nabla c) = (g - m)c \qquad (1)$$

where $c = c(x, y, t)$ is the concentration averaged over a presumed surface mixed layer of depth of 20 m. Notation used throughout is summarized in Table 1. The diffusion coefficient D is chosen to give a uniform Peclet number, $\mathrm{Pe} = \frac{|v|l}{D}$, where l is the local length scale of the finite element mesh. In this application we set Pe = 10 to ensure numerical stability of the advection–diffusion equation, yet avoiding excessive diffusivity. By definition, D is dependent on the flow v and the pair (v, D) defines the full transport effect on the biological field. Eq. (1) is solved with a Galerkin finite element method with implicit time stepping (Lynch et al., 1996). Natural boundary conditions are assumed.

The population dynamics of *A. fundyense* is described in McGillicuddy et al. (2005) and Stock et al. (2005). The vegetative cells emerge through excystment from cyst beds (Anderson et al., 2005), the abundance and distribution of which were derived from measurements of the sediments. The spatially variable growth rate, g, is a function of local temperature, salinity, light, and nutrients. The best prior estimate of mortality (0.1 d^{-1}) is imposed in a spatially uniform manner; perturbations to that prior, specifying our uncertainty in its value, are discussed further below.

Table 1
Notation used herein

Symbol	Definition
c	*A. fundyense* concentration
g	*A. fundyense* growth rate
m	*A. fundyense* mortality rate
c_i, m_i	i^{th} ensemble *A. fundyense* concentration and mortality
H	Linear measurement operator for *A. fundyense* concentration
$\mathbf{W}$	Observational error covariance
P	Model error covariance for *A. fundyense* concentration
P_{cm}	Joint error covariance for mortality rate and *A. fundyense* concentration
K_0	Kalman gain matrix for *A. fundyense* initial conditions
K_m	Kalman gain matrix for *A. fundyense* mortality rate
d	Data $d = H c^{\mathrm{true}} + \xi$ where ξ is the measurement error
$E[\cdot]$	Ensemble expectation operator, $E[x] = \frac{1}{n}\sum_{i=1}^{n} x_i$
c^a	Estimate of *A. fundyense* based on data, $c^a = E[c\|d]$
m^a	Estimate of mortality based on data, $m^a = E[m\|d]$
t_0	Start time of the simulation
c^b	Estimate based on posterior mortality and initial conditions, $c^b = E[c\|c^a(t_0), m^a]$

We consider three different representations of the velocity field ν (and by implication the corresponding diffusion field D). The first field assumes no motion,$\nu_0 \equiv 0$. The second velocity field, ν_c, is extracted from a climatological depth-averaged velocity simulation of the Gulf of Maine for the May–June period (Aretxabaleta et al., submitted for publication; Lynch et al., 1996). This climatological simulation was initialized using mean temperature, salinity, and density fields for the May–June time period and forced by baroclinic pressure gradients, bimonthly mean wind stress, and the semidiurnal M_2 tide. The third velocity field, ν_h, is extracted from a data-assimilative hindcast simulation of the same area during the cruise period (Aretxabaleta et al., in preparation). This hindcast simulation included density initialization derived from objective analysis of hydrographic data, observed wind stress during the cruise period, and elevation boundary conditions inferred from the assimilation of shipboard and moored velocity observations in the interior (Lynch and Naimie, 2002; Lynch et al., 1998). In the hindcast, the root mean square misfit of the predicted velocity is reduced from 14 cm s^{-1} in the prior to 11 cm s^{-1} after assimilation. The hindcast also improves the prediction of unassimilated drifter trajectories. There is a logical preference for the hindcast velocity field over the climatology, and for the climatology over no motion. However, we choose not to formalize this preference here; instead we treat each of these three cases in independent sensitivity analyses.

2.2. Bayesian parameter estimation

Consider the generic data assimilation problem. There is a set of observations, d, that will be used to estimate the uncertain parameters of a dynamical model. Let θ denote the unknown model parameters, $f(\theta)$ the prior distribution for the parameters, and Ψ_θ the dynamical model solution given parameter choice θ. The data d is an imperfect observation of the true state of the system, $d = H\Psi_{\text{true}} + \xi$ where ξ is the observational error and H is the measurement operator for the observations. Bayes theorem allows us to compute the posterior likelihood over θ

$$f(\theta|d) = \frac{f(d|\theta)f(\theta)}{\int_\theta f(d|\theta)f(\theta)d\theta} = \frac{f(d|\theta)f(\theta)}{B(d)} = \frac{f(d|\psi_\theta)f(\theta)}{B(d)} \tag{2}$$

where $B(d)$ is the overall likelihood of the model given the data. Assuming $f(\theta)$ is Gaussian, let $\bar{\theta}$ and $C_{\theta\theta}$ denote the mean and covariance of θ. If the observations are not biased and ξ has a Gaussian distribution with covariance W,

$$f(\theta|d) = \frac{\exp\left(-\frac{1}{2}(H\psi_\theta - d)^T W^{-1}(H\psi_\theta - d)\right)\exp\left(-\frac{1}{2}\left(\theta - \bar{\theta}\right)^T C_{\theta\theta}^{-1}\left(\theta - \bar{\theta}\right)\right)}{\sqrt{(2\pi)^{n_m n_d}|W|^{n_d}|C_{\theta\theta}|^{n_m}}} \tag{3}$$

where $n_m$is the dimension of θ, and n_d the dimension of d. We can define a cost function proportional to the log of the conditional likelihood function $J(\theta) = -2\log(f(\theta|d))$,

$$J(\theta) = (H\psi_\theta - d)^T W^{-1}(H\psi_\theta - d) + \left(\theta - \bar{\theta}\right)^T C_{\theta\theta}^{-1}\left(\theta - \bar{\theta}\right) + q \tag{4}$$

where $q = \log\left((2\pi)^{n_m n_d}|W|^{n_d}|C_{\theta\theta}|^{n_m}\right)$ is a constant. This cost function is typical of strong constraint data assimilation problems usually solved with variational methods. The first quadratic form penalizes the misfit to data, and the second enforces a regularity constraint, penalizing departures from the best prior estimates of the parameters. The inverse covariance matrices, W^{-1} and $C_{\theta\theta}^{-1}$, play the role of weighting matrices. The value of θ that minimizes Eq. (4) will also be the maximum likelihood estimate.

We introduce a phantom linear operator $F\theta \cong H\Psi_\theta$ by assuming that the state of the system (as reflected in the data) will respond linearly over the range of plausible variations in the parameters. Rewriting Eq. (4) with F we have,

$$J(\theta) = (F\theta - d)^T W^{-1}(F\theta - d) + \left(\theta - \bar{\theta}\right)^T C_{\theta\theta}^{-1}\left(\theta - \bar{\theta}\right) + q \tag{5}$$

Minimizing this quadratic form over θ leads to the normal equations,

$$0 = \frac{\partial J(\theta)}{\partial\theta} = 2F^T W^{-1}(F\theta - d) + 2C_{\theta\theta}^{-1}\left(\theta - \bar{\theta}\right) \tag{6}$$

which can be solved for θ,

$$\theta = \bar{\theta} + C_{\theta\theta}F^T\left(FC_{\theta\theta}F^T + W\right)^{-1}\left(d - F\bar{\theta}\right) \tag{7}$$

yielding the familiar Kalman analysis. In our application to the *A. fundyense* model, the linear operator F is never constructed explicitly but rather implicitly estimated from a Monte Carlo procedure. The term $FC_{\theta\theta}F^T$ is the model error covariance sampled at the observation points. The matrix $\mathbf{C}_{\theta\theta}F^T$ is the model error covariance between the uncertain parameters, θ, and the observation points. Thus the inverse problem can be solved in this manner without the construction of tangent linear model or adjoint.

2.3. Prior model for A. fundyense initial conditions

The starting point for any Bayesian estimation is the specification of a prior model (distribution) for the unknown parameters, in this case the initial conditions of *A. fundyense* concentration and in Section 4 the mortality rate as well. We assume the mean *A. fundyense* field $c_{\text{clim}}(x,y,t)$ is the climatological simulation computed in McGillicuddy et al. (2005), which runs through the period March to August. This mean seasonal simulation does not contain any of the interannual variability or small-scale patchiness inherent in the survey observations to be assimilated, which sample *A. fundyense* concentration in a non-synoptic manner over a period of 11 days (Fig. 1).

We seek an ensemble of initial conditions for the start of the survey t_0 (June 6, 2006). Perturbations about the mean state at time t_0 are computed assuming that both the mean and variability are related to the climatological value. We construct n random initial conditions, and the i^{th} random initial condition for the *A. fundyense* model is,

$$c_i(t_0) = \max\left(0, \frac{c_c \lim^{(t_0)}}{\sqrt{e}}\right)\exp\left(\xi_i^a\right) + 100\frac{\text{cells}}{\text{liter}}\xi_i^b \tag{8}$$

where $c_{\text{clim}}(x,y,t_0)$is the climatological *A. fundyense* estimate at the beginning of the survey and e is base of the natural log (the term $\frac{1}{\sqrt{e}}$ is a normalization constant). The background uncertainty of 100 cells l^{-1} reflects a subjective choice of what constitutes a “significant” concentration of the organism,

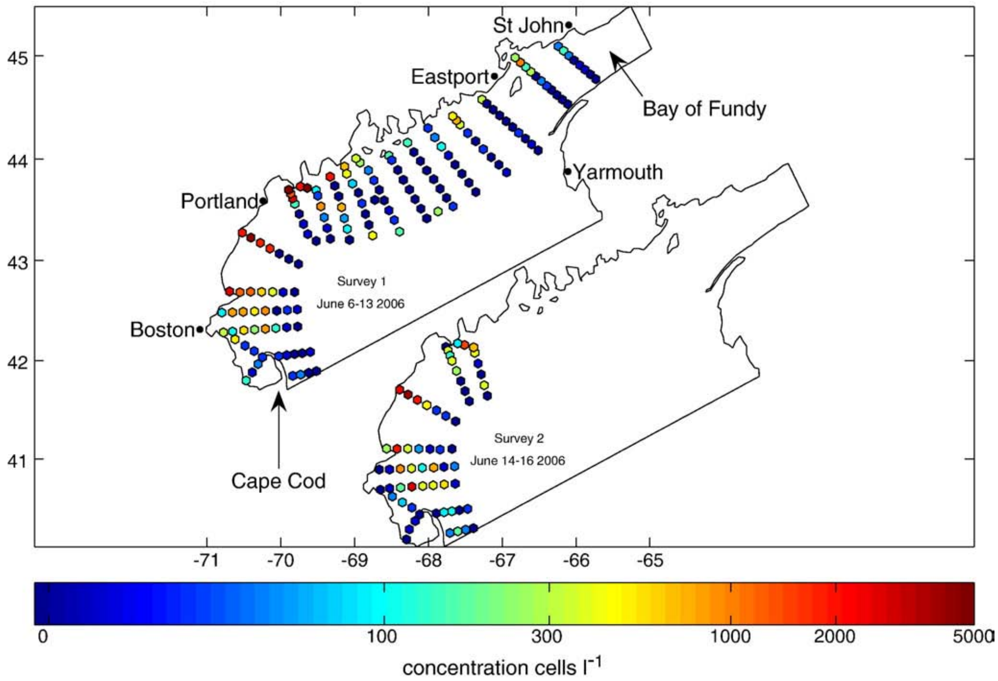

Fig. 1. Spatial map of data and boundary of computational domain. The first survey is depicted on the left and the second survey on the right. The *A. fundyense* samples were taken from June 1 2006 to June 15 2006. The measurements followed the coast from Cape Cod to the Bay of Fundy totaling 214 casts. The axes labels corresponding to the figure on the left are longitude (degrees W) along the abscissa and latitude (degrees N) along the ordinate.

chosen in this case to be one half of the concentration generally required to lead to toxicity in shellfish along the coast (Keafer et al., 2005). ξ_i^a and ξ_i^b are independent Gaussian random variables with zero mean, unit standard deviation and anisotropic covariance

$$\operatorname{cov}(x_1, x_2) = \exp\left(-\frac{|x_1 - x_2|^2}{\lambda_x^2} - \frac{|h_1 - h_2|^2}{\lambda_h^2}\right) \tag{9}$$

where $h_1$and h_2 are the bathymetric depth at x_1 and x_2 respectively. The bathymetric decorrelation scale, λ_h, forces shorter length scales in the cross isobath direction while allowing correlation along isobaths to stretch to λ_x. Stock et al. (2005) found along shore misfit decorrelation length scales of 20–25 km and cross shore decorrelation scales of 15 km. Here we choose λ_x=50 km and the bathymetric decorrelation scale λ_h=50 km, as the form of the anisotropy is slightly different than was used in Stock et al. (2005).

The resulting distribution has high variance where climatological values are high and a standard deviation of approximately 100 cells l^{-1} where the climatological values are low (Fig. 2, right panel). Note that this form induces a positive bias relative to the prior at low concentrations due to the positive definite requirement on concentration and the perturbations to the *A. fundyense* field specified in Eq. (8). It should also be pointed out the distribution implicit in Eq. (8) is neither normal

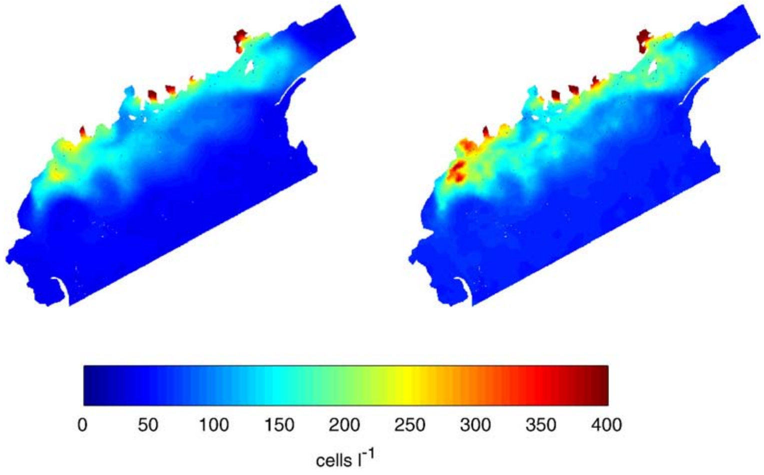

Fig. 2. Mean (left panel) and standard deviation (right panel) of initial conditions for *A. fundyense*. Units for the color scale are cells l^{-1}.

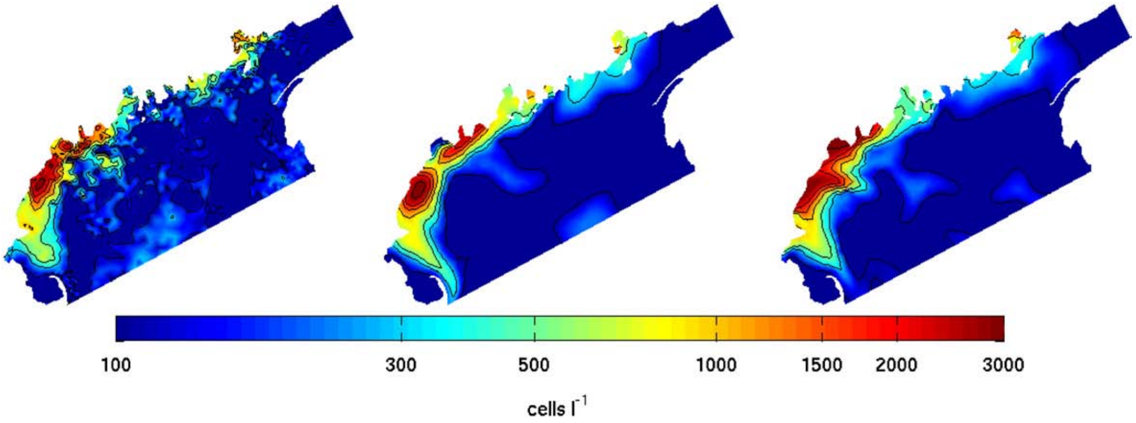

Fig. 3. Estimates of *A. fundyense* concentration on June 10 (2/3 of the way through the survey) for three different estimates of *A. fundyense* initial conditions on June 1. Left: experiment A_0 (no flow); middle: experiment A_c (climatological velocity); right: experiment A_h (hindcast velocity).

nor lognormal but close to the sum of a normal and lognormal distribution. A Monte Carlo approximation to the distribution of the prior model error at later times is obtained by solving Eq. (1) with the initial conditions given by Eq. (8).

2.4. Statistical model for A. fundyense measurements

In addition to a distribution for model errors, the other prerequisite for Bayesian estimation is a statistical model for measurement error. The *A. fundyense* data utilized here consist of shipboard microscope counts from water sampled between 1 m depth and the surface. Ten liters of water were sieved with a 20 µm filter and 1/140th of the filtered material was counted under the microscope. Other morphologically similar species, such as *A. ostenfeldii*, may be confused with *A. fundyense* cells and introduce positive bias in the data. For simplicity, we assume the measurements are unbiased with uncorrelated Gaussian standard errors of 100 cells l^{-1}, although other forms are possible.

3. Estimation of initial conditions

The EnS is a simple Monte Carlo procedure to estimate the parameters of a model. The ensemble of initial conditions $c_i(x,y,t_0)$ are simulated from the prior distribution as described above. The temporal evolution of the *A. fundyense* field is determined by solving Eq. (1), for each of the initial conditions, producing the ensemble forecast c_i. Eq. (1) is solved on the time interval $[t_0, t_f]$ where t_f is the time of the last *A. fundyense* observation. The ensemble covariances between the initial conditions and misfits to the measurements are used to estimate the initial conditions, using the standard ensemble Kalman gain update (Evensen, 2006). An ensemble of size $n=500$ is used throughout.

The analysis requires estimating the following matrices from the ensemble integrations,

$$R_0PH^T = \frac{n}{n-1}E[(c(t_0)-E[c(t_0)])(Hc-E[Hc])] \tag{10}$$

and

$$\mathbf{HPH}^T = \frac{n}{n-1}E[(Hc-E[Hc])(Hc-E[Hc])] \tag{11}$$

where $c(t_0)$ is the initial *A. fundyense* concentration, and c is the *A. fundyense* concentration at all times. R_0 is the linear operator corresponding to initial conditions, $c(t_0)=R_0c$, and thus R_0PH^T is the ensemble covariance between the initial conditions and the observations. The matrix HPH^T is the ensemble covariance between the observation points. The Kalman gain matrix for the initial conditions is,

$$K_0 = R_0PH^T\left(\mathbf{HPH}^T + \mathbf{W}\right)^{-1} \tag{12}$$

where W is the measurement error covariance matrix. The posterior ensemble of initial conditions are then estimated as

$$c_i^a(t_0) = c_i(t_0) + K_0(d-Hc_i-n_i) \tag{13}$$

where $\xi_i \sim G(0,\text{W})$, a Gaussian random variable with mean 0 and covariance W.

A Monte Carlo approximation to the posterior distribution of the *A. fundyense* population is obtained by solving Eq. (1) with initial conditions given by Eq. (13). The prediction is taken to be the ensemble mean of the posterior ensemble c^a henceforth.

3.1. Results

We present three experiments in which the initial conditions for *A. fundyense* are estimated, each of which utilizes a different circulation estimate. Velocity fields for experiments A_0, A_c, and A_h are no flow, climatology, and hindcast, respectively. As expected, the spatial structure of the estimated *A. fundyense* fields reveals dependence on the velocity field (Fig. 3). In particular A_0 is much less smooth than A_c and A_h due to the absence of diffusion (via the Peclet relationship). The RMS of the misfit of the posterior estimate, $r=d-Hc^a$, reveals relatively little dependence on the velocity fields (Table 2). The lowest RMS misfit is attained with the assumption of no motion, due to the less stringent constraint on the spatial structure of the estimate. In fact, a perfect fit to the observations could be attained in the no-flow case if the decorrelation length scales of the initial condition covariance function (Eq. (9)) were sufficiently small, the variance sufficiently large, and all measurements were distinct in space/time.

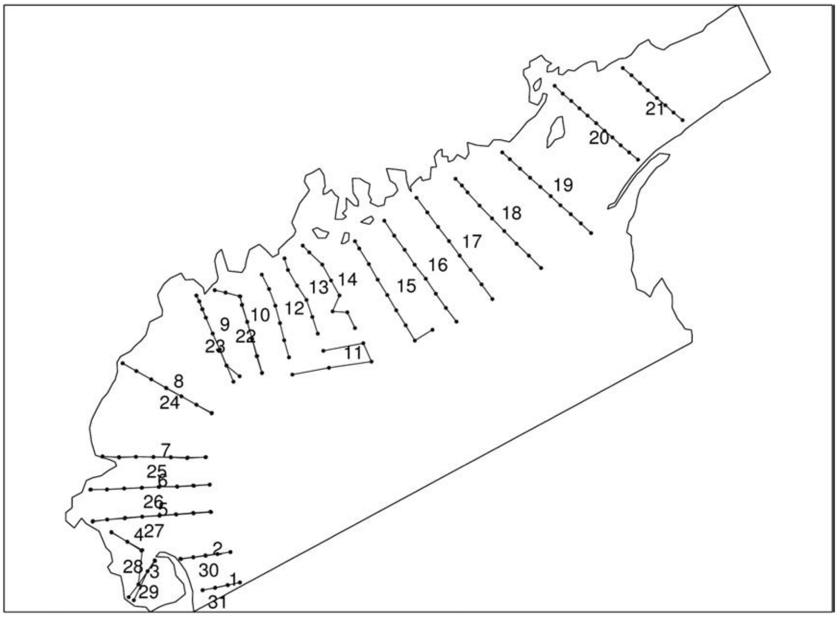

Fig. 4. Transects of data used in the cross-validation experiments. Transects 22–31 lie on top of transects 9–1 and are sampled approximately seven days apart.

3.2. Cross-validation

We seek a means to evaluate the skill of the data assimilative model with respect to data not used in the estimation. A set of thirty-one experiments was carried out, in which each one of the survey transects (Fig. 4) was systematically omitted from the EnS. The union of the misfits at the unused data points from each of the thirty-one experiments is used to assess the skill of the estimation.

Formally the procedure is a K-fold cross-validation scheme (Efron and Tibshirani, 1993). Let t^i denote the data from the missing transect (test set), s^i the remaining data (training set) and T^i and S^i the corresponding measurement operators. For each of the K=31 transects and velocity fields $\nu\in[\nu_0, \nu_c, \nu_h]$, we compute the ensemble smoother estimate of *A. fundyense* $c^a(v, s^i)$. The full cross-validation prediction vector is

$$t(v) = \left[T^1 E\left[c^a\left(v, s^1\right), T^2 E\left[c^a\left(v, s^2\right)\right], \ldots, T^K E\left[c^a\left(v, s^K\right)\right]\right] \qquad (14)$$

and the cross-validation residual is $r_{cv}(\nu)= d-t(\nu)$. We define $\mathrm{RMS}_{\mathrm{CV}}(v_h) = \sqrt{\sum_{i=1}^{n_d}(d_i-t_i(v_h))^2}$ as the RMS of the misfit from the cross-validation predictions. In the same manner we define $\mathrm{RMS}_{\mathrm{CV}}(\nu_c)$ and $\mathrm{RMS}_{\mathrm{CV}}(\nu_0)$ corresponding to estimates based on the climatological and no-flow velocity fields.

Results of the cross-validation (Tables 2 and 3) reveal that the skill of the no-motion case is barely better than that of the prior estimate. Use of the climatological and hindcast velocity fields results in better skill than with no-flow. However, the

Table 2
Table of misfit to data from the experiment estimating *A. fundyense* initial conditions

Experiment	Velocity field	Estimated parameters	Prior RMS misfit	Posterior RMS misfit	RMSCV
Data mean		$c = \frac{1}{n}\sum_{i=1}^{n} d_i$	861	791	798
A_0	None	$c(t_0)$	811	463	756
A_c	Climatological	$c(t_0)$	816	484	627
A_h	2006 hindcast	$c(t_0)$	807	501	662
B_0	None	$c(t_0),m$	809	418	831
B_c	Climatological	$c(t_0),m$	815	447	629
B_h	2006 hindcast	$c(t_0),m$	811	463	593

All entries are RMS misfits in units of cells l−1. The RMS of the data is 861 cells l^{-1}. $\mathrm{RMS}_{\mathrm{CV}}$ is the rms of the cross-validated residual. The differences between the prior RMS misfits when the same velocity field is used are due to the Monte Carlo nature of the calculation and non-linearity of the mortality affect. The row titled data mean uses the data mean as the predictor, rather than a dynamic model. The cross-validation prediction for the data mean is the mean of the training set and the prior is assumed to be zero.

Table 3
Table of significance level at which order of skill can be reversed

Experiment	A_0	A_c	A_h	B_0	B_c	B_h
Data mean	41	13	20	43	14	11
A_0		21	29	35	22	17
A_c			40	12	50	40
A_h				35	41	32
B_0					12	10
B_c						40

Based on a two sample *t*-test on the squared cross-validated misfits,$(t(v)-d)^2$, whose mean is the square of the skill estimate employed here. The test gives the probability that the true mean of the two sets of squared misfits have opposite rank of the sample mean rank, under the assumption that the squared misfits are independent and have a Gaussian distribution. The significance levels are not lower because of the variability of the squared residuals and relatively small sample size.

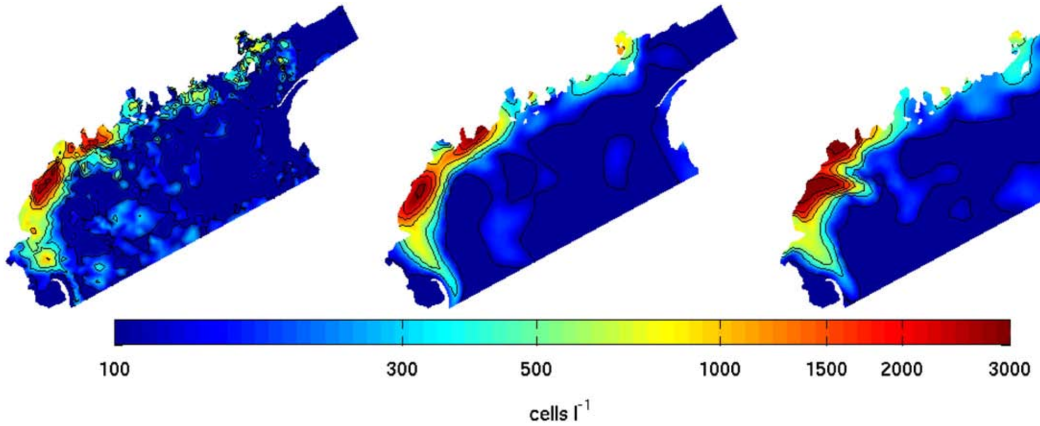

Fig. 5. Estimates of *A. fundyense* concentration on June 10 (2/3 of the way through the survey) for joint estimation of *A. fundyense* initial conditions and mortality. Left: experiment B_0 (no flow); middle: experiment B_c (climatological velocity); right: experiment B_h (hindcast velocity).

difference in skill between A_c and A_h is less significant than the differences between A_0 and either A_c or A_h.

4. Estimation of initial conditions and mortality

In this second set of computational experiments, we augment the stochastic model for *A. fundyense* to include a spatially variable mortality field. The mortality field is estimated from the *A. fundyense* measurements in the same manner as the initial conditions. With the uncertain mortality, the estimation problem becomes strongly nonlinear due the term m and c both being unknown and their product appearing on the right hand side of Eq. (1). With so little known about the spatial structure of the mortality field, we simply introduce a distribution of perturbations similar to that used for the *A. fundyense* initial conditions. We simulate the ith ensemble mortality field:

$$m_i = .1 + .025\xi_i^c \tag{15}$$

where ξ_i^c is a Gaussian random variable with mean zero and covariance given by Eq. (9). The resulting prior distribution for mortality is Gaussian with a spatially uniform mean of 0.1 d^{-1}, standard deviation of 0.025 d^{-1}, and spatial covariance described above.

We experimented with simultaneous estimation of both initial conditions and mortality, but found that approach does not work well. Because overestimation is correlated with both a high initial population and low mortality (which are assumed to be independent here), the direct EnS procedure, applied to the joint estimate of *A. fundyense* initial conditions and mortality rate, tends to overshoot the data. Instead we employ a two-step procedure. Random initial conditions and mortality rate are simulated from Eqs. (8) and (15) respectively, and the *A. fundyense* initial conditions are estimated using the algorithm described above. Second, the solution to Eq. (1) is computed again with the corrected *A. fundyense* initial conditions. Misfits from this second simulation are then used to estimate the mortality field. The posterior estimate for the mortality field is:

$$m_i^a = m_i + K_m\left(d - Hc_i^a - \xi_i\right) \tag{16}$$

where

$$K_m = R_m P_{cm}[H|0]^T\left(HPH^T + \mathbf{W}\right)^{-1}. \tag{17}$$

The matrix R_m is the operator corresponding to mortality rate, $m = R_m\begin{bmatrix} c \\ m \end{bmatrix}$ and P_{cm} is the joint error covariance over mortality rate and *A. fundyense* concentration. Thus,

$$R_m P_{cm}[H|0]^T = \frac{n}{n-1}E[(m - E[m])(Hc^a - E[Hc^a])] \tag{18}$$

is the ensemble covariance between the mortality field and the *A. fundyense* field at the observation points. The posterior estimates of the initial conditions and mortality field, $c_i^a(t_0)$ and m_i^a, are then used to generate the posterior ensemble, c_i^b, via the numerical solution of Eq. (1). As in experiment A, the estimate is the mean of the posterior ensemble, c^b.

4.1. Results

Three experiments B_0, B_c, and B_h were conducted with estimation of both the initial conditions for *A. fundyense* and the mortality field, using the same velocities as those in experiments A_0, A_c, and A_h. The joint initial condition and mortality estimation experiments reveal strong dependence on the velocity field (Figs. 5 and 6). Similar to the initial condition experiments, the RMS of the misfit of the posterior estimate $r = d - Hc^b$ achieves its lowest value in the no-flow case, with misfits increasing in the climatological and hindcast velocity cases (Table 2).

Cross-validation of the three experiments with joint estimation of the initial conditions and mortality indicates the best skill for the hindcast velocity field (Table 2). The lowest RMS for the cross-validation prediction is found with the hindcast velocity. The climatological velocity gives a slightly higher RMS for the cross-validated residuals. The low RMS misfit obtained for the no-motion case produces a high RMS_{CV}, i.e. low skill. As in experiment A, more significant differences in skill are found between B_0 and both B_c and B_h, than between B_c and B_h (Table 3).

Analysis of the distribution of posterior mortality fields provides an additional test of the model. Strictly speaking, the "mortality" term in this simple model represents all processes

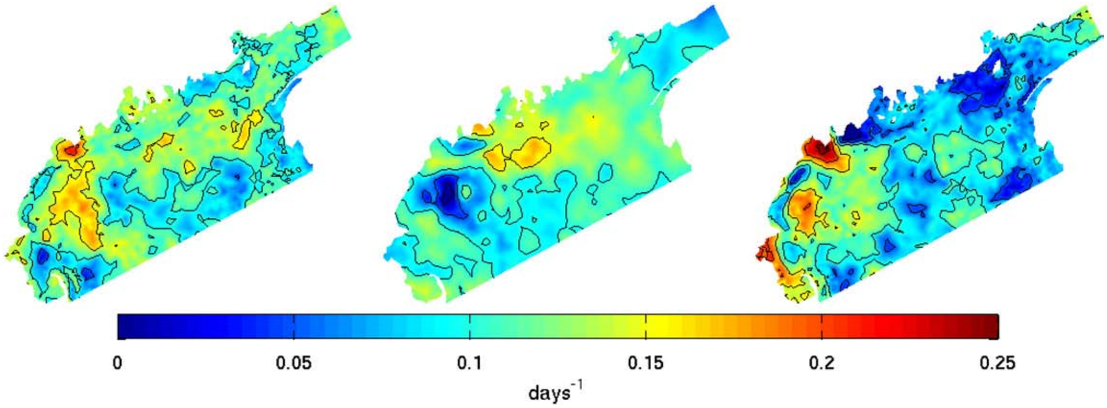

Fig. 6. Mortality estimates. Left: experiment B_0 (no flow); middle: experiment B_c (climatological velocity); right: experiment B_h (hindcast velocity). The best prior estimate of mortality is .1 d^{-1}.

that lead to removal of cells from the water column, including zooplankton grazing, encystment, and cell death. We focus on zooplankton grazing as it is the best known of the three. Studies in this region have revealed that grazing on *A. fundyense* is a complex interaction of both predator and prey, involving many different size classes of zooplankton (Campbell et al., 2005; Teegarden et al., 2001; Teegarden and Cembella, 1996; Turner and Borkman, 2005; Turner and Tester, 1997). Further complicating matters, the toxin produced by *A. fundyense* can influence its suitability as a food item for some predators (Colin and Dam, 2002). Direct measurements document highly variable grazing rates, ranging from near zero to over 600% of the standing stock of *A. fundyense* per day (e.g. Turner and Borkman (2005), Table 4). Because such measurements are so difficult and time-consuming to make, the existing database is not nearly sufficient to construct space-time maps of grazing impact on *A. fundyense* populations. However, it is clear that the mortality rates inferred from the EnS estimation procedure (ranging from 0 to 0.25 d^{-1}; Fig. 6) fall well within the envelope of observations. The inferred mortality fields are thus plausible, but their spatial structure is not strictly testable at this time.

5. Conclusions

We have demonstrated the utility of a Monte Carlo linear minimum variance estimation procedure in two strong constraint data assimilation problems. The first is a linear estimation problem: the estimation of initial conditions. The second is a nonlinear estimation problem: the joint estimation of initial conditions and mortality rate. Although the success of the method is not surprising given the broad success of the closely related EnKF approach, the application herein demonstrates that the EnS estimation methodology can be used to tackle strong constraint data assimilation problems for coupled physical–biological systems.

Our results demonstrate that estimates of biological parameters are highly dependent on the hydrodynamic fields, a finding noted earlier by Matear and Holloway (1995). In both sets of experiments (estimation of initial conditions only and joint estimation of initial conditions and mortality), lower misfits to assimilated data were achieved in cases with no flow. However, those solutions exhibited the lower skill as measured by cross-validation experiments in which subsets of the data were systematically withheld from the estimation. In contrast, use of climatological and hindcast velocity fields resulted in higher misfits to assimilated data, but better skill in predicting unassimilated data. These trends were most pronounced in the joint estimation of initial conditions and mortality. In that context, the no-flow condition resulted in the worst skill, whereas the best skill of all resulted from use of the hindcast velocity field.

Posterior analysis of the mortality field confirms that the inferred rates fall well within the envelope of observations. However, the extant database is insufficient to assess the accuracy of its spatial structure. The inferred mortality field is thus credible, but not presently testable. It should be noted that this approach tends to compensate for errors in the prescribed circulation with corresponding adjustments to the mortality field. The fact that the least skillful estimate of the mortality field (assuming no-flow) cannot be rejected in favor of the most skillful estimate (using the hindcast velocity) on the basis of direct grazing measurements highlights the intricate relationship between physical and biological uncertainties in testing coupled physical–biological models with observations.

Acknowledgments

This work was supported by grants NSF DMS-0417845, NSF OCE-0430724, NIEHS 1P50-ES01274201 and NOAA NA06N054780245. The authors thank Alfredo Aretxabaleta for the velocity estimates, Bruce Keafer, Kerry Norton and others for providing the live cell counts and Andrew Solow for valuable discussions. Additionally two anonymous reviewers made valuable suggestions that led to improvements in this paper. This is ECOHAB contribution #265.

References

Allen, J.I., Eknes, M., Evensen, G., 2002. An ensemble Kalman filter with a complex marine ecosystem model: hindcasting phytoplankton in the Cretan Sea. Annales Geophysicae 20, 1–13.

Anderson, D.M., 1997. Bloom dynamics of toxic *Alexandrium* species in the northeastern U.S. Limnology and Oceanography 42, 1009–1022.

Anderson, D.M., et al., 2005. *Alexandrium fundyense* cyst dynamics in the Gulf of Maine. Deep-Sea Research II 52, 2522–2542.

Annan, J.D., Hargreaves, J.C., 2004. Efficient parameter estimation for a highly chaotic system. Tellus A 56 (5), 520–526.

Aretxabaleta, A.L., McGillicuddy, D.J., Smith, K.W., Lynch, D.R., in preparation. Model Simulations of the Bay of Fundy Gyre: 2. Hindcasts for 2005–2007 reveal interannual variability in retentiveness. Journal of Geophysical Research.

Aretxabaleta, A.L., McGillicuddy, D.J., Smith, K.W., Lynch, D.R., submitted for publication. Model simulations of the Bay of Fundy Gyre: 1. Climatological results. Journal of Geophysical Research.

Campbell, R.G., Teegarden, G.J., Cembella, A.D., Durbin, E.G., 2005. Zooplankton grazing impacts on *Alexandrium* spp. in the nearshore environment of the Gulf of Maine. Deep Sea Research Part II 52 (19–21), 2817–2833.

Colin, S.P., Dam, H.G., 2002. Latitudinal differentiation in the effects of the toxic dinoflagellate *Alexandrium spp*. on the feeding and reproduction of populations of the copepod *Acartia hudsonica*. Harmful Algae 1, 113–125.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman & Hall, New York. 436 pp.

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

Evensen, G., 2006. Data Assimilation: the Ensemble Kalman Filter. Springer-Verlag, Berlin Heidelberg. 279 pp.

Friedrichs, M.A.M., et al., 2007. Assessment of skill and portability in regional marine biogeochemical models: the role of multiple planktonic groups. Journal of Geophysical Research 112. doi:10.1029/2006JC003852 (C08001).

Friedrichs, M.A.M., Hood, R.R., Wiggert, J.D., 2006. Ecosystem model complexity versus physical forcing: quantification of the relative impact with assimilated Arabian Sea data. Deep-Sea Research II 53, 576–600.

Keafer, B.A., Churchill, J.H., Anderson, D.M., 2005. Blooms of the toxic dinoflagellate, *Alexandrium fundyense* in the Casco Bay region of the western Gulf of Maine: advection from offshore source populations and interactions with the Kennebec River plume. Deep Sea Research II 52 (19–21), 2631–2655.

Li, X., McGillicuddy, D.J., Durbin, E.G., Wiebe, P.H., 2006. Biological control of the vernal population increase of *Calanus finmarchicus* on Georges Bank. Deep-Sea Research II 53 (23–24), 2632–2655.

Lynch, D.R., Ip, J.T.C., Naimie, C.E., Werner, F.E., 1996. Comprehensive coastal circulation model with application to the Gulf of Maine. Continental Shelf Research 16, 875–906.

Lynch, D.R., Naimie, C.E., 2002. Hindcasting the Georges Bank Circulation, Part II: Wind-Band inversion. Continental Shelf Research 22, 2191–2224.

Lynch, D.R., Naimie, C.E., Hannah, C.G., 1998. Hindcasting Georges Bank circulation, Part I: detiding. Continental Shelf Research 18, 607–639.

Matear, R.J., Holloway, G., 1995. Modeling the inorganic phosphorus cycle of the North Pacific using an adjoint data assimilation model to assess the role of dissolved organic phosphorus. Global Biogeochemical Cycles 9, 101–119.

McGillicuddy, D.J., Anderson, D.M., Lynch, D.R., Townsend, D.W., 2005. Mechanisms regulating the large-scale seasonal fluctuations in *Alexandrium fundyense* populations in the Gulf of Maine: results from a physical–biological model. Deep-Sea Research II 52, 2698–2714.

McGillicuddy, D.J., Bucklin, A., 2002. Intermingling of two *Pseudocalanus* species on Georges Bank. Journal of Marine Research 60, 583–604.

McGillicuddy, D.J., et al., 1998. An adjoint data assimilation approach to diagnosis of physical and biological controls on *Pseudocalanus* spp. in the Gulf of Maine—Georges Bank Region. Fisheries Oceanography 7, 205–218.

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

Natvik, L.J., Evensen, G., 2003b. Assimilation of ocean colour data into a biochemical model of the North Atlantic. Part 2: Statistical analysis. Journal of Marine Systems 40–41, 155–169.

Stock, C.A., McGillicuddy, D.J., Solow, A.R., Anderson, D.M., 2005. Evaluating hypotheses for the initiation and development of *Alexandrium fundyense* blooms in the western Gulf of Maine using a coupled physical–biological model. Deep-Sea Research II 52 (19–21), 2715–2744.

Tarantola, A., 2005. Inverse Problem Theory and Methods for Model Parameter Estimation. SIAM, Philadelphia, PA. 342 pp.

Teegarden, G.J., Campbell, R.G., Durbin, E.G., 2001. Zooplankton feeding behavior and particle selection in natural plankton assemblages containing toxic *Alexandrium* spp. Marine Ecology-Progress Series 218, 213–226.

Teegarden, G.J., Cembella, A.D., 1996. Grazing of toxic dinoflagellates, *Alexandrium* spp. by adult copepods of coastal Maine: implications for the fate of paralytic shellfish toxins in marine food webs. Journal of Experimental Marine Biology and Ecology 196, 145–176.

Turner, J.T., Borkman, D.G., 2005. Impact of zooplankton grazing on *Alexandrium* blooms in the offshore Gulf of Maine. Deep-Sea Research II 52 (19–21), 2801–2816.

Turner, J.T., Tester, P.A., 1997. Toxic marine phytoplankton, zooplankton grazers, and pelagic food webs. Limnology and Oceanography 5 (2), 1203–1214.