Data assimilation in a coupled physical-biogeochemical model of the California Current System using an incremental lognormal 4-dimensional variational approach: Part 1, Model formulation and biological data assimilation twin experiments

Hajoon Song[a,*], Christopher A. Edwards[b], Andrew M. Moore[b], Jerome Fiechter[b]

[a]*Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02137, U.S.A.*

[b]*Ocean Sciences Department, University of California, 1156 High Street, Santa Cruz, CA 96064, U.S.A.*

Abstract

A quadratic formulation for an incremental lognormal 4-dimensional variational assimilation method (incremental L4DVar) is introduced for assimilation of biogeochemical observations into a 3-dimensional ocean circulation model. L4DVar assumes that errors in the model state are lognormally rather than Gaussian distributed, and implicitly ensures that state estimates are positive definite, making this approach attractive for biogeochemical variables. The method is made practical for a realistic implementation having a large state vector through linear assumptions that render the cost function quadratic and allow application of existing minimization techniques. A simple nutrient-phytoplankton-zooplankton-detritus (NPZD) model is coupled

*Corresponding author, Tel. : +1 617 253 0098
Email address: hajsong@mit.edu (Hajoon Song)

Preprint submitted to Ocean Modelling *March 22, 2016*

to the Regional Ocean Modeling System (ROMS) and configured for the California Current System. Quadratic incremental L4DVar is evaluated in a twin model framework in which biological fields only are in error and compared to G4DVar which assumes Gaussian distributed errors. Five-day assimilation cycles are used and statistics from four years of model integration analyzed. The quadratic incremental L4DVar results in smaller root-mean-squared errors and better statistical agreement with reference states than G4DVar while maintaining a positive state vector. The additional computational cost and implementation effort are trivial compared to the G4DVar system, making quadratic incremental L4DVar a practical and beneficial option for realistic biogeochemical state estimation in the ocean.

Keywords: Data assimilation, Biogeochemical model, Positive-definite variables, quadratic incremental lognormal 4DVar

---

1. Introduction

In atmospheric and ocean sciences, data assimilation refers to the rigorous adjustment of model control variables to reduce inconsistencies between model state estimates and data from observations. The practice of state estimation has matured considerably in the last few decades owing to improvements in algorithmic methods and increases in computational resources and observational data collection. To date, the majority of oceanic data assimilation efforts have focused on physical state estimation. Indeed, several groups now routinely offer data assimilative output on global and regional scales in both hindcast and near-realtime systems (Oke et al. (2015a,b) and references therein).

Efforts to similarly constrain biogeochemical/ecosystem models to improve ocean state estimates of biological and chemical variables have begun to emerge and are summarized in recent reviews (Gregg, 2008, Edwards et al., 2015). Multiple approaches have been explored, including nudging (Armstrong et al., 1995, Moisan et al., 1996), optimal interpolation (Anderson et al., 2000, Popova et al., 2002), various forms of Kalman filter (Natvik et al., 2001, Allen et al., 2002, Hoteit et al., 2003, Natvik and Evensen, 2003, Hu et al., 2012) and variational methods (McGillicuddy et al., 1998, Schlitzer, 2000, Fennel et al., 2001, Friedrichs, 2001, Tijputra et al., 2007, Fiechter et al., 2011). Variational methods in biogeochemical applications have been popular for model parameter estimation (Gregg et al., 2009), though their use for state estimation is more common in physical applications (Stammer et al., 2002, Powell et al., 2008, Forget, 2010). In some cases, model deficiencies or inconsistencies have been identified through unsuccessful parameter estimation when the model is ultimately unable to represent observed features (Fennel et al., 2001).

Although estimating state variables and model parameters using variational methods is similar, one important difference exists for biogeochemical problems. In both cases, control variables are optimally adjusted to minimize a cost function that is often defined as a quadratic misfit between the observations and corresponding model states. The difference lies in the statistics of the control variables and their errors. In parameter estimation, it is generally assumed *a priori* that the parameters are consistent with a Gaussian distribution, although recent work suggests this is not always the case (Mattern et al., 2012, Fiechter et al., 2013). However, the probability density function (PDF) of biogeochemical state variables is not Gaussian but better represented by a lognormal distribution (e.g., see Campbell (1995) for analysis of satellite chlorophyll). In addition, biogeochemical variables are positive-definite. If a prior Gaussian distribution is assumed to estimate the state variables, it is possible that the maximum likelihood value of the posterior PDF may be negative. This means that the prior Gaussian distribution assumption can lead to a negative posterior concentrations for biogeochemical state variables after fitting the observations. In contrast, a lognormal distribution constrains the optimal posterior estimation to be always positive. Thus, it is desirable to reformulate the variational method using the assumption of a lognormal distribution for biogeochemical variables for computing posterior model state estimation.

Fletcher and Zupanski (2006a) introduce a 3-dimensional variational method based on the assumption that variables are lognormally distributed, and it is expanded to a 4-dimensional variational method (4DVar) in Fletcher (2010). Song et al. (2012) transform biological variables to log-space where their distribution is more Gaussian and apply an incremental form of this method to a one dimensional nutrient-phytoplankton-zooplankton (NPZ) model in a twin experiment. In the incremental approach, small adjustments, or increments, to the state vector (in this case, model initial conditions) are determined using a tangent linear assumption (Courtier et al., 1994). A maximum likelihood value of the posterior PDF is determined in log-space and then transformed back to the original space using the exponential function. Their results show significant improvement in ecosystem model state estimates for both observed and unobserved variables. This method implicitly preserves the positive-definite property because the exponential function maps any input to a positive value. Fletcher and Jones (2014) introduce a multiplicative incremental variational data assimilation method in which the optimization problem is expressed with geometric tangent linear model and does not go through the transformation to log-space.

Although 4DVar with the assumption of lognormally distributed variables and errors (L4DVar) is more appropriate for biogeochemical data assimilation, its practical implementation in a realistic configuration can be problematic. In conventional 4DVar that *a priori* assumes variables and errors are Gaussian distributed (G4DVar), the optimal state estimates are often obtained from the incremental formulation that seeks the optimal increment to the background state. In this case, the increment is assumed to be small compared to the prior (or background) and its evolution reasonably approximated by linearized model dynamics about a nonlinear model trajectory. This incremental approach reduces the optimization problem to finding the minimum of a quadratic cost function and is formally equivalent to a truncated Gauss-Newton approach (Lawless et al., 2005). However, in the incremental formulation of L4DVar, the cost function remains non-quadratic under the incremental assumption because of the logarithmic conversion of variables. The multiplicative incremental cost function in Fletcher and Jones (2014) is also non-quadratic. Consequently, the minimization algorithm requires several times more computation than incremental G4DVar.

In this study, we formulate an incremental L4DVar in quadratic form by making a first order, linear approximation for the nonlinear terms using a Taylor expansion. The quadratic form of incremental L4DVar uses the same tangent linear model, adjoint model and minimization algorithm as incremental G4DVar, making the implementation straightforward. We evaluate its performance based on a nutrient-phytoplankton-zooplankton-detritus (NPZD) model coupled to an ocean circulation model, the Regional Ocean Modeling System (ROMS), in a twin experiment framework configured for the California Current System (CCS). Results of quadratic form of incremental L4DVar from the twin experiment is compared with that of G4DVar and the discussion about the properties of quadratic incremental L4DVar follows.

2. Incremental 4DVAR

2.1. Gaussian 4DVar

One fundamental assumption in variational methods, though not always rigorously correct (Wunsch and Heimbach, 2007), is that the distributions of observational errors and control variables are close to Gaussian. Bayes' theorem can be used to derive the cost function for variables having a Gaussian distribution (Lorenc, 1986).

$$\begin{aligned} J_G(\mathbf{x}_0) &= \frac{1}{2}(\mathbf{x}_0 - \mathbf{x}_{b,0})^T \mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}_{b,0}) \\ &+ \frac{1}{2}\sum_{i=1}^{N_o}(\mathbf{y}_i - \mathbf{x}_i^o)^T \mathbf{R}_i^{-1}(\mathbf{y}_i - \mathbf{x}_i^o), \end{aligned} \tag{1}$$

where $\mathbf{x}_0 = [x_1,\ x_2,\ \ldots,\ x_n]_0^T$ is a state vector at the initial time, $\mathbf{x}_{b,0}$ represents the background initial condition, $\mathbf{y}_i = [y_1,\ y_2,\ \ldots,\ y_{m_i}]_i^T$ is the i^{th} observation set out of a total number of N_o, and $\mathbf{x}_i^o = [x_1^o,\ x_2^o,\ \ldots,\ x_{m_i}^o]_i^T$ represents the model state evaluated at the observation points. Matrices, $\mathbf{B}$ and $\mathbf{R}_i$, represent background and observational error covariance matrices, respectively. In general, the control variables may include surface and lateral boundary conditions and model errors, but in the case considered the control vector comprises only the model initial conditions. The vector, $\mathbf{x}_i^o$, can be expressed in terms of the nonlinear model $\mathcal{M}_{i,0}$ that integrates the initial condition to $t = t_i$, and the observation operator $\mathcal{H}_i$ that maps integrated model solutions from the model space to the observation locations. Thus $\mathbf{x}_i^o = \mathcal{H}_i(\mathcal{M}_{i,0}(\mathbf{x}_0))$, and we seek the solution $\mathbf{x}_{a,0}$ that minimizes (1).

The cost function J_G can be rewritten in the incremental form (Courtier et al., 1994),

$$\begin{aligned} J_G(\delta\mathbf{x}_0) &= \frac{1}{2}\delta\mathbf{x}_0^T\mathbf{B}^{-1}\delta\mathbf{x}_0 \\ &+\frac{1}{2}\sum_{i=1}^{N_o}(\mathbf{d}_i - \mathbf{H}_i\mathbf{M}_{i,0}\delta\mathbf{x}_0)^T\mathbf{R}_i^{-1}(\mathbf{d}_i - \mathbf{H}_i\mathbf{M}_{i,0}\delta\mathbf{x}_0), \end{aligned} \quad (2)$$

where $\mathbf{d}_i = \mathbf{y}_i - \mathcal{H}_i(\mathcal{M}_{i,0}(\mathbf{x}_{b,0}))$, and matrices, $\mathbf{H}_i$ and $\mathbf{M}_{i,0}$, are tangent linear representations of $\mathcal{H}_i$ and $\mathcal{M}_{i,0}$, respectively. The cost function J_G is now quadratic in $\delta\mathbf{x}_0$, and the computation for $\delta\mathbf{x}_0$ reduces to the linear problem, $\mathbf{A}\delta\mathbf{x}_0 = \mathbf{h}$, where $\mathbf{A} = \mathbf{B}^{-1} + \sum_{i=1}^{N_o}\mathbf{M}_{i,0}^T\mathbf{H}_i^T\mathbf{R}_i^{-1}\mathbf{H}_i\mathbf{M}_{i,0}$ is the Hessian matrix of J_G in (2) and $\mathbf{h} = \sum_{i=1}^{N_o}\mathbf{M}_{i,0}^T\mathbf{H}_i^T\mathbf{R}_i^{-1}\mathbf{d}_i$. In realistic atmospheric and oceanic problems, the size of $\mathbf{A}$ often exceeds $10^8 \sim 10^9$, which makes computation of the inverse of $\mathbf{A}$ difficult or impossible. However, the direct inverse computation can be avoided using an iterative, optimization procedure. A conjugate gradient descent algorithm is one optimization algorithm appropriate for quadratic cost functions.

In ROMS 4DVar the Lanczos formulation of the conjugate gradient algorithm is used whereby the inverse of the Hessian matrix is estimated using a sequence of orthonormal Lanczos vectors to factorize $\mathbf{A}$ (Fisher and Courtier, 1995, Tshimanga et al., 2008, Moore et al., 2011b). The Lanczos recurrence relation is

$$\mathbf{A}\mathbf{q}_k = \gamma_k \mathbf{q}_{k+1} + \delta_k \mathbf{q}_k + \gamma_{k-1}\mathbf{q}_{k-1}, \tag{3}$$

where $\mathbf{q}_k$ is the k^{th} Lanczos vector. The orthonormality of Lanczos vectors allows us to write the following expressions for γ_k and δ_k: $\delta_k = \mathbf{q}_k^T \mathbf{A}\mathbf{q}_k$ and $\gamma_k^2 = \mathbf{a}_k^T \mathbf{a}_k$, where $\mathbf{a}_k = \mathbf{A}\mathbf{q}_k - \delta_k \mathbf{q}_k - \gamma_{k-1}\mathbf{q}_{k-1}$. According to Equation (3), a new Lanczos vector $\mathbf{q}_{k+1}$ can be computed using the two Lanczos vectors $\mathbf{q}_k$ and $\mathbf{q}_{k-1}$, and $\mathbf{A}\mathbf{q}_k$, where $\mathbf{A}\mathbf{q}_k$ can be computed by

$$\mathbf{A}\mathbf{q}_k = \left.\frac{\partial J_G}{\partial \mathbf{x}_0}\right|_{\mathbf{q}_k} - \left.\frac{\partial J_G}{\partial \mathbf{x}_0}\right|_{0}. \tag{4}$$

Thus it is unnecessary to handle the explicit form of Hessian. Instead, only a vector $\mathbf{A}\mathbf{q}_k$ of size of $(n \times 1)$ is required, and it is easily computed using the gradient of the cost function at the k^{th} and at the first iteration. After all iterations, an orthonormal matrix $\mathbf{V}_m = [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_m]$ can be constructed, and the inverse of the Hessian matrix $\tilde{\mathbf{A}}_m^{-1}$, estimated with m Lanczos vectors, is

$$\tilde{\mathbf{A}}_m^{-1} = \mathbf{V}_m \mathbf{T}_m^{-1} \mathbf{V}_m^T, \tag{5}$$

where a symmetric tridiagonal matrix $\mathbf{T}_m$ is

$$\begin{bmatrix} \delta_1 & \gamma_1 & 0 & \cdots & 0 & 0 \\ \gamma_1 & \delta_2 & \gamma_2 & \cdots & 0 & 0 \\ 0 & \gamma_2 & \delta_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \delta_{m-1} & \gamma_{m-1} \\ 0 & 0 & 0 & \cdots & \gamma_{m-1} & \delta_m \end{bmatrix}. \tag{6}$$

Then the solution of the linear problem $\mathbf{A}\delta\mathbf{x}_0 = \mathbf{h}$ is estimated as $\delta\mathbf{x}_0 = \mathbf{V}_m\mathbf{T}_m^{-1}\mathbf{V}_m^T\mathbf{h}$.

2.2. Lognormal 4DVar

As discussed in section 1, the statistics of some biogeochemical variables such as phytoplankton or zooplankton concentrations will generally be non-Gaussian, and are generally better described by a lognormal distributions, which respects the positive nature of the concentration. The maximum likelihood value (mode) in a Gaussian distribution also represents the unbiased (median) and the minimum variance (mean) value. Thus the solution that minimizes (1) represents the maximum likelihood value or the mode of the posterior PDF as well as the mean and the median. In a lognormal distribution, however, the mode is different from the median and the mean because the concentration distribution is skewed. When fitting the mode, one can derive the cost function to compute the maximum likelihood value of the posterior PDF by combining the prior and observation conditional PDFs using Bayes' theorem (Fletcher and Zupanski, 2006a, Fletcher, 2010). One can also choose to fit the mean of prior and observation conditional PDF (Fletcher, 2010).

In this study of incremental L4DVar, we consider fitting of the median. Although the median solution may not be as optimal as the modal solution, Song et al. (2012) show that median fitting is more robust than mode fitting as uncertainties grow. In biogeochemical data assimilation we often encounter high levels of error both in the models and the observations (Anderson et al., 2000, Popova et al., 2002, Hu et al., 2012). Additionally, the incremental lognormal cost function for the median solution provides a relatively easy conversion to the quadratic form that is of interest here.

If $\ln \mathbf{x}$ represents a state vector whose elements are the logarithm of the elements of $\mathbf{x}$, the cost function for L4DVar is

$$\begin{aligned} J_L(\mathbf{x}_0) &= \frac{1}{2}(\ln \mathbf{x}_0 - \ln \mathbf{x}_{b,0})^T \mathbf{B}_L^{-1}(\ln \mathbf{x}_0 - \ln \mathbf{x}_{b,0}) \\ &\quad + \frac{1}{2}\sum_{i=1}^{N_o} (\ln \mathbf{y}_i - \ln \mathbf{x}_i^o)^T \mathbf{R}_{L,i}^{-1} (\ln \mathbf{y}_i - \ln \mathbf{x}_i^o) , \end{aligned} \tag{7}$$

where $\mathbf{B}_L$ and $\mathbf{R}_{L,i}$ are the background and observation error covariances in the transformed space, respectively. For the incremental formulation, (7) can be rewritten with respect to $\delta\mathbf{g}_0 = \ln \mathbf{x}_0 - \ln \mathbf{x}_{b,0}$

$$J_L(\delta\mathbf{g}_0) = \frac{1}{2}\delta\mathbf{g}_0^T \mathbf{B}_L^{-1}\delta\mathbf{g}_0 + \frac{1}{2}\sum_{i=1}^{N_o} (\ln \mathbf{y}_i - \ln \mathbf{x}_i^o)^T \mathbf{R}_{L,i}^{-1} (\ln \mathbf{y}_i - \ln \mathbf{x}_i^o) . \tag{8}$$

Once the optimal $\delta\mathbf{g}_0$ is obtained, the analysis $\mathbf{x}_{a,0}$ can be written in terms of $\delta\mathbf{g}_0$ as follows:

$$\begin{aligned} \mathbf{x}_{a,0} &= \exp(\ln \mathbf{x}_{b,0} + \delta\mathbf{g}_0) \\ &= \mathbf{x}_{b,0} \circ \exp(\delta\mathbf{g}_0), \end{aligned} \tag{9}$$

where operator $\circ$ represents a Hadamard product (i.e. the element-wise multiplication, also known as the Schur product) such that

$$\mathbf{a} \circ \mathbf{b} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \circ \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} = \begin{bmatrix} a_1 b_1 \\ a_2 b_2 \\ \vdots \\ a_n b_n \end{bmatrix} . \tag{10}$$

$\mathbf{x}_i^o$ is the model states in the observation space and approximated with the tangent linear assumption

$$\begin{aligned}\mathbf{x}_i^o &\approx \mathcal{H}_i(\mathcal{M}_{i,0}(\mathbf{x}_{b,0})) + \mathbf{H}_i\mathbf{M}_{i,0}\delta\mathbf{x}_0 \\ &\equiv \mathbf{x}_{b,i}^o + \delta\mathbf{x}_i^o. \end{aligned} \tag{11}$$

It is noted that the cost function (8) is identical to the one in Fletcher and Jones (2014) (their equation (31) without the last two terms for the median solution) despite the different treatment of the problem (additive in this study versus geometric in Fletcher and Jones (2014)).

Even after the tangent linear assumption, the incremental L4DVar cost function (8) is not quadratic in $\delta\mathbf{g}$ because of the logarithm function $\ln \mathbf{x}_i^o$. Among possible minimization algorithms, one can apply Newton-Raphson method or quasi Newton method to solve this problem in an iterative manner. However, these methods either calculate or estimate the inverse of Hessian that is updated in every iteration, which makes the minimization of the cost function non-trivial. The Lanczos formulation cannot be applied to non-quadratic cost functions because (4) does not apply. Hence, it is desirable to further linearize (8) as a quadratic form so that incremental L4DVar is more affordable in realistic problems.

2.3. Quadratic L4DVar

The cost function (8) is non-quadratic with respect to $\delta\mathbf{g}_0$ after applying tangent linear assumption because of $\ln \mathbf{x}_i^o = \ln(\mathbf{x}_{b,i}^o + \delta\mathbf{x}_i^o)$. However, the natural logarithm function can be linearized using a Taylor expansion,

$$\begin{aligned}\ln\left(\mathbf{x}_{b,i}^o + \delta\mathbf{x}_i^o\right) &\approx \ln \mathbf{x}_{b,i}^o + \mathbf{L}_i\delta\mathbf{x}_i^o \\ &\approx \ln \mathbf{x}_{b,i}^o + \mathbf{L}_i\mathbf{H}_i\mathbf{M}_{i,0}\delta\mathbf{x}_0, \end{aligned} \tag{12}$$

where

$$
\begin{aligned}
\mathbf{L}_i &\equiv \left.\frac{\partial \ln \mathbf{x}_i^o}{\partial \mathbf{x}_i^o}\right|_{\mathbf{x}_i^o=\mathbf{x}_{b,i}^o} \\
&= \begin{bmatrix} (\mathbf{x}_{b,i}^o)_1 & 0 & \cdots & 0 \\ 0 & (\mathbf{x}_{b,i}^o)_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & (\mathbf{x}_{b,i}^o)_{m_i} \end{bmatrix}^{-1} \qquad (13)
\end{aligned}
$$

and $(\mathbf{x}_{b,i}^o)_j$ is the j^{th} element of the vector $\mathbf{x}_{b,i}^o$. Equation (12) can then be expanded as

$$
\begin{aligned}
\ln\left(\mathbf{x}_{b,i}^o + \delta\mathbf{x}_i^o\right) &\approx \ln \mathbf{x}_{b,i}^o + \mathbf{L}_i\mathbf{H}_i\mathbf{M}_{i,0}(\mathbf{x}_{a,0} - \mathbf{x}_{b,0}) \\
&= \ln \mathbf{x}_{b,i}^o + \mathbf{L}_i\mathbf{H}_i\mathbf{M}_{i,0}(\mathbf{x}_{b,0} \circ \exp(\delta\mathbf{g}_0) - \mathbf{x}_{b,0}), \qquad (14)
\end{aligned}
$$

and can be further linearized as

$$
\begin{aligned}
\ln\left(\mathbf{x}_{b,i}^o + \delta\mathbf{x}_i^o\right) &\approx \ln \mathbf{x}_{b,i}^o + \mathbf{L}_i\mathbf{H}_i\mathbf{M}_{i,0}(\mathbf{x}_{b,0} \circ (\mathbf{1}_n + \delta\mathbf{g}_0) - \mathbf{x}_{b,0}) \\
&= \ln \mathbf{x}_{b,i}^o + \mathbf{L}_i\mathbf{H}_i\mathbf{M}_{i,0}\mathbf{x}_{b,0} \circ \delta\mathbf{g}_0 \\
&= \ln \mathbf{x}_{b,i}^o + \mathbf{L}_i\mathbf{H}_i\mathbf{M}_{i,0}\mathbf{X}_{b,0}\delta\mathbf{g}_0, \qquad (15)
\end{aligned}
$$

where $\mathbf{X}_{b,0}$ is a diagonal matrix comprised of the elements of $\mathbf{x}_{b,0}$.

As a result, the cost function for incremental L4DVar in (8) can be written

$$
\begin{aligned}
&J_L(\delta\mathbf{g}_0) \\
&\quad = \frac{1}{2}\delta\mathbf{g}_0^T\mathbf{B}_L^{-1}\delta\mathbf{g}_0 \\
&\qquad + \frac{1}{2}\sum_{i=1}^{N_o}\left(\mathbf{p}_i - \mathbf{L}_i\mathbf{H}_i\mathbf{M}_{i,0}\mathbf{X}_{b,0}\delta\mathbf{g}_0\right)^T\mathbf{R}_{L,i}^{-1}\left(\mathbf{p}_i - \mathbf{L}_i\mathbf{H}_i\mathbf{M}_{i,0}\mathbf{X}_{b,0}\delta\mathbf{g}_0\right), (16)
\end{aligned}
$$

where $\mathbf{p}_i = \ln \mathbf{y}_i - \ln \mathbf{x}^o_{b,i}$, and (16) is now quadratic with respect to $\delta\mathbf{g}_0$. The gradient of J_L with respect to $\delta\mathbf{g}_0$ is

$$\frac{\partial J_L}{\partial \delta \mathbf{g}_0} = \mathbf{B}_L^{-1}\delta\mathbf{g}_0 - \mathbf{X}_{b,0}^T \sum_{i=1}^{N_o} \mathbf{M}_{0,i}^T \mathbf{H}_i^T \mathbf{L}_i^T \mathbf{R}_{L,i}^{-1} \left(\mathbf{p}_i - \mathbf{L}_i \mathbf{H}_i \mathbf{M}_{i,0} \mathbf{X}_{b,0} \delta\mathbf{g}_0\right) \quad (17)$$

and the Hessian is

$$\frac{\partial^2 J_L}{\partial \delta \mathbf{g}_0^2} = \mathbf{B}_L^{-1} + \mathbf{X}_{b,0}^T \left(\sum_{i=1}^{N_o} \mathbf{M}_{0,i}^T \mathbf{H}_i^T \mathbf{L}_i^T \mathbf{R}_{L,i}^{-1} \mathbf{L}_i \mathbf{H}_i \mathbf{M}_{i,0} \right) \mathbf{X}_{b,0}. \quad (18)$$

The optimal solution $\delta\mathbf{g}_0$ can be estimated using the Lanczos form of conjugate gradient algorithm as described in section 2.1. After all iterations, the solution in log-space can be easily converted to $\mathbf{x}_{a,0}$ using (9).

The quadratic cost function (16) has two additional matrices $\mathbf{X}_{b,0}$, $\mathbf{L}_i$ compared to the cost function of incremental G4DVar in (2). These two matrices, however, are trivial to handle because they are diagonal matrices and represent weighting factors for each vector element. Thus the additional computational expense resulting from these two matrices is negligible.

3. Data assimilation of surface chlorophyll data

3.1. Model

In this section, we compare the performance of incremental G4DVar and quadratic incremental L4DVar within the twin experiment framework using a NPZD model coupled to ROMS. The NPZD model has four, nonlinearly interacting components: phytoplankton (P), zooplankton (Z), nutrient (N) and detritus (D) (Powell et al., 2006, Fiechter et al., 2009). Specifically, P uptakes nutrient (N) and grows following a Michaelis-Menten formulation; it is consumed by Z with an Ivlev formulation. The mortality rate of both P

and Z are linearly proportional to their concentrations and their loss is added to D. The concentration of D decreases with the remineralization of D to N that is linearly proportional to its concentration. It also redistributes vertically by sinking with prescribed vertical sinking velocity. The parameters used in the NPZD model are listed in Table 1.

3.2. Setting

The CCS region was chosen for the twin experiment. Our domain covers the region ranging 134-115.5°W and 30-48°N with a horizontal resolution of 1/3° and 30 vertical levels. This model domain has been used in other studies for ROMS 4DVar, and it is described in detail by Broquet et al. (2009, 2011) and Moore et al. (2011a).

To prepare the initial condition for NPZD variables and the background error covariance matrix, a 45-year physical-biological coupled forward run was executed. The model was forced using fluxes derived from CORE2 (Common Ocean-Ice Reference Experiments; Large and Yeager (2009)), and open boundary condition data was taken from monthly output from the Simple Ocean Data Assimilation (SODA, version 2.1.6) data set with half degree resolution (Carton and Giese, 2008). The initial condition for N was taken from monthly climatological values (World Ocean Atlas 2001). Other variables, for which climatological data is not available, had uniform concentrations horizontally and vertically with a constant value (0.1 mmol N m^{-3}). Similar to the initial conditions, the open boundary condition for N was derived from climatology and a constant boundary value was chosen for P, Z and D.

The simulations for incremental G4DVar and quadratic incremental L4DVar started from January 1st, 2001. The initial conditions for the physical circulation were taken from a data assimilation run described by Broquet et al. (2009) (i.e., a physical data assimilation product on the same model domain within the same model framework). Surface forcing fields were derived from daily averaged atmospheric conditions produced by the Coupled Ocean Atmosphere Mesoscale Prediction System (COAMPS) (Doyle et al., 2009). Open boundary conditions for physical variables were taken from the monthly SODA data set. The initial and boundary conditions for the NPZD variables were obtained from the 45-year forward run. The coupled NPZD-ROMS model was integrated for 4 years from 2001 to 2004.

Fig. 1 compares the model simulation with the Sea-viewing Wide Field-of-view Sensor (SeaWiFS) chlorophyll data during those 4 years. The simulated P is converted to carbon using a C:N=(106 mol C):(16 mol N) Redfield ratio and then to chlorophyll using a fixed C:Chl ratio of (50 g C):(1 g Chl), although a spatially dependent C:Chl ratio may be desirable to reflect variability in this value within the diverse phytoplankton of the CCS (Goebel et al., 2010). The annually averaged chlorophyll data from the satellite shows that coastal areas north of 40°N have higher chlorophyll than other areas; it has been argued that the Strait of Juan de Fuca and Columbia River supply macro- and micronutrients, fuel primary production as well as local upwelling, possibly associated with submarine canyons (Hickey and Banas, 2008, Bruland et al., 2008, Banas et al., 2009, Davis et al., 2014). In contrast, our roughly 30 km resolution model simulation, which does not include river outflow or represent shelf/slope topography well, does not represent these high levels of chlorophyll in the northern coastal areas. The model simulation also underestimates offshore chlorophyll values compared to the satellite data. This shortcoming is presumably associated with having only one P box to represent the natural phytoplankton diversity of the CCS and using a constant C:Chl conversion ratio. The ratio used represents diatoms which dominate the coastal upwelling system, but smaller phytoplankton contribute more to offshore populations in nature. Furthermore, diatoms typically have a higher N half-saturation constant, which hinders biomass production in N-limited offshore waters where smaller phytoplankton types with lower N requirements for growth can thrive.

The latitude-time plots show seasonal variability for the coastal chlorophyll concentration averaged over the areas from the coast to about 100 km offshore in both satellite data and model simulation. Along the central California coast (34°N to 42°N), modeled chlorophyll has higher variability than the data, showing higher peak concentration during bloom periods and lower concentrations in between. At higher latitudes, modeled chlorophyll variability is also weaker than in nature, in part owing to the omission of the Strait of Juan de Fuca and Columbia River outflow.

Despite these differences between model and data, the model produces a realistic mean geographic pattern in the phytoplankton field along with a vigorous annual cycle and higher frequency variability with reasonable amplitude and spatial structure. Improvements, through alteration of model resolution, biological dynamics or further tuning of parameters, are possible, but not required for the evaluation of the the quadratic form of incremental L4DVar within a realistic configuration, which is the purpose of this paper. In our twin experiment framework, this 4-year integration is taken to represent the "true" NPZD, time-varying state (hereafter referred to as the "true" run) from which pseudo-observations are drawn.

To investigate biological data assimilation in isolation, experiments consisted of assimilation cycles in which the background state for physical variables was equivalent to the true run, but perturbations were introduced for biological variables. Forcing and lateral boundary conditions were also identical to the true run. We conducted multiple, 30-day sequences of 5-day assimilation cycles. The background initial condition for the first 5-day cycle of a sequence was created by averaging fields on that day from the 4 year output of the "true" run. For example, the background initial condition for January 1^{st} was the mean states of January 1^{st} from 2001 to 2004. We applied 10 iterations of the conjugate gradient algorithm (or 10 inner loops) to estimate the inverse of the Hessian matrix, and the final state was determined after 4 repetitions of the minimization process (or 4 outer loops) with updated background model states. After the data assimilation adjusts to the initial condition for the NPZD model, the physical-biological coupled model was integrated to generate the analysis, and further integrated for another 5 days to yield a background for the next 5-day cycle. This procedure was repeated 6 times, spanning 30 days, and then restarted at the first day of the following month by resetting the NPZD prior initial condition to the 4-year mean value for that day. Using the true physical circulation, we observed that even a forward (non-data assimilative) ecosystem model run over the course of time approached the "true" run, regardless of any initial condition consistent with climatology. Thirty day sequences were sufficiently long to investigate the benefits of sequential assimilation without loss of initial condition memory. In our analysis, we treated the first 5 days as a spinup period and considered only the last 25 days of each month.

The background error covariance was estimated according to $\mathbf{\Sigma C \Sigma}^T$, where $\mathbf{\Sigma}$ is a diagonal matrix of error standard deviations and $\mathbf{C}$ is a univariate correlation matrix. The correlation in $\mathbf{C}$ is the normalized solution of the diffusion equation (Weaver and Courtier, 2001, Bennett, 2002, Moore et al., 2011b) with horizontal and vertical length scales of 50 km and 30 m, respectively. Incremental G4DVar and quadratic incremental L4DVar share the same $\mathbf{C}$ with the assumption that the ranges of observation influence are the same in both methods. However, they have different $\mathbf{\Sigma}$. The matrix $\mathbf{\Sigma}$ was computed for each month using the 45-year forward simulation following Broquet et al. (2009) but in different spaces. The $\mathbf{\Sigma}$ in the physical space was used for incremental G4DVar and the $\mathbf{\Sigma}$ in log-space was used for quadratic incremental L4DVar. We further used preconditioning using Ritz vectors of $\mathbf{A}$ to expedite the search for the cost function minimum (Tshimanga et al., 2008, Moore et al., 2011b).

The 45-year forward simulation was forced by CORE2, while the simulation for the "true" states were forced by COAMPS. Ideally, the surface forcing for two simulations should be consistent. We choose COAMPS for our experiments because of its high resolution in the California Current region, but its record is shorter than CORE2, starting only in 1999. For the calculation of the model variability, which contributes to the background error covariance, we felt that generating statistics from a longer model run was advantageous. We acknowledge that the assimilation system could function with many other background error covariance estimates, and that the one deriving from this particular run is inevitably different from the true matrix

B. Nonetheless, it is a reasonable choice for a proof of concept experiment such as carried out here.

Pseudo-observations were sampled daily from the surface P field of the true run, then perturbed in log-space by adding random error sampled from $\mathcal{N}(0, \sigma^2)$ with $\sigma = 0.2$, which corresponds approximately 20% of multiplicative error. Thus the observation error covariance for quadratic incremental L4DVar is a diagonal matrix with $(0.2)^2$ on its diagonal. This uncertainty level is smaller than that for global chlorophyll data (±35%, Moore et al. (2009)) but optimistically chosen. In Song et al. (2016), real satellite observations are assimilated and we increase the observational errors to be more consistent with estimates of those errors. The uncertainty level for incremental G4DVar was determined after transforming the perturbations to the original space and fitting them to the Gaussian distribution. These estimated additive observational error levels are 0.2 ± 0.02 in incremental G4DVar. Thus its observational error covariance matrix is comparable to that for quadratic incremental L4DVar.

3.3. Evaluation of the linear approximation

3.3.1. Tangent linear approximation

Both incremental G4DVar and quadratic incremental L4DVar make the tangent linear approximation such that the model states can be decomposed into a background state and a perturbation. Thus a check of the time scale over which the tangent linear approximation is valid is appropriate, and we used the proportion of perturbation growth associated with the nonlinear dynamics to the total perturbation growth in the data assimilated state as a metric. The total perturbation growth is computed as $\Delta = \mathcal{M}(\mathbf{x}_{b,0} +$

$\delta\mathbf{x}_0) - \mathcal{M}(\mathbf{x}_{b,0})$ and the perturbation growth by nonlinear dynamics is $\delta = \mathcal{M}(\mathbf{x}_{b,0} + \delta\mathbf{x}_{b,0}) - \mathcal{M}(\mathbf{x}_{b,0}) - \mathbf{M}\delta\mathbf{x}_0$. If the ratio $\delta/\Delta = 0$, total perturbation growth can be explained solely by the linear dynamics.

Fig. 3 shows the ratio δ/Δ for the surface P, Z and N in time for 48 experiments corresponding to the first cycle of each 30-day sequence and using the actual perturbation determined by assimilation for $\delta\mathbf{x}_0$. Although some months show a rapid increase in the ratio such that δ/Δ exceeds a value of 1 within 5 days, the majority of ensemble members show δ/Δ is smaller than 1 for more than 5 days. The ensemble mean ratios (black lines) also remain below 1 up to 5 days. We conclude that a 5-day assimilation cycles is reasonably consistent with the linear assumptions of the tangent linear approximation for this model configuration and application.

3.3.2. Taylor series approximation for ln *and* exp *function*

The cost function for incremental G4DVar is quadratic, and as a result, the Lanczos form of conjugate gradient minimization can be applied. In incremental L4DVar, however, we need to consider further linear approximations for ln and exp functions as shown in (12) and (15) for a quadratic cost function.

The first order linear approximation in (12) is equivalent to the Taylor series approximation of ln function, $\ln\mathbf{x}_i^o \equiv \ln\left(\mathbf{x}_{b,i}^o + \delta\mathbf{x}_i^o\right)$. To be valid, the perturbation term $\delta\mathbf{x}_i^o$ should be considerably smaller than $\mathbf{x}_{b,i}^o$. Their relative sizes can be evaluated at run-time when quadratic incremental L4DVar processes the observations. In our experiments, we added a filter to remove any observations that invalidate this approximation.

For a given element in $\ln\left(\mathbf{x}_{b,i}^o + \delta\mathbf{x}_i^o\right)$, the more complete series expansion is written

$$\ln(x_b^o + \delta x^o) = \ln x_b^o + \frac{\delta x^o}{x_b^o} - \frac{1}{2}\left(\frac{\delta x^o}{x_b^o}\right)^2 + \cdots \tag{19}$$

where the error associated with the first order truncation is $O(\left(\frac{\delta x^o}{x_b^o}\right)^2)$. It is desirable for this error to be small. Typically, the updated state is located between the background state and the observation. Thus, we argue that $|\delta x^o| = |x_a^o - x_b^o| < |y - x_b^o|$ in general. It is useful then to require

$$\left(\frac{\delta x^o}{x_b^o}\right)^2 < \left(\frac{y - x_b^o}{x_b^o}\right)^2 < \alpha^2, \tag{20}$$

where α is a positive constant to be chosen. The equation $|y - x_b^o|/x_b^o < \alpha$ is equivalent to

$$(1-\alpha)x_b^o < y < (1+\alpha)x_b^o. \tag{21}$$

Since y and x_b^o are both positive-definite, α should be chosen between 0 and 1. In this experiment, we set $\alpha = 1$ and discard observations outside of the range in (21). Although this approach reduces the number of available observations, it produces a more robust analysis and one that is more consistent with the formulation. This filtering also expedites the convergence of the cost function (not shown).

Fig. 4(a,e) plots $\ln\left(\mathbf{x}_{b,i}^o + \delta\mathbf{x}_i^o\right)$ and $\ln \mathbf{x}_{b,i}^o + \mathbf{L}_i\mathbf{H}_i\mathbf{M}_{i,0}\delta\mathbf{x}_0$. If the first order approximation is valid, the slope should be closer to 1. In the first assimilation cycle, the slope is 0.98 and R^2 coefficient is 0.92, which shows that the approximation is reasonably good. The linear approximation becomes more accurate with cycles as the model states get closer to the truth. In the last cycle, the slope is 1 and R^2 coefficient is 0.98.

The second linear approximation is made when writing the $\exp(\delta \mathbf{g}_0) \approx (\mathbf{1}_n + \delta \mathbf{g}_0)$ using a Taylor expansion. In order for this approximation to be valid, $\delta \mathbf{g}_0$ should be small relative to 1. The increment, $\delta \mathbf{g}_0$, is determined by the assimilation procedure, and a consistency check is possible at that time.

Fig. 4(b,c,d) show the surface $\delta \mathbf{g}$ for P, Z and N from the first assimilation cycle, respectively. The magnitude of $\delta \mathbf{g}$ elements are generally smaller than 1 in most areas west of 126°W. However, large areas closer to the coast have elements of $\delta \mathbf{g}$ with magnitude greater than 1, leading to a less accurate linear approximation there. Fortunately, the increment amplitude generally decreases through sequential assimilation as the assimilated state approaches truth. In the last cycle, elements of $\delta \mathbf{g}_0$ have magnitude less than 1 (and mostly less than 0.3) in all areas, making the quadratic form of L4DVar closer to the non-quadratic form of L4DVar. At present, we implement no filter to handle cases where this second approximation is significantly violated, but instead rely on the fact that the correction is generally in the appropriate direction, even when the tangent linear assumption is violated, and that subsequent cycles can make further corrections in the state estimate. Indeed, the quadratic form of L4DVar converges to the "true" states without filter as to be shown in the following subsection.

3.4. Results

The performance of the quadratic form of the incremental L4DVar was first evaluated in terms of the RMSE at the surface from five simulations: a free run (no assimilation), a background (or prior) and analysis by incremental G4DVar, and a background and analysis by quadratic incremental L4DVar, respectively (Fig. 5). With a 4-year experiment, error calculations are based on 12 ensembles of 25-day assimilation runs.

Both incremental G4DVar and the quadratic form of incremental L4DVar generally improve the model's state estimation for both the observed variable P and unobserved variables Z, N and D, showing the smallest RMSE in their analysis. Among the five simulations, the smallest RMSE is that for the analysis by quadratic incremental L4DVar (red bars) in all cases. The RMSE differences in P between the two analyses are not statistically significant, showing that they are both equally effective in improving the estimation for the observed variable. For unobserved variables, however, quadratic incremental L4DVar shows statistically better performance than incremental G4DVar.

The RMSEs of the background by quadratic incremental L4DVar (orange bars) are also significantly smaller than the free run RMSE for all variables, indicating that the benefits of assimilation outlast the cycle period within which data is available. We note that the background states of the quadratic incremental L4DVar has smaller RMSEs than the analysis using incremental G4DVar. This result suggests that finding the optimal solution in log-space is more accurate and desirable because the main difference between the two methods is the log-transformation. Both methods use the same tangent linear and adjoint model (hence the same dynamics), but the fitting occurs in different spaces.

A Taylor diagram is used to compare the reference states and model estimates using three statistical properties: standard deviation, correlation coefficient and root-mean-squared (RMS) difference. Fig. 6 shows the normalized improvements by incremental G4DVar (open arrowhead) and quadratic incremental L4DVar (filled arrowhead) at the surface for four seasons. If the arrowhead is closer to the reference point, the variance of the posterior state estimate is more similar to the reference state (truth) and the two have a higher correlation.

Both methods show meaningful improvements in the observed variable P (Fig. 6, blue arrows). Quadratic incremental L4DVar performs slightly better with a higher correlation coefficient and smaller RMS difference than incremental G4DVar in all seasons. The variance of incremental G4DVar is usually closer to the reference value. Significant improvements in D are also shown from both methods in all seasons (cyan arrows). Quadratic incremental L4DVar gives slightly better statistics with smaller RMS differences and higher correlation. Although its actual RMSE reduction is the smallest ($O(10^{-3})$), the normalized statistics show the second best improvement. Improvements in Z (red arrows) are not as substantial as in P or D, but both methods improve the estimation of this variable. Consistent with the non-normalized RMSE (Fig. 5), normalized improvements for N (Fig. 6, green arrows) are smallest, with the shortest arrow lengths. Although small, adjustment by quadratic incremental L4DVar in all seasons is generally more toward the reference than for G4DVar.

The advantage of the quadratic form of incremental L4DVar is also seen in the adjusted initial fields. Fig. 7 shows the initial conditions of P and Z on a log-scale for June 6^{th} 2001, in the midst of a phytoplankton bloom (Fig. 1). Initial conditions for P from incremental G4DVar (Fig. 7(c)) and quadratic incremental L4DVar (Fig. 7(d)) visually are both closer to the true initial condition (Fig. 7(a)) than the background (Fig. 7(b)). As expected, all values from the quadratic incremental L4DVar analysis are positive through the domain. However, incremental G4DVar creates areas (shown in black) with negative concentration after fitting the observations. Furthermore, quadratic incremental L4DVar represents areas with small concentrations better than incremental G4DVar.

Improvement in Z on a log-scale (Fig. 7(e-h)) is not as clear as for P, but the reduction of RMSE is statistically significant in the original space (Fig. 5). Negative concentrations for Z result from incremental G4DVar as with P. Negative values have $O(10^{-1})$, which is not negligible. For example, the reference P state near areas at 34°N and 125°W ($\sim$ 2.5 mmol N m^{-3}) have higher concentration than the background state ($\sim$ 0.5 mmol N m^{-3}). This positive innovation can be reduced by increasing the initial P concentration or decreasing the initial Z concentration so that grazing is reduced and the concentration of P increases. In practice, both adjustments occur, consistent with the model dynamics and model error covariances. Here, both incremental G4DVar and quadratic incremental L4DVar increase the initial P concentration to roughly 1.5 mmol N m^{-3} and 2.4 mmol N m^{-3}, respectively. Incremental G4DVar reduces the initial Z concentration more than its background value resulting in a negative concentration. In contrast, quadratic incremental L4DVar analysis keeps initial Z concentrations positive even if smaller than the background value.

We note that the bias is also reduced by both approaches, although the improvement is not clear in the analysis due to a small background bias (not shown). This fact results from our choice of climatology as the background, which has a small bias when averaged over four cycles. It is possible that in a more realistic setting, there may be a considerable change in the bias improvement by the two approaches which must be considered.

Fig. 8 shows differences between initial true state and free run state at three vertical cross-sections on June 16^{th} 2001, along with the adjustments by incremental G4DVar and quadratic incremental L4DVar. Differences between the initial true state and free run state represent the changes required for the analysis to match truth, and we refer to them as desirable adjustments. We pick three cross-sections at 37°N, 40°N and 43°N, where interesting vertical features can be observed.

The desirable adjustments at 43°N are negative at the coast and this signal reaches down to −50 m depth. Both methods make negative adjustments over similar regions as in $P_{true} - P_b$. Offshore, the desirable adjustments are positive at the surface and weakly negative below −30 m. Both methods are able to make positive adjustments at the surface. However, they are not able to capture the negative subsurface misfit correctly using the surface observations. At 40°N, negative desired adjustments near the coast extend from the surface to about −75 m. Incremental G4DVar makes adjustments with a similar horizontal scale, but the depth of the negative adjustments are shallower (−30 m) than desired, with positive adjustments deeper in the water column. The quadratic form of incremental L4DVar also makes shallower (−50 m) adjustments than desired, but it does not have positive adjustments below −50 m. At 37°N, the desirable adjustments are well captured in both horizontal and vertical scales by both quadratic incremental L4DVar and incremental G4DVar at both coastal and offshore areas, though incremental G4DVar is slightly inferior near −127.5°W.

As stated earlier, both methods are based on the same dynamics by using the same tangent linear and adjoint models. Thus the differences of the adjustment come from the log-transformation. Since the observational error matrices in original space and log-space differ only within 10%, the assumption of variable's PDF and corresponding representation of the background error have a significant impact on the accuracy of state estimation.

Fig. 9 shows the STD of P at the surface as well as three vertical sections used to generate diagonal elements in the model error covariances for incremental G4DVar and quadratic incremental L4DVar. In the original space (Fig. 9a), high variations can be found near the coast with the STD greater than 3, and low variation can be found at offshore with the STD close to zero. Thus little adjustment offshore is allowed when using this STD field. When computed in log-space (Fig. 9b), the STD field shows different horizontal characteristics. The STD values are in the same order over most of the domain, with largest values in a coastal transition zone near 128°W. This STD field leads to large (logarithmic) adjustment over all areas at the surface by using the quadratic incremental L4DVar as shown in Fig. 7.

The vertical structure of STD also differs dramatically between the two spaces. Variances in the original space are close to zero below -80 m depth, while the maximum variance can be found below -50 m depth in log-space. Although it is difficult to conclude what methods result in better estimation of vertical structure with surface observations from Fig. 8, we can anticipate that incremental G4DVar is more effective at adjusting large amplitude concentrations (e.g., in coastal regions) than low amplitude signals (e.g., offshore and at depth) while quadratic incremental L4DVar should be able to adjust a range of amplitudes in both coastal and offshore regions. We note also that the substantially higher model uncertainty in log-space at depths below -50 m imply that quadratic incremental L4DVar is very sensitive in these regions, and may lead in some circumstances to overly large adjustments at depth. We have not fully investigated the implications of this large log-space uncertainty at depth with the present experiments.

4. Discussion

The non-Gaussian statistics and non-negative character of biogeochemical variables suggests that data assimilation of these variables can be improved by adjustment of the underlying statistics. Fletcher (2010), Song et al. (2012) and Fletcher and Jones (2014) formulate the 4DVar for lognormally distributed variables, which can be applied to biogeochemical models.

Although incremental 4DVar with a lognormal distribution assumption (L4DVar) improves the estimation of states for lognormally distributed variables, the non-quadratic cost function limits its practical implementation to problems with small dimension. The incremental form for 4DVar with Gaussian distribution assumption (G4DVar) has a quadratic cost function and it is widely used in realistic problems because it is computationally more efficient than the nonlinear cost function formulation. In this study, incremental L4DVar is linearized with respect to the increment in log-space so that it has a quadratic cost function and can be easily implemented in realistic biogeochemical data assimilation problems in the ocean. Two additional linearization approximations for the nonlinear terms in the L4DVar cost function avoid any modification of the forward ecosystem model and made the computational cost of L4DVar comparable to that of G4DVar.

Twin experiments for the California Current System showed that the quadratic form of incremental L4DVar used here generally outperforms incremental G4DVar, with smaller posterior RMSE and better statistical representation of the true state. Quadratic incremental L4DVar allows appropriate adjustments at low concentrations, where incremental G4DVar struggles because the variance is close to zero in the original space. For example, the variance in log-space shows considerable model uncertainty at low levels offshore, and quadratic incremental L4DVar successfully reduced model data misfits there. Quadratic incremental L4DVar implicitly ensures positive concentrations, while incremental G4DVar can generate negative concentrations.

It is not obvious that negative concentrations resulting from assimilation are in practice a major problem. Most forward ecosystem models have the potential for negative values either due to losses associated with biological interactions (e.g., grazing of phytoplankton) or resulting numerically from the advection-diffusion implementation. Biological losses can be restricted to positive concentrations by using an implicit scheme (as is done in many ROMS ecosystem models) and advection-diffusion issues can be avoided by using a positive definite algorithm such as MPDATA (Margolin and Smolarkiewicz, 1998). Many ecosystem models address this issue with artificial corrections that simply make negative values positive. Such a crude fix could also be used with G4DVAR. Indeed, in our experiments, such a correction was imposed on the G4DVar analysis; that is, while the control variables (model initial conditions) determined by G4DVar included negative concentrations, the first step of the nonlinear model in each outer-loop resets these values to a small positive value, and the resulting output over the full cycle was reasonable overall. However, it is clearly desirable to avoid this numerical fix, and to accurately estimate small concentrations which quadratic incremental L4DVar does.

While the quadratic form of incremental L4DVar fits observations well and does so in a computationally efficient manner compared to non-quadratic incremental L4DVar, some caution is warranted. This approach requires two additional linearization approximations.The first is a Taylor expansion of the natural logarithm in observational space. If the prior model/data discrepancy is too large, the linear assumption is not accurate and leads to cost function convergence problems. We have found it better to exclude these observations from our procedure, though alternate approaches are possible. The second is a Taylor expansion of the exponential function in model space. We were not able to introduce a filter or method to ensure consistency with this approximation because validation can be examined only after assimilation, when the increment in log-space is determined. While it is possible that discrepancies between the background state and observations can result from a long assimilation window in which the tangent linear assumption is stretched, we do not believe that this issue is the major cause in this case. For an accurate background estimate resulting in small increments, the first order approximation is valid. Rather, model-data misfits occur sometimes simply because the background state is a poor estimate of truth, with model estimates in places far from observed values. We found that the accuracy of the background estimate is improved through sequential assimilation cycles, and that the last of 6 cycles was considerably more linear in this regard than the first. We note that our twin model experiment configuration may overestimate the improvement by sequential cycles, and we will have to revisit this issue in a more realistic setting.

This study used a twin experiment framework to investigate biogeochemical assimilation in isolation of errors in the physical circulation environment, by allowing erroneous fields at the start of each assimilation cycle only in bioegeochemical variables. In nature, uncertainties exist in the physical environment as well and further study is required to evaluate the quadratic incremental L4DVar developed here in a more general context. A natural next step is to consider the assimilation of both physical and biological fields simultaneously. Coupling of physical and ecosystem dynamics through the tangent linear and adjoint models and potentially through covariances would enable observations of biogeochemical variables to influence physical state estimates, and vice versa. For example, better biological estimates can result from by improving representation of oceanic mesoscale (i.e. eddies and current fields; Miller et al. (2000), Berline et al. (2007), Fiechter et al. (2011)) and lead to feedback to physical states (Murtugudde et al., 2002, Sweeney et al., 2005). However, unbalanced physical states at the start of each assimilation cycle can also drive erroneous biological fluctuations (Anderson et al., 2000). In one study, it was shown that assimilating biological variables did not substantially adjust the physical state estimates (Anderson et al., 2000), but additional investigation of this potential is warranted. From a practical point of view, a coupled physical-biological data assimilation system is desired because ocean observing systems are increasingly collecting both physical and biological information. A hybrid assimilation scheme including both G4DVar and L4DVar for different variables was introduced by Fletcher and Zupanski (2006b) and Fletcher and Jones (2014). In a companion paper, we develop this hybrid scheme for our oceanic application and explore the hybrid of incremental G4DVar and quadratic incremental L4DVar for physical and biological data assimilation, respectively.

5. Acknowledgement

We are grateful for grants from the Gordon and Betty Moore Foundation as well as from the National Oceanic and Atmospheric Administration (NA10OAR4320156) that supported this research. The authors would like to thank four anonymous reviewers for valuable comments and suggestions, which significantly improved the manuscript.

6. References

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

Anderson, L. A., Robinson, A. R., Lozano, C. J., 2000. Physical and biological modeling in the Gulf Stream region: I. Data assimilation methodology. Deep Sea Res. Pt I 47, 1787 – 1827.

Armstrong, R. A., Sarmiento, J. L., Slater, R. D., 1995. Monitoring ocean productivity by assimilating satellite chlorophyll into ecosystem models. In: Powell, S. (Ed.), Ecological Time Series. Chapman and Hall, London, pp. 371–390.

Banas, N. S., Lessard, E. J., Kudela, R. M., MacCready, P., Peterson, T. D., Hickey, B. M., Frame, E., 2009. Planktonic growth and grazing in the Columbia River plume region: A biophysical model study. J. Geophys. Res. 114, C00B06.

Bennett, A. F., 2002. Inverse Modeling of the Ocean and Atmosphere. Cambridge University Press.

Berline, L., Brankart, J. M., Brasseur, P., Ourmières, Y., Verron, J., 2007. Improving the physics of a coupled physical-biogeochemical model of the North Atlantic through data assimilation: Impact on the ecosystem. J. Marine Syst. 64, 153 – 172.

Broquet, G., Edwards, C. A., Moore, A. M., Powell, B. S., Veneziani, M., Doyle, J. D., 2009. Application of 4D-Variational data assimilation to the California Current System. Dynam. Atmos. Oceans 48, 69–92.

Broquet, G., Moore, A. M., Arango, H. G., Edwards, C. A., 2011. Corrections to ocean surface forcing in the California Current System using 4D variational data assimilation. Ocean Modell. 36, 116–132.

Bruland, K. W., Lohan, M. C., Aguilar-Islas, A. M., Smith, G. J., Sohst, B., Baptista, A., 2008. Factors influencing the chemistry of the nearfield Columbia River plume: Nitrate, silicic acid, dissolved Fe, and dissolved Mn. J. Geophys. Res. 113, C00B02.

Campbell, J. W., 1995. The lognormal distribution as a model for bio-optical variability in the sea. J. Geophys. Res. 100 (C7), 13237–13254.

Carton, J., Giese, B., 2008. A reanalysis of ocean climate using Simple Ocean Data Assimilation (SODA). Mon. Wea. Rev. 136, 29993017.

Courtier, P., Thépaut, J., Hollingsworth, A., 1994. A strategy for operational implementation of 4D-Var, using an incremental approach. Quart. J. Roy. Meteor. Soc. 120, 1367–1387.

Davis, K. A., Banas, N. S., Giddings, S. N., Siedlecki, S. A., MacCready, P., Lessard, E. J., Kudela, R. M., Hickey, B. M., 2014. Estuary-enhanced upwelling of marine nutrients fuels coastal productivity in the U.S. Pacific Northwest. J. Marine Syst. 119.

Doyle, J. D., Jiang, Q., Chao, Y., Farrara, J., 2009. High-resolution real-time modeling of the marine atmospheric boundary layer in support of the AOSN-II field campaign. Deep-Sea Res. Pt. II 56, 87–99.

Edwards, C. A., Moore, A. M., Hoteit, I., Cornuelle, B. D., 2015. Regional ocean data assimilation. Annu. Rev. Mar. Sci. 7, 6.1–6.22.

Fennel, K., Losch, M., Schröter, J., Wenzel, M., 2001. Testing a marine ecosystem model: sensitivity analysis and parameter optimazation. J. Marine Syst. 28, 45–63.

Fiechter, J., Broquet, G., Moore, A. M., Arango, H. G., 2011. A data assimilative, coupled physical-biological model for the coastal Gulf of Alaska. Dynam. Atmos. Oceans 51, 75–98.

Fiechter, J., Herbei, R., Leeds, W., Brown, J., Milliff, R., Wikle, C., Powell, T., Moore, A. M., 2013. A Bayesian parameter estimation method applied to a marine ecosystem model for the coastal Gulf of Alaska. Ecological Modelling Accepted.

Fiechter, J., Moore, A. M., Edwards, C. A., Bruland, K. W., Lorenzo, E. D., Lewis, C. V., Powell, T. M., Curchitser, E. N., Hedstrom, K., 2009. Modeling iron limitation of primary production in the coastal Gulf of Alaska. Deep Sea Res. Pt II 56 (24), 2503 – 2519.

Fisher, M., Courtier, P., 1995. Estimating the Covariance Matrices of Analysis and Forecast Error in Variational Data Assimilation. ECMWF Technical Memorandum 220, European Centre for Medium Range Weather Forecasts, Shinfield Park, Reading, UK.

Fletcher, S. J., 2010. Mixed Gaussian-lognormal four-dimensional data assimilation. Tellus A 62, 266–287.

Fletcher, S. J., Jones, A. S., 2014. Multiplicative and additive incremental variational data assimilation for mixed lognormal-gaussian errors. Mon. Wea. Rev. 142, 2521–2544.

Fletcher, S. J., Zupanski, M., 2006a. A data assimilation method for lognormally distributed observational errors. Quart. J.Roy. Meteor. Soc. 132, 2505–2519.

Fletcher, S. J., Zupanski, M., 2006b. A hybrid multivariate normal and lognormal distribution for data assimilation. Atmosph. Sci. Lett. 7, 43–46.

Forget, G., 2010. Mapping ocean observations in a dynamical framework: A 2004-06 ocean atlas. J. Phys. Oceanogr. 40, 1201–1221.

Friedrichs, M. A. M., 2001. Assimilation of JGOFS EqPac and SeaWiFS data into a marine ecosystem model of the central equatorial Pacific ocean. Deep Sea Res. Pt II 49, 289 – 319.

Goebel, N. L., Edwards, C. A., Zehr, J. P., Follows, M. J., 2010. An emergent community ecosystem model applied to the California Current System. J. Marine Syst. 83.

Gregg, W. W., 2008. Assimilation of SeaWiFS ocean chlorophyll data into a three-dimensional global ocean model. J. Marine Syst. 69, 205 – 225.

Gregg, W. W., Friedrichs, M. A., Robinson, A. R., Rose, K. A., Schlitzer, R., Thompson, K. R., Doney, S. C., 2009. Skill assessment in ocean biological data assimilation. J. Marine Syst. 76 (1-2), 16 – 33.

Hickey, B. M., Banas, N. S., 2008. Why is the northern end of the California current system so productive? Oceanography 21 (4), 90–107.

Hoteit, I., Triantafyllou, G., Petihakis, G., Allen, J., 2003. A singular evolutive extended Kalman filter to assimilate real in-situ data in a 1-D marine ecosystem model. Ann. Geophys. 21, 389–397.

Hu, J., Fennel, K., Mattern, J. P., Wilkin, J., 2012. Data assimilation with a local ensemble Kalman filter applied to a three-dimensional biological model of the Middle Atlantic Bight. J. Marine Syst. 94, 145 – 156.

Large, W. G., Yeager, S. G., 2009. The global climatology of an interannually varying air-sea flux data set. Climate Dyn. 33, 341–364, 10.1007/s00382-008-0441-3.

Lawless, A. S., Gratton, S., Nichlos, N. K., 2005. Approximate iterative methods for variational data assimilation. Int. J. Numer. Methods Fluids 47, 1129–1135.

Lorenc, A. C., 1986. Analysis methods for numerical weather prediction. Q. J. R. Meteorol. Soc. 112, 1177–1194.

Margolin, L., Smolarkiewicz, P. K., 1998. Antidiffusive velocities for multipass donor cell advection. SIAM J. Sci. Comput. 20, 907–929.

Mattern, J. P., Fennel, K., Dowd, M., 2012. Estimating time-dependent parameters for a biological ocean model using an emulator approach. J. Marine Syst. 96–97, 32–47.

McGillicuddy, D. J. J., Lynch, D. R., Moore, A. M., Gentleman, W. C., Davis, C. S., Meise, C. J., 1998. An adjoint data assimilation approach to diagnosis of physical and biological controls on Pseudocalanus spp. in the Gulf of MaineGeorges Bank region. Fish. Oceanogr. 7 (3-4), 205–218.

Miller, A. J., Di Lorenzo, E., Neilson, D. J., Cornuelle, B. D., Moisan, J. R., 2000. Modeling CalCOFI observations during El Niño: Fitting physics and biology. Calif. Coop. Ocean. Fish. Invest. Rep. 41, 87–97.

Moisan, J. R., Hofmann, E. E., Haidvogel, D. B., 1996. Modeling nutrient and plankton processes in the California coastal transition zone 2. A three-dimensional physical-bio-optical model. J. Geophys. Res. 101, 226677–22691.

Moore, A. M., Arango, H. G., Broquet, G., Edwards, C. A., Veneziani, M., Powell, B. S., Foley, D., Doyle, J., Costa, D., Robinson, P., 2011a. The

Regional Ocean Modeling System (ROMS) 4-dimensional variational data assimilation systems, Part II: Performance and application to the California Current System. Prog. Oceanogr. 91, 50–73.

Moore, A. M., Arango, H. G., Broquet, G., Powell, B. S., Zavala-Garay, J., Weaver, A. T., 2011b. The Regional Ocean Modeling System (ROMS) 4-dimensional variational data assimilation systems, Part I: Formulation and Overview. Prog. Oceanogr. 91, 34–49.

Moore, T. S., Campbell, J. W., Dowell, M. D., 2009. A class-based approach to characterizing and mapping the uncertainty of the MODIS ocean chlorophyll product. Remote Sens. Environ. 113 (11), 2424 – 2430.

Murtugudde, R., Beauchamp, J., McClain, C. R., Lewis, M., Busalacchi, A. J., 2002. Effects of Penetrative Radiation on the Upper Tropical Ocean Circulation. J. Climate 15 (5), 470–486.

Natvik, L. J., Eknes, M., Evensen, G., 2001. A weak constraint inverse for a zero-dimensional marine ecosystem model. J. Marine Syst. 28 (1-2), 19 – 44.

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

Oke, P., Larnicol, G., Fujii, Y., Smith, G., Lea, D., Guinehut, S., Remy, E., Balmaseda, M. A., Rykova, T., Surcel-Colan, D., Martin, M., Sellar, A., Mulet, S., Turpin, V., 2015a. Assessing the impact of observations on ocean forecasts and reanalyses: Part 1, Global studies. Journal of Operational Oceanography 8 (sup1), s49–s62.

Oke, P. R., Larnicol, G., Jones, E. M., Kourafalou, V., Sperrevik, A., Carse, F., Tanajura, C. A. S., Mourre, B., Tonani, M., Brassington, G. B., Henaff, M. L., Halliwell, G. R., Atlas, R., Moore, A. M., Edwards, C. A., Martin, M. J., Stellar, A. A., Alvarez, A., Mey, P. D., Iskandarani, M., 2015b. Assessing the impact of observations on ocean forecasts and reanalyses: Part 2, Regional applications. Journal of Operational Oceanography 8 (sup1), s63–s79.

Popova, E., Lozano, C., Srokosz, M., Fasham, M., Haley, P., Robinson, A., 2002. Coupled 3D physical and biological modelling of the mesoscale variability observed in North-East Atlantic in spring 1997: biological processes. Deep Sea Res. Pt I 49 (10), 1741 – 1768.

Powell, B. S., Arango, H. G., Moore, A. M., Di Lorenzo, E., Milliff, R. F., Foley, D., 2008. 4DVAR data assimilation in the Intra-Americas Sea with the Regional Ocean Modeling System (ROMS). Ocean Modell. 23, 130–145.

Powell, T., Lewis, C., Curchitser, E., Haidvogel, D., Hermann, A., Dobbins, E., 2006. Results from a three-dimensional, nested, biologicalphysical model of the california current system and comparisons with statistics from satellite imagery. J. Geophys. Res. 111, C07018.

Schlitzer, R., 2000. Applying the adjoint method for global biogeochemical modeling. In: Kasibhatla, P., Heimann, M., Hartley, D., Mahowald, N., Prinn, R., Rayner, P. (Eds.), Inverse Methods in Global Biogeochemical Cycles. Vol. 114, of Geophys. Monograph Series. American Geophysical Union, Washington, D. C., pp. 107–124.

Song, H., Edwards, C. A., Moore, A. M., Fiechter, J., 2012. Four-dimensional variational data assimilation of positive-definite oceanic variables using a logarithm transformation. Ocean Modell. 54–55, 1–17.

Song, H., Edwards, C. A., Moore, A. M., Fiechter, J., 2016. Data assimilation in a coupled physical-biogeochemical model of the California Current System using an incremental lognormal 4-dimensional variational approach: Part 3, Assimilation in a realistic context using satellite and in situ observations. Ocean Modell. submitted.

Stammer, D., Wunsch, C., Giering, R., Eckert, C., Heimbach, P., Marotzke, J., Adcroft, A., Hill, C. N., Marshall, J., 2002. The global ocean circulation during 1992-1997, estimated from ocean observations and a general circulation model. J. Geophys. Res. 107 (C9), 3118.

Sweeney, C., Gnanadesikan, A., Griffies, S. M., Harrison, M. J., Rosati, A. J., Samuels, B. L., Jun. 2005. Impacts of Shortwave Penetration Depth on Large-Scale Ocean Circulation and Heat Transport. J. Phys. Oceanogr. 35 (6), 1103–1119.

Tijputra, J., Polzin, D., Winguth, A., 2007. Assimilation of seasonal chlorophyll and nutrient data into an adjoint three-dimensional ocean carbon cycle model: sensitivity analysis and ecosystem parameter optimization. Global Biogeochem. Cycles 21, GB1001.

Tshimanga, J., Gratton, S., Weaver, A. T., Sartenaer, A., 2008. Limited-memory preconditioners, with application to incremental four-dimensional variational data assimilation. Q. J. R. Meteorol. Soc. 134, 751–769.

Weaver, A., Courtier, P., 2001. Correlation modelling on the sphere using a generalized diffusion equation. Quart. J. Roy. Meteorol. Soc. 127, 1815–1846.

Wunsch, C., Heimbach, P., 2007. Practical global oceanic state estimation. Physica D 230, 197–208.

Table 1: Parameter names, values and units for the NPZD model

Parameter name	Value	Units
Light		
Extinction coefficient for sea water	0.067	m^{-1}
Photosynthetically active radiation (PAR)	0.43	Nondimensional
Phytoplankton		
Self-shading coefficient	0.02	m^2 mmol N^{-1}
Initial slope of P-I curve	0.02	m^2 W^{-1}
Uptake rate for nitrate	1.0	day^{-1}
Half-saturation constant for nitrate	1.0	mmol N m^{-3}
Mortality rate	0.1	day^{-1}
Zooplankton		
Grazing rate	0.65	day^{-1}
Ivlev constant	1.4	Nondimensional
Excretion efficiency	0.3	Nondimensional
Mortality rate	0.145	day^{-1}
Detritus		
remineralization rate	0.1	day^{-1}
Sinking velocity	40	m day^{-1}

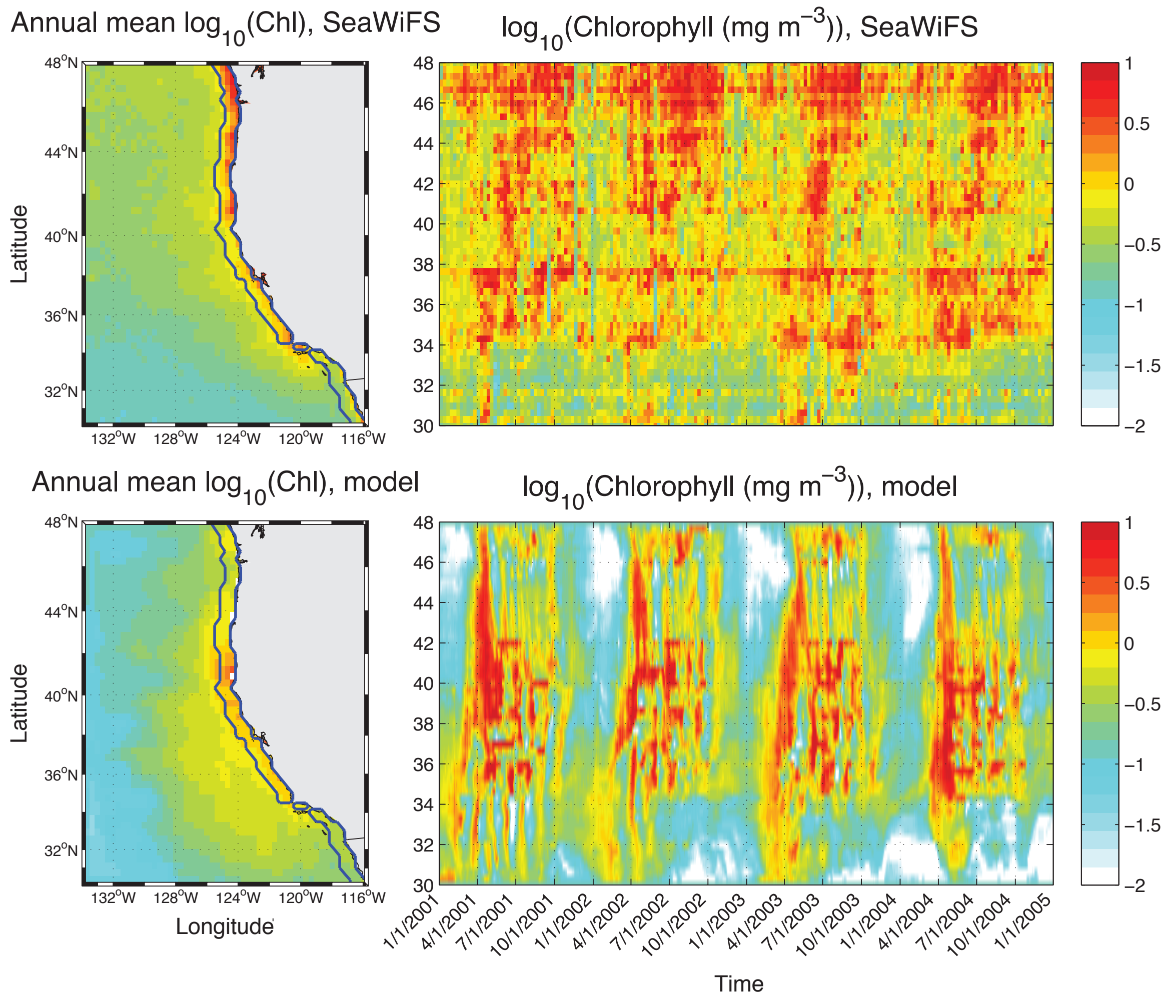

Figure 1: Panels in the left column show annual averaged surface $\log_{10}$(chlorophyll (mg m^{-3})) from the SeaWiFS (top) and from the model simulation (bottom). Blue contours bound an area from the coast to about 100 km offshore. Panels in the right column are latitude-time plots of surface $\log_{10}$(chlorophyll (mg m^{-3})) averaged over the area within the blue contours for the SeaWiFS (top) and the model simulation (bottom).

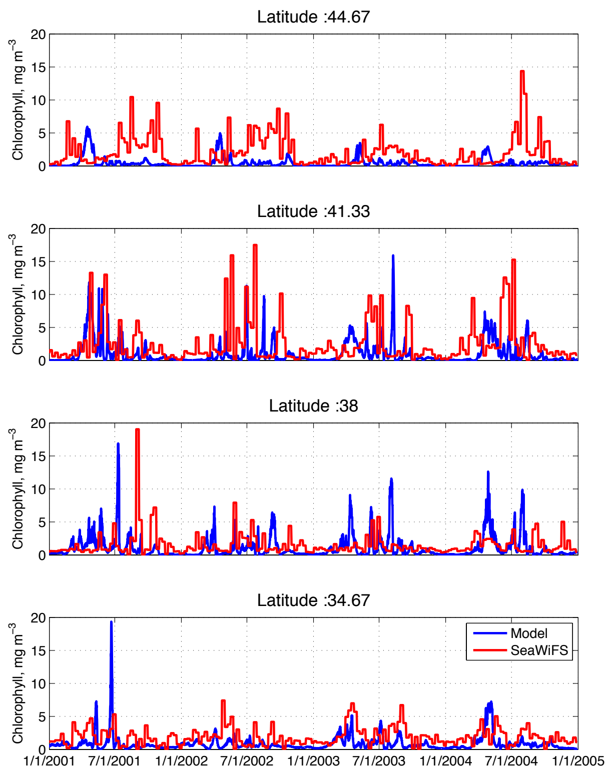

Figure 2: Time series of chlorophyll at the coast at four different latitudes: 34.67°N, 38°N, 41.33°N and 44.67°N. Chlorophyll from the SeaWiFS data and the model are plotted in red and blue, respectively.

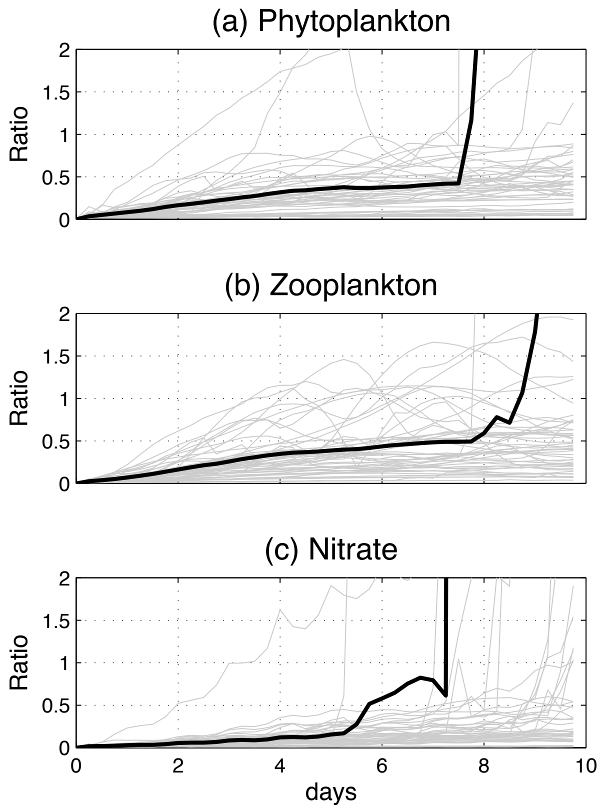

Figure 3: The growth of the proportion of nonlinear dynamics to the total perturbation, δ/Δ, in time for surface (a) phytoplankton, (b) zooplankton and (c) nitrate. Forty eight grey lines represent each month during a 4-year simulation, and black lines are the ensemble mean.

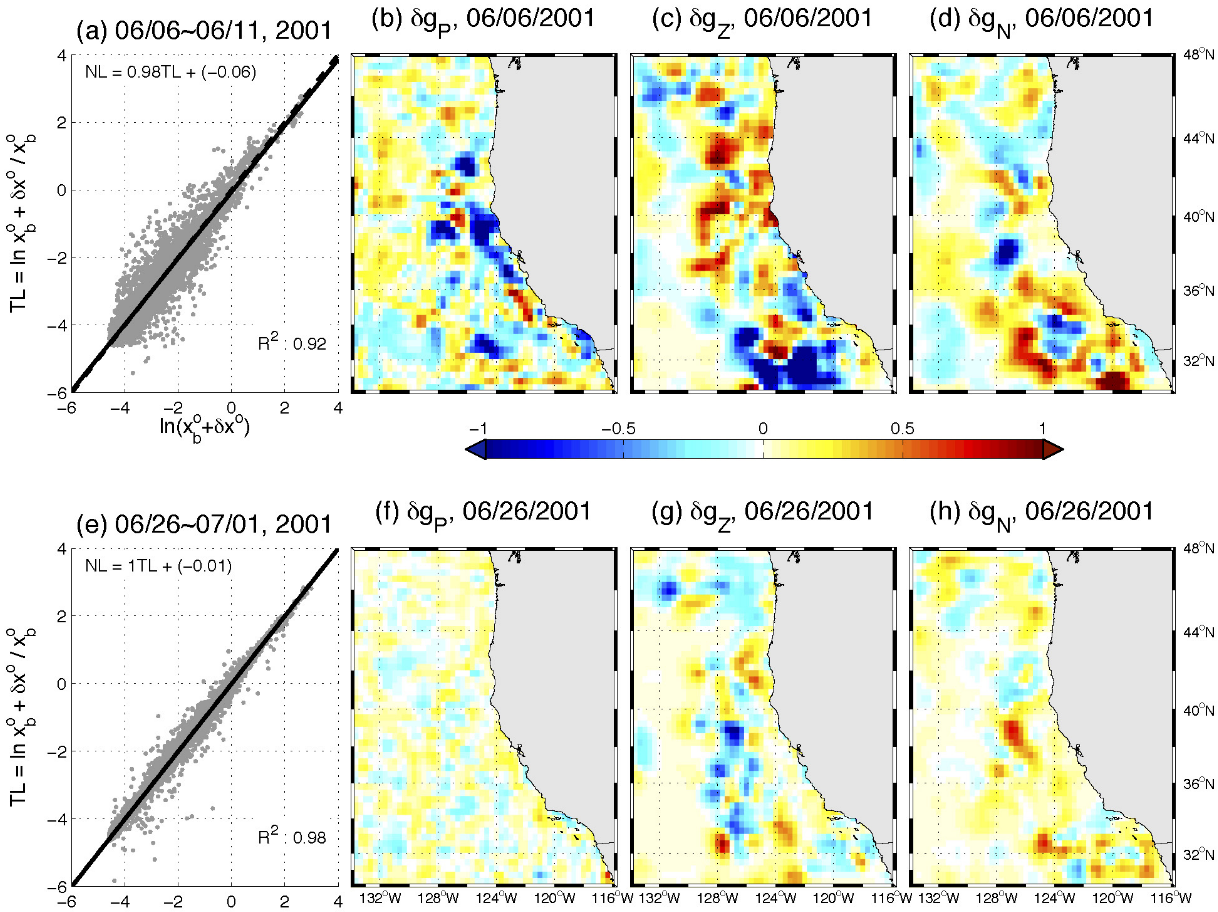

Figure 4: Comparison between $\ln(\mathbf{x}_b^o + \delta\mathbf{x}^o)$ and its first order linear approximation during the first cycle after the 5-day spinup period (a) and the last cycle (e) in June 2001. Solid and dashed lines represent the linear fit and straight lines with slope 1, respectively. The increments in log-space are plotted for P (b, f), Z (c, g) and N (d, h) during the first cycle (b, c, d) and the last cycle (f, g, h) in June 2001. It is noted that the dashed lines are not clearly visible because they are under the solid lines.

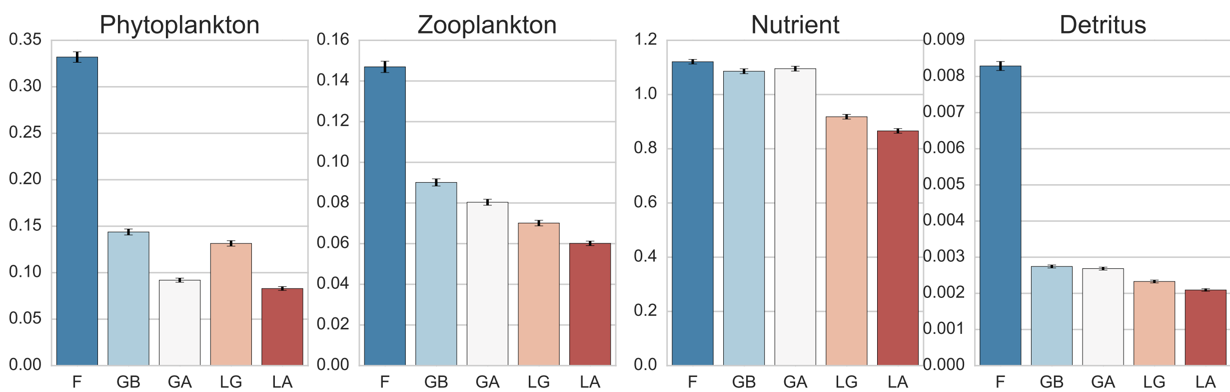

Figure 5: Seasonal RMSEs for the free run state (F), background (GB) and analysis (GA) by incremental G4DVar, and background (LG) and analysis (LA) by quadratic form of incremental L4DVar. The error bars (black) represent the standard error.

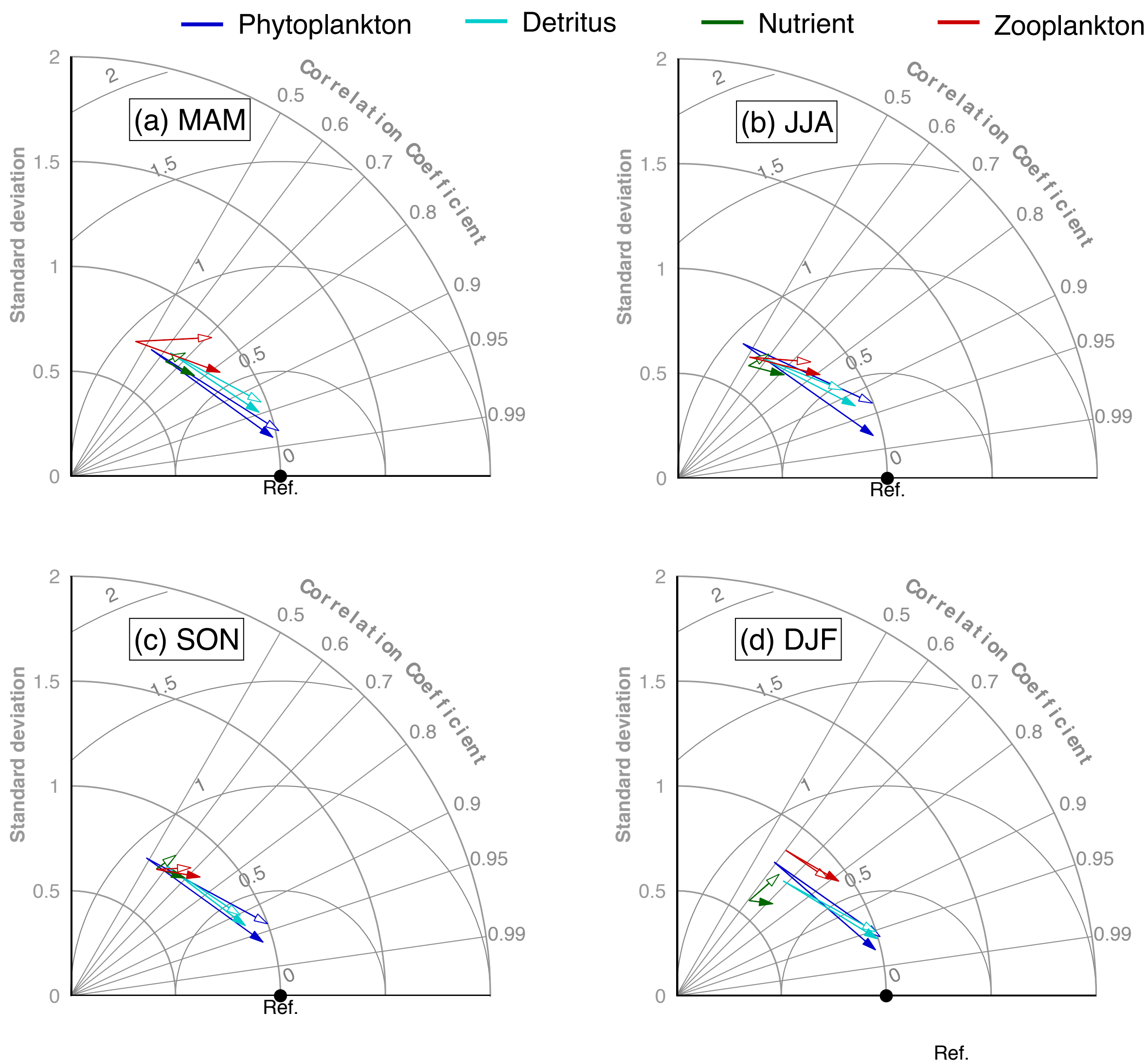

Figure 6: Seasonal Taylor diagrams showing the statistical improvements in surface P (blue), Z (red), N (green) and D (cyan) by G4DVar (open arrowhead) and L4DVar (closed arrowhead). Arrows start at the background state and point to the analysis state. The reference state (Ref, black dots) indicates the direction of statistical improvement for the assimilation system.

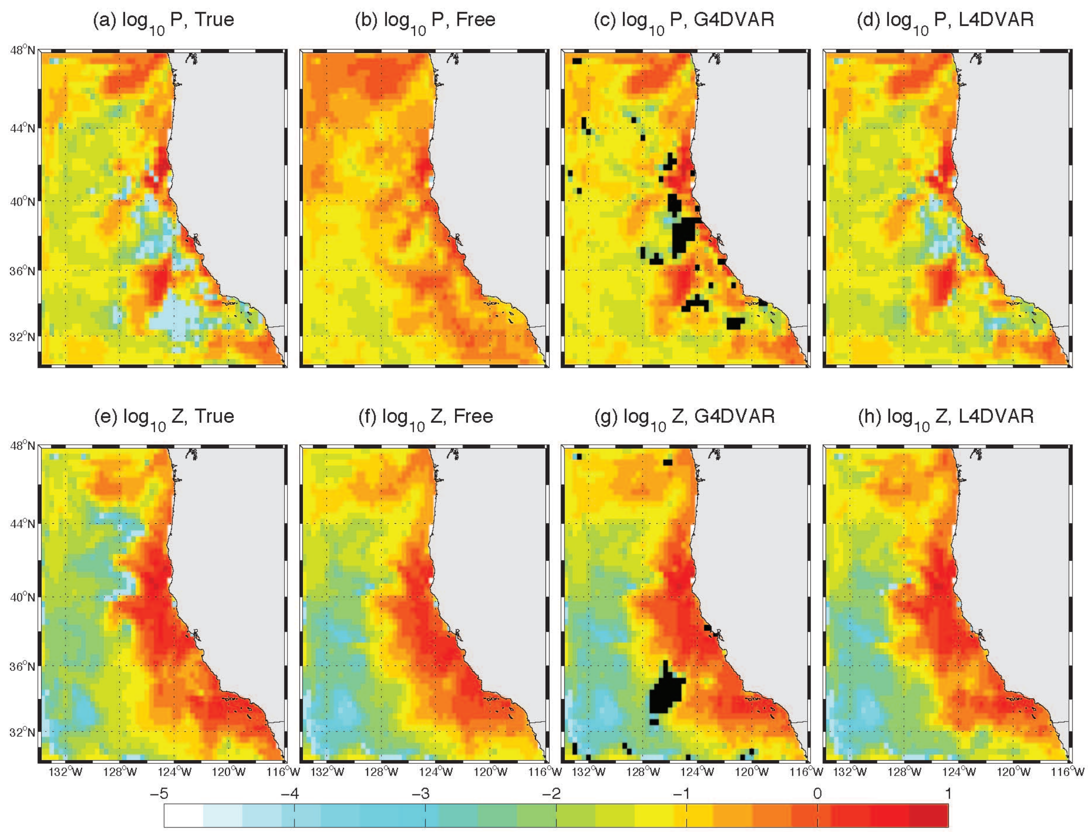

Figure 7: The initial condition of surface P (a-d) and Z (e-h) on a log-scale from four simulations: truth (a, e), free run (b, f), incremental G4DVar posterior (c, g) and quadratic incremental L4DVar posterior (d, h) on June 6^{st}, 2001. Black represent areas with negative concentration.

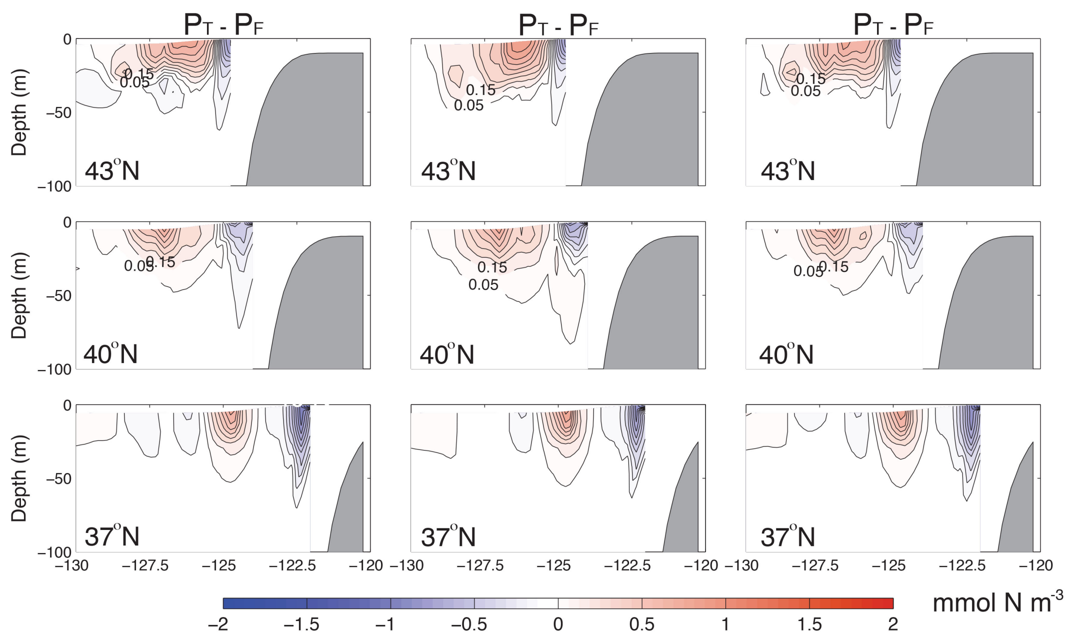

Figure 8: The vertical cross-sections of P differences between in the truth and free run at three latitudes (37°N, 40°N and 43°N) on June 16^{st}, 2001. The first, second and third column on the right show the desirable adjustment, the realized adjustment by incremental G4DVar and the realized adjustment by quadratic incremental L4DVar, respectively.

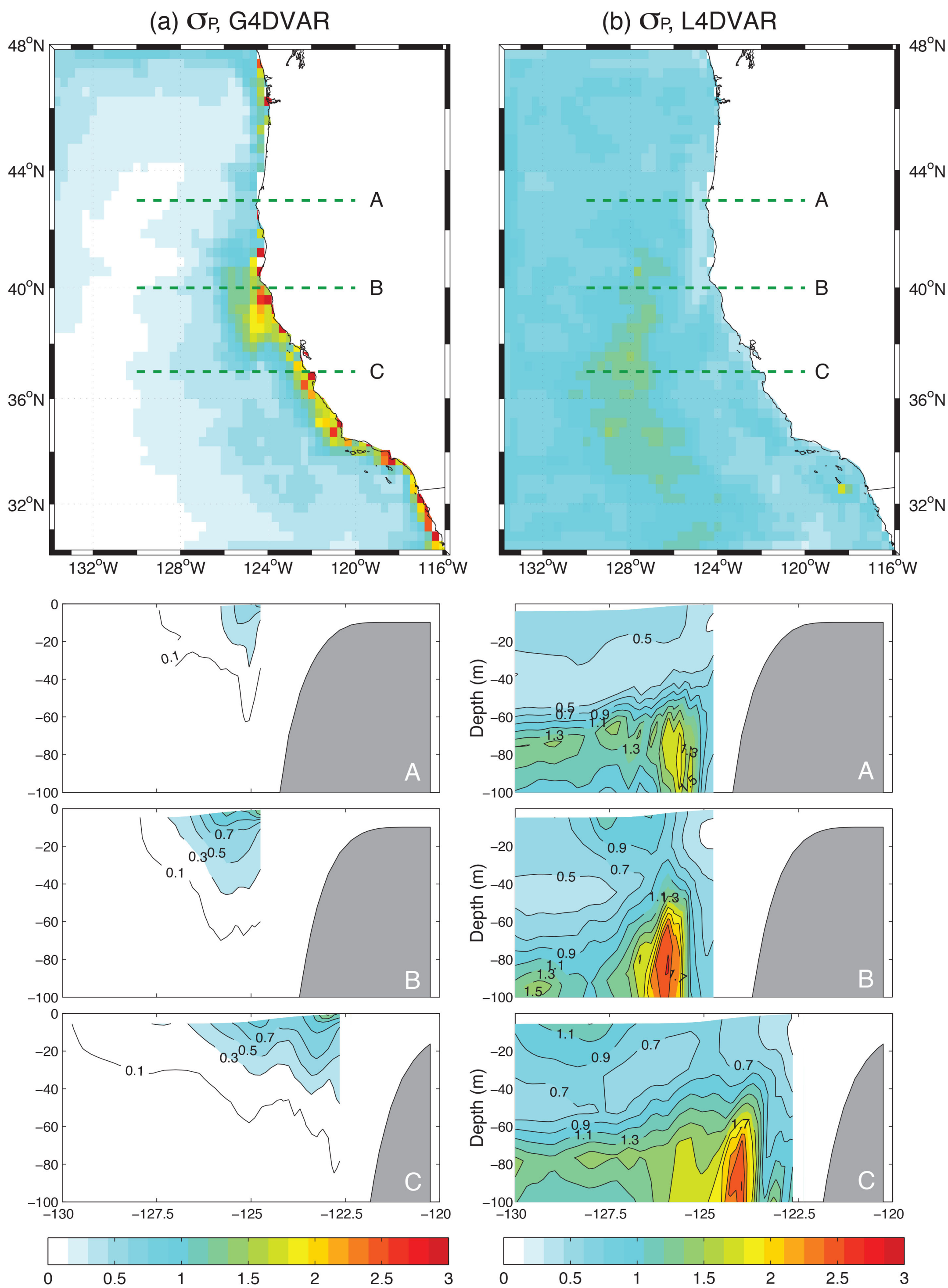

Figure 9: The standard deviation of surface P, used to generate the diagonal components of the model error covariances, in the original linear space (a) and in log-space (b). Panels below show the vertical cross-sections at three latitudes (37°N, 40°N and 43°N) in linear space (left column) and in the log-space (right column).