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

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 02139, U.S.A.*

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

Abstract

A fully coupled physical and biogeochemical ocean data assimilation system is tested in a realistic configuration of the California Current System using the Regional Ocean Modeling System. In situ measurements for sea surface temperature and salinity as well as satellite observations for temperature, sea level and chlorophyll are used for the year 2000. Initial conditions of the combined physical and biogeochemical state are adjusted at the start of each 3-day assimilation cycle. Data assimilation results in substantial reduction of root-mean-square error (RMSE) over unconstrained model output. RMSE for physical variables is slightly lower when assimilating only physical variables than when assimilating both physical variables and sur-

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

Preprint submitted to Ocean Modelling *June 5, 2016*

face chlorophyll. Surface chlorophyll RMSE is lowest when assimilating both physical variables and surface chlorophyll. Estimates of subsurface, nitrate and chlorophyll show modest improvements over the unconstrained model run relative to independent, unassimilated in situ data. Assimilation adjustments to the biogeochemical initial conditions are investigated within different regions of the California Current System. The incremental, lognormal 4-dimensional data assimilation method tested here represents a viable approach to coupled physical biogeochemical state estimation at practical computational cost.

Keywords: Coupled Data assimilation, 4DVar, Biogeochemical model, California Current System, Coastal upwelling

1. Introduction

The study of marine ecosystems in regional environments is motivated by a wide range of topics, spanning fundamental questions concerning controls on primary production, community structure and carbon export to more applied problems in fisheries management, harmful algal blooms, and habitat monitoring, to name but a few. Investigations generally require quantification of various elements of the physical and/or biogeochemical constituents, such as temperature, salinity, phytoplankton biomass, and processes such as nutrient uptake or grazing. Short space and time scales of variability in the coastal ocean present a challenge for direct and comprehensive observation of key variables, though real progress in observing sensors and platforms has been accomplished over the last decade.

Coupled physical and biogeochemical models provide a complementary approach to direct observation for the study of marine ecosystems. Worldwide, a handful of advanced physical circulation models are widely used as backbones for a much larger assortment of biogeochemical models that range in complexity and purpose. Such coupled models show increasing skill in representing marine ecosystems, but discrepancies between model predictions and observations are inevitable. Such errors arise from multiple unavoidable issues such as uncertainty in model initialization and forcing as well as incomplete or incorrect parameterization of basic model processes.

One approach to reduce discrepancies between ocean model output and observations is through data assimilation, where observations are used to rigorously constrain ocean model trajectories. Data assimilation of the physical circulation is well-established and carried out routinely on global and regional scales. The assimilation of ecosystem variables into coupled physical-biogeochemical models is less advanced, although considerable progress has been made over the last two decades. Biogeochemical data assimilation has been used to constrain model parameters, some of which are poorly known, and to improve estimates of the biogeochemical state, and sometimes for both purposes (See Gregg (2008) and Edwards et al. (2015) for recent reviews).

In two companion papers, we implemented a new formulation for biogeochemical and coupled physical-biogeochemical data assimilation for ocean state estimation (Song et al., 2016,). Our approach is an incremental form of lognormal 4-dimensional variational assimilation (4DVar), first proposed by Song et al. (2012) and described further by Fletcher and Jones (2014). We choose a lognormal formulation because of the skewed statistical distributions of biological variables that are clearly non-Gaussian and better represented as lognormal (Campbell, 1995). We have implemented this capability within the Regional Ocean Modeling System (ROMS; Shcheptkin and McWilliams, 2004), building on its existing 4DVar capabilities developed for physical variables (Moore et al., 2011,).

In idealized model twin experiments, Song et al. (2016) show that the lognormal form of 4DVar produces superior state estimates with lower root-mean-square errors (RMSEs) for biological fields relative to those derived assuming Gaussian error distributions. Song et al. (2016) implemented a fully coupled physical and biogeochemical system allowing the simultaneous assimilation of both Gaussian and lognormally distributed errors following Fletcher (2010) and Fletcher and Jones (2014). Tests in an idealized model twin experiment compared data assimilation of only physical variables, only biological variables, and both physical and biological variables. The lowest RMSE for both the physical and biogeochemical variables of the modeled ocean state resulted from the assimilation of both physical and biological observations.

Model twin experiment is a useful guide for understanding model performance, but ultimately is limited because the assimilation model is identical to that used as a surrogate for the true state. In a real application, the model is imperfect and thus unable to exactly match nature. It is the purpose of this paper to test the fully coupled 4DVar data assimilation system in a realistic environment, and we choose the California Current System (CCS) as our testbed.

The CCS refers to a collection of ocean currents and other circulation features in the northeastern subtropical Pacific. As with other eastern boundary regions, the CCS experiences seasonally vigorous upwelling driven by equatorward wind stress near the coast. The wind-driven upwelling supplies nutrients to the euphotic zone and drives substantial primary production, ultimately supporting a disproportionately rich and complex ecosystem relative to its small area (Carr, 2002). The present investigation builds on several previous modeling studies of the CCS, including efforts to describe the physical circulation using forward, adjoint, and data assimilative models (Veneziani et al., 2009,, Broquet et al., 2009, 2011), and various aspects of the CCS ecosystem using non-data assimilative coupled physical-biogeochemical models of varying complexity (Goebel et al., 2010, Fiechter et al., 2014).

Here we evaluate coupled physical-biogeochemical data assimilation using ROMS and a simple 4-component Nutrient-Phytoplankton-Zooplankton-Detritus (NPZD) model (Powell et al., 2006) for one calendar year (2000). Physical data assimilated includes sea surface height, sea surface temperature, and in situ temperature and salinity. For biogeochemical data, we assimilate satellite-derived sea surface chlorophyll. In situ nitrate and chlorophyll observations from two field programs are withheld for independent evaluation. Model initial conditions at the start of each assimilation cycle are adjusted. We demonstrate both the utility of this approach in a realistic and practical implementation and also investigate how the assimilation system functions in different regions of the CCS for which different unconstrained (prior) model deficiencies are identified with respect to the observations.

2. Coupled data assimilation system

Song et al. (2016) present a full description of the physical and biogeochemical data assimilation (PBDA) procedure, and we include here only an abbreviated version. Using the 4-dimensional variational method, updates to a control vector are based on all available observations within an assimilation window. In general, the control vector can include multiple elements, such as model forcing fields and open boundary conditions (Moore et al., 2011), but for the present investigation, we consider for simplicity only a control vector consisting of model initial conditions.Describing the additional impact of adjustments to model forcing and lateral boundary conditions is left to future studies. The increment to the background initial state is denoted $\delta\mathbf{z}_0$ and consists of both physical and biological elements; more specifically, $\delta\mathbf{z}_0^T = [(\delta\mathbf{x}_0^{phy})^T(\delta\mathbf{x}_0^{bio})^T]$, where $\delta\mathbf{x}_0^{phy} = (\mathbf{x}_a^{phy} - \mathbf{x}_b^{phy})_0$ and $\delta\mathbf{x}_0^{bio} = (\mathbf{x}_a^{bio} - \mathbf{x}_b^{bio})_0$ are the $(n_g \times 1)$ and $(n_l \times 1)$ increment vectors of physical and biological variables at the initial time, respectively. The subscript a/b represents the posterior/prior solution.

Some biogeochemical variables are known to have non-Gaussian distributions, with better consistency with lognormal distributions (Campbell, 1995, Campbell et al., 1995). As a result, the increments $\delta\mathbf{z}_0$ will not be Gaussian-distributed, and a solution assuming Gaussian errors for all variables will not be optimal. We proceed with the assumption that physical variables have Gaussian distributed errors while errors in biogeochemical variables are lognormally distributed. Though the lognormal assumption is likely also imperfect, it allows a straightforward solution to the assimilation problem, and this solution has been shown in model twin experiments to be superior to the Gaussian assumption for biogeochemical variables (Song et al., 2016).

By definition, a logarithm transformation of lognormally distributed variables results in Gaussian distributed values, and the difference between Gaussian distributed variables also has a Gaussian distribution. As a result, we define $\delta\mathbf{g}_0^{bio} = (\ln \mathbf{x}_a^{bio} - \ln \mathbf{x}_b^{bio})_0$ whose distribution is Gaussian. In addition, if $\delta\mathbf{x}_0^{phy}$ is Gaussian distributed, the new control vector $\delta\mathbf{z}_0^T = [(\delta\mathbf{x}_0^{phy})^T(\delta\mathbf{g}_0^{bio})^T]$ will also be drawn from a Gaussian distribution.

The optimal solution for $\delta\mathbf{z}_0$ minimizes the cost function J:

$$\begin{aligned} J(\delta\mathbf{z}_0) &= \frac{1}{2}\delta\mathbf{z}_0^T\mathbf{B}^{-1}\delta\mathbf{z}_0 \\ &\quad + \frac{1}{2}\sum_{i=1}^{N_o}\left(\mathbf{d}_i - \mathbf{O}_i\mathbf{H}_i\mathbf{M}_{i,0}\mathbf{X}\delta\mathbf{z}_0\right)^T\mathbf{R}_i^{-1} \\ &\quad \left(\mathbf{d}_i - \mathbf{O}_i\mathbf{H}_i\mathbf{M}_{i,0}\mathbf{X}\delta\mathbf{z}_0\right). \end{aligned} \tag{1}$$

Here, $\mathbf{d}_i$ defines the innovations that can be partitioned into linear and log-space. More specifically, $\mathbf{d}_i = \mathbf{y}_i - \mathbf{x}_{b,i}^o$ for Gaussian distributed variables and $\mathbf{d}_i = \ln \mathbf{y}_i - \ln \mathbf{x}_{b,i}^o$ for lognormally distributed variables, where $\mathbf{y}_i$ represents the i^{th} set of observations, and $\mathbf{x}_{b,i}^o$ indicates the corresponding background model estimates. $\mathbf{d}_i$ can be a mixture of both Gaussian and lognormally distributed variables. The matrices $\mathbf{H}_i$ and $\mathbf{M}_{i,0}$ are tangent linear forms of the observation operator and nonlinear model, respectively. Diagonal matrices $\mathbf{O}_i$ and $\mathbf{X}$ have diagonal elements $[1,\ 1,\ \ldots,\ 1,\ (\mathbf{x}_{b,i}^o)_1,\ (\mathbf{x}_{b,i}^o)_2,\ \ldots,\ (\mathbf{x}_{b,i}^o)_{m_l}]^{-1}$ and $[1,\ 1,\ \ldots,\ 1,\ (\mathbf{x}_{b,0})_1,\ (\mathbf{x}_{b,0})_2,\ \ldots,\ (\mathbf{x}_{b,0})_{n_l}]$, respectively, where m_l is the number of observations for lognormally distributed variables. Matrices $\mathbf{B}$ and $\mathbf{R}$ denote the background and observation error covariance matrices, respectively, and will be discussed further below. This cost function is quadratic, hence its optimal solution can be found using traditional methods such as conjugate gradient. The Jacobian and Hessian of this cost function can be found in a companion paper Song et al. (2016) along with more details.

3. Observations

Physical and biological observations were used to constrain the model during the year 2000. More than 3 million physical observations including sea surface height (SSH), sea surface temperature (SST), subsurface T and S are used. In addition, more than a million surface chlorophyll data are available for coupled physical and biological state estimation (Table 1).

3.1. Physical observations

A brief introduction to the physical observations is provided here, but detailed descriptions about the data set and preprocessing can be found in Moore et al. (2011) for the same collection of physical observations used in this study. For SSH observations, we use the sum of the mapped sea level anomaly product from Ssalto-Duacs system and the mean dynamic topography estimated by Rio et al. (2004). Mean sea level is adjusted so that the unconstrained model and data have the same spatio-temperal mean value. Temporal and spatial resolution of the data are 7 days and 1/3°, respectively, while the observation error is set to 0.02 m. Daily assimilated SST data is from the Advanced Very High Resolution Radiometer (AVHRR) with a horizontal resolution of approximately 0.04° (Kilpatrick et al., 2001). The observation error assumed for SST is set to 0.4 °C. In situ T and S observations come from the quality controlled data prepared by the European Union ENSEMBLES project (EN3) (Ingleby and Huddleston, 2007). This data set includes CTD profiles sampled during the CalCOFI program from the southern and central CCS, and GLOBEC-LTOP survey cruises from the northern CCS. Observation errors for in situ T and S are assumed to be 0.1

°C and 0.01, respectively.

3.2. Biological observations

The biological model, NPZD (nutrients, phytoplankton, zooplankton and detritus), solves for phytoplankton biomass, instead of chlorophyll. We first import the SeaWiFS level 3 Standard Mapped Image (SMI) products with roughly 9 km horizontal resolution and then convert chlorophyll observations in units of mg m^{-3} to phytoplankton units of mmol N m^{-3}. The carbon to nitrogen conversion is based on a Redfield ratio (C:N=106:16), and a chlorophyll to carbon ratio of C:Chl=50:1, which is reasonable for diatoms (i.e., the dominant phytoplankton species associated with coastal upwelling) in the California Current System (Goebel et al., 2010). Although satellite observations represent an integral over an optical depth, we choose for this study the more simple approach of assimilating satellite-derived estimates of phytoplankton biomass into the uppermost model level. The error level for phytoplankton biomass data in log-transformed space is set to 0.3, which is approximately 30 % of the observed value (±35% for chlorophyll in Moore et al. (2009)).

The SeaWiFS daily chlorophyll data does not provide good temporal coverage in the coastal regions during 2000. Temporal data coverage in coastal areas (which we define here as a nearshore strip approximately 100 km wide and indicated by the blue line in Figures 1a and c) is in fact less than 30%. As shown in Figure 1b, the temporal coverage for coastal chlorophyll is particularly low in winter at higher latitudes, which is most likely associated with the passage of storm systems. NASA's Ocean Biology Processing Group also provides an 8-day composite product with close to 100% data coverage after computing a temporal and spatial weighted mean (Campbell et al., 1995). In this product, chlorophyll values are fixed for 8 days, whereas both physical and biological processes in coastal regions generally vary considerably on shorter time-scales. Although the spatial coverage of the 8-day product is good, the temporal variations captured are questionable. Therefore, we use the daily chlorophyll data with low spatial coverage and with the expectation that the data assimilation system will estimate missing observations using model dynamics and error covariances. This interpolation capability is one potential benefit of the 4DVar data assimilation.

For our investigation, we also consider subsurface biological data. Specifically, chlorophyll and nitrate (NO_3) from the CalCOFI and GLOBEC-LTOP programs were available during the time-period of our experiment, and their locations are shown in dots in Figure 1(a). These data are not assimilated but used only for the evaluation of the coupled state estimates.

If more than one observation of a single type (e.g., temperature) is available in a model grid cell within one day, all observations of this type are merged into a single value. This creation of "super observations" reduces data redundancy, and an appropriate level of error for the merged data is determined by the uncertainties of all observations within a model grid cell (Moore et al., 2011).

3.3. Observation filter

Not all the biological observations were used in the system. As reported in Song et al. (2016), our quadratic lognormal 4DVar formulation requires a linear approximation to the logarithm transform as follows:

$$\begin{aligned}\ln\left(\mathbf{x}^o_{b,i}+\delta\mathbf{x}^o_i\right) &\approx \ln\mathbf{x}^o_{b,i}+\mathbf{L}_i\delta\mathbf{x}^o_i \\ &\approx \ln\mathbf{x}^o_{b,i}+\mathbf{L}_i\mathbf{H}_i\mathbf{M}_{i,0}\delta\mathbf{x}_0,\end{aligned} \tag{2}$$

where $\delta\mathbf{x}^o_i$ represents the increment for the i^{th} observation set, and

$$\begin{aligned}\mathbf{L}_i &\equiv \left.\frac{\partial\ln\mathbf{x}^o_i}{\partial\mathbf{x}^o_i}\right|_{\mathbf{x}^o_i=\mathbf{x}^o_{b,i}} \\ &= \begin{bmatrix}(\mathbf{x}^o_{b,i})_1 & 0 & \cdots & 0 \\ 0 & (\mathbf{x}^o_{b,i})_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & (\mathbf{x}^o_{b,i})_{m_i}\end{bmatrix}^{-1}.\end{aligned} \tag{3}$$

Equation (2) results from a Taylor expansion of the logarithm function for $\mathbf{x}^o_i$. In the simplest case when there is only one observation, Song et al. (2016) suggest that the observation y should satisfy

$$(1-\alpha)x^o_b < y < (1+\alpha)x^o_b \tag{4}$$

for the Taylor series approximation to be valid, where α is a constant between 0 and 1. In this experiment, we choose $\alpha = 0.5$ and discard observations outside the range in (4). This filtering reduces the number of observations that are used in the assimilation process, but helps to prevent the model from diverging due to violation of a linearity condition embedded in our formulation. It can also be thought of as a form of background quality control.

4. Model settings

Coupled PBDA for the year 2000 was performed with the NPZD model coupled to ROMS. ROMS is a 3D ocean circulation regional model with terrain following vertical coordinate (Haidvogel et al., 2000, Shcheptkin and McWilliams, 2004, Haidvogel et al., 2008). The model is configured for the CCS with 1/10° horizontal resolution and 42 vertical levels. This configuration has been widely used in other studies and proven to reproduce the mean CCS circulation as well as its seasonal variability (Veneziani et al., 2009,, Moore et al., 2011). The model also captures the circulation by mesoscale eddies with a length scale larger than 30 km, which imposes a greater challenge for coupled state estimation at the same time because of highly nonlinear features in the system. We note that this configuration has higher horizontal resolution than the one used in the companion papers for the model twin experiments (1/3° in Song et al. (2016,)).

The NPZD model has relatively simple dynamics linking the 4 components (nutrient, phytoplankton, zooplankton and detritus) (Powell et al., 2006). All components are budgeted in terms of nitrogen. Phytoplankton grows with nutrient uptake using Michaelis-Menten kinetics and is consumed by grazing and mortality. Zooplankton biomass increases by grazing phytoplankton (using an Ivlev formulation) and decreases through mortality. Concentrations of detritus increase through phytoplankton and zooplankton mortality as well as through unassimilated grazing. Remineralization reduces detrital concentrations, returning nitrogen to its inorganic form. Table 2 lists the parameter values tuned for the CCS region.

The NPZD model dynamics are critical for PBDA to determine the increment to the biological initial condition from the misfit between observations and model estimates during the assimilation cycle. For example, if the model prior has lower phytoplankton biomass than observed, the assimilation procedure has several mechanisms by which it can increase the modeled phytoplankton biomass. Model initial conditions can be adjusted to (a) increase phytoplankton concentrations directly, (b) increase nutrient levels, (c) decrease zooplankton biomass, or (d) because this is a fully coupled assimilation system, alter flux divergences of phytoplankton, nutrients or zooplankton via the velocity field resulting in the desired increase in phytoplankton at the observation point. In practice, a combination of all these mechanisms occurs, where the relative proportion of each is based on the model dynamics and the prescribed uncertainties in the observations and model variables.

The target year for the PBDA experiment is the year 2000 during which the CCS was close to the climatological norm despite La Niña conditions (Durazo et al., 2001). The initial condition for physical variables was prepared from the CCS 31-year historical reanalysis, CCSRA31 (Neveu et al., 2015), a product using the ROMS-4DVar procedure on the same model grid as this study. The initial condition for biological variables was derived from a 10-year spin-up of the coupled model. Physical boundary conditions and surface forcing were taken from SODA (Carton and Giese, 2008) and COAMPS (Hodur et al., 2002, Doyle et al., 2009), respectively. Biological boundary conditions for nutrients are the nitrate field extracted from World Ocean Atlas 2009 climatology (Garcia et al., 2006). Other variables are set to a small, constant value $C_0 = 0.1$ mmol N m^{-3}.

The background error covariance matrix $\mathbf{B}$ in (1) is a block matrix,

$$\mathbf{B} = \begin{bmatrix} \mathbf{B}_G & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_L \end{bmatrix}, \tag{5}$$

where $\mathbf{B}_G$ and $\mathbf{B}_L$ represent background error covariances for physical and log-transformed biological variables, respectively. $\mathbf{B}_G$ is adopted from the error covariance matrix used in the CCSRA31. $\mathbf{B}_L$ is estimated as $\mathbf{\Sigma C \Sigma}^T$ as in Broquet et al. (2009). The diagonal matrix of standard deviations, $\mathbf{\Sigma}$, is computed using the log-transformed biological variables from the 10-year spin-up run. The univariate correlation matrix $\mathbf{C}$ is constructed with the horizontal and vertical decorrelation length scale of 30 km and 7 m, respectively. We assume that observation errors are independent and uncorrelated, yielding an observational error covariance matrix $\mathbf{R}_i$ in (1) that is diagonal with error levels that appear in section 3.

We perform three experiments. The first simulation is a free run without any constraints and is referred to here as FREE. The second experiment, referred to as PDA, includes physical data assimilation only, and the model solution is constrained by physical observations alone. The third integration is called PBDA, assimilating both physical and biological observations in a fully coupled sense. Comparison between these three simulations highlights the impact of assimilating both physical and biological observations.

Each assimilation cycle spans 3 days. The chosen window length depends on the time scale for which the tangent linear approximation is valid. In the companion papers, the tangent linear approximation was found to be valid for at least 5 days (Song et al., 2016,). Here, at higher horizontal resolution and using realistic observations, we choose a more conservative

3-day window. Sensitivity experiments run with longer 4-day cycles achieved comparable performance. Even 8-day cycles produce quite acceptable results but are less favorable than 3- or 4-day cycles. Although the NPZD nonlinear model conserves total nitrogen during forward integration, data assimilation results in instantaneous adjustments to this quantity, as it does to physical variables such as heat content and momentum.

The local minimum of the quadratic cost function J in (1) is found using the Lanczos formulation (Fisher and Courtier, 1995, Tshimanga et al., 2008, Moore et al., 2011). After the local minimum is identified using 10 inner loops, the nonlinear coupled system is integrated forward with an updated initial condition to start another optimization cycle (2 outer loops). There is no model spin-up associated with each cycle, and no dynamical balance of biological variables is imposed on that initial condition. The final model trajectory is determined using the initial condition resulting from this second optimization cycle.

5. Evaluation

Table 3 summarizes the RMSEs for the three experiments with respect to assimilated observations (SSH, SST, $T_{in\ situ}$, $S_{in\ situ}$ and surface chlorophyll (SChl)). In addition, the biological ocean states from each experiment are evaluated against the independent observations of chlorophyll and NO_3 in the upper 200 m from GLOBEC-LTOP and CalCOFI in Table 4. The RMSEs for biological variables were computed without log-transformation.

5.1. Physical variables

For all physical variables, PDA yields the smallest RMSEs (Table 3). PBDA also yields small RMSEs, with values comparable to but slightly larger than those by PDA. This result differs from that of Song et al. (2016). In that idealized model twin experiment, the smallest RMSE of physical variables occurred using PBDA. In both Song et al. (2016) and this study, the quantitative differences between PDA and PBDA RMSEs were small relative to their improvement over the FREE run. We note also that as should be expected, the RMSE for physical variables can be reduced further in PBDA with more iterations, and thus at somewhat higher computational cost (not shown).

5.2. Biological variables

As in Song et al. (2016), PBDA results in the smallest RMSE for biological variables (Table 3). The RMSE for surface chlorophyll is reduced by 40% with respect to that of the FREE run. The observed annual mean surface chlorophyll and Hovmöller diagram in Figure 1 present at least three characteristics by which to evaluate assimilation performance: (1) high chlorophyll biomass in coastal regions with an initially sharp and then much more gradual decrease in the offshore direction (Figure 1(a,c)); (2) high chlorophyll biomass near the northern Washington coast (46°N-48°N) throughout the year (Figure 1(b,d)); and (3) episodic blooms of chlorophyll biomass along the central California coast (34°N-46°N) that appear throughout spring and summer, presumably responding to variable alongshore wind stress forcing (Figure 2).

5.2.1. Crosshore structure

To some degree, all three model experiments capture the sharp, then more gradual decrease of annual mean chlorophyll biomass in the offshore direction (Figure 3(a,c,e)). However, compared to observations, offshore concentrations of chlorophyll biomass are too low in the FREE run and too high in PDA (Figure 3(a,c)). In this experiment, high chlorophyll biomass in PDA must be driven by changes in physical properties alone, and two mechanisms have been identified by Raghukumar et al. (2015). Because updated initial conditions during each assimilation cycle are not required to be dynamically balanced, assimilation cycles exhibit initialization shocks in which gravity waves are released as part of their adjustment. These numerically-driven waves transiently increase nutrient concentrations in the euphotic zone, leading to increased primary production and in turn phytoplankton biomass. The second mechanism results from the update of subsurface physical temperature and/or salinity with no associated update to biological fields. Increased nutrient variance on isopycnal surfaces results also in increased primary production where density surfaces reach well-lit waters. Increased chlorophyll biomass is most noticeable in regions of very low concentration (i.e., offshore), though it is also visible in the coastal transition zone 100-200 km from shore. In our experiment, PDA resulted in the highest RMSE against the surface chlorophyll (Table 3). In contrast, the estimated offshore chlorophyll biomass in the PBDA experiment is comparable to observations in both magnitude and spatial distribution (Figure 3(e-f)). PBDA does not eliminate waves produced through initialization shock, and it does impose changes in stratification; however, PBDA also adjusts biogeochemical variables with the result that systematically higher chlorophyll concentrations than observed do not occur.

5.2.2. Northern U.S. west coast (44°N-48°N)

Along the U.S. west coast, like other eastern boundary upwelling systems, equatorward wind stress brings nutrient-rich subsurface water to the surface, supporting high chlorophyll biomass near coastal boundaries. Upwelling favorable wind stress is stronger along the central coast than the northern coast (Figure 2). As a result, the FREE run (Figure 3(a,b)) shows lower simulated chlorophyll biomass in the northern coastal region than along the central coast because wind-driven upwelling is the main driver for nutrient supply in the model. Indeed, wind-driven upwelling precedes high phytoplankton biomass by about 1 week along the northern U.S. west coast in the FREE run (Figure 4), offering support that the simulated chlorophyll biomass is mainly associated with the nutrient supply due to wind-driven upwelling. However, elevated levels of chlorophyll observed along the northern coast are not well explained by Ekman transport alone (Figure 1(d) and 2), suggesting that the current model configuration misses the key (either physical or biological) mechanisms in that region. Similarly, low phytoplankton levels in this northern coastal region have been noted in other forward modeling studies (e.g., Goebel et al., 2010).

Hickey and Banas (2008) suggested several mechanisms that support a highly productive north coast zone. Among them are a continuous nutrient supply from the Strait of Juan de Fuca, localized canyon enhanced upwelling, poleward coastally trapped wave and iron supply by the Columbia river. Recently, Davis et al. (2014) have shown that the first of these is a major factor.

Tidal mixing within the Strait of Juan de Fuca and Puget Sound results in a substantial nutrient flux to surface waters outside of the sound and ultimately along the Washington coast. The present model configuration includes neither Puget Sound nor tidal forcing, and it uses climatological nutrient boundary conditions along the northern boundary that are not especially elevated. This deficiency suggests an erroneous representation of the ecosystem in this region, including a systematically lower phytoplankton biomass. Since this issue is locally the result of low nutrients and remote physical process that occur outside of the model domain, it can not be improved by physical data assimilation. Indeed PDA (Figure 3(c,d)) results in a quantitatively different circulation in the region and an altered ecosystem response, including somewhat higher phytoplankton levels in spring, fall and winter, and lower levels in summer relative to the FREE run. But qualitatively, PDA is also deficient in the northern coastal region.

On the other hand, PBDA allows for chlorophyll observations to constrain the model, and thus it can improve on low prior estimates. In this system, model initial conditions are adjusted such that the misfit between model chlorophyll estimates and those observed is reduced. Modeled PBDA chlorophyll levels in the northern coastal region (Figure 3(e,f)) are mostly higher than in either FREE or PDA simulations, and these higher levels are relatively sustained through much of the year.

As mentioned above, PBDA can accomplish this adjustment through multiple mechanisms. The system can increase phytoplankton biomass directly, increase nutrients (NO_3) to drive primary production, and decrease zooplankton that grazes on phytoplankton. Figure 5 presents the PBDA increments to phytoplankton, nutrient and zooplankton prior estimates. It reveals that all three mechanisms occur in the northern coastal zone, which together elevate, phytoplankton biomass in this region relative to the FREE run. It is also possible that changes to physical properties alter transport and mixing and thus overall phytoplankton levels. Although Raghukumar et al. (2015) showed that adjustments to the physical circulation can improve spatial positioning of features (e.g., a higher correlation coefficient), we find that spatially and temporally averaged phytoplankton biomass is not substantially altered by the physical adjustments.

It is important to note that the magnitude of each increment is quantitatively determined by the prior model-data misfit, underlying model dynamics and prescribed values of observation and model uncertainty. The quantitative contribution of each increment can be assessed through analysis of source/sink terms in the phytoplankton budget (e.g., primary production, grazing, and mortality) that contribute to changes in phytoplankton biomass in the forward model. These changes can also be compared to the direct adjustment of phytoplankton itself, and we present this information averaged over the upwelling season (April to September) and nearshore 100 km in Figure 6 as a function of latitude. North of 44°N, direct phytoplankton adjustments overwhelmingly dominate, with lesser contributions by productivity (associated with changes to nutrients) and negligible contributions by grazing (associated with changes to zooplankton). We note that increases in phytoplankton biomass also result in non-negligible increases in mortality, which contribute negatively to phytoplankton concentrations during each assimilation cycle.

5.2.3. Central U.S. west coast (34°N-44°N)

Along the central U.S. west coast, the FREE simulation overestimates chlorophyll biomass from spring to early fall and underestimates it during other times of the year (Figure 1). With upwelling favorable wind stress starting in March (Figure 2), wind-driven phytoplankton blooms in the FREE run occur regularly in this region (Figure 3b). The chlorophyll biomass responds to wind-driven coastal upwelling after approximately 1 or 2 weeks (Figure 4). Assimilation of physical variables (PDA) does not change the simulated nearshore chlorophyll biomass significantly (Figure 3d); better agreement with the observation occurs in winter, but the timing and magnitude of phytoplankton blooms along the central coast disagree clearly with observed values. In contrast, PBDA successfully reduces the overall chlorophyll biomass modeled during the upwelling season and increases the biomass during fall and winter. Overall, PBDA alters the modeled structure of phytoplankton stock substantially, both in space and in time, matching observations considerably better than either FREE or PDA.

As in the northern region, multiple reasons may possibly exist for the discrepancies in the FREE run. One simple explanation for the higher chlorophyll biomass in the model is the suboptimal choice for parameter values. Another possible reason is the absence of iron limitation. In the central and northern CCS, high macronutrient levels with intermediate or low chlorophyll concentrations have been shown to result from iron limitation (Hutchins and Bruland, 1998, Hutchins et al., 1998, Bruland et al., 2001, Firme et al., 2003, Chase et al., 2007). Our simple NPZD model, with only one nutrient compartment arguably representing nitrate, neglects iron biochemistry altogether. With no potential for iron limitation, our model may overestimate phytoplankton growth, leading to higher chlorophyll biomass during the upwelling season. Following the upwelling season, California central and northern coast waters receive nutrients through riverine input (Chase et al., 2007), and nearshore chlorophyll biomass is generally above 0.1 mg m^{-3} (Figure 1(d))). Our model does not include riverine input, and this omission may reduce nutrient supply in wintertime and early spring relative to nature, resulting in lower levels of modeled chlorophyll biomass.

As discussed previously, changes in phytoplankton biomass can be influenced by changes in multiple state vector components. We find that along the central coast during the upwelling season, PBDA results in negative adjustments to phytoplankton and to nutrients, and more variable positive or negative increments to zooplankton depending on latitude and specific time-window (Figure 5). During fall and winter, changes to nutrients and phytoplankton are reversed, and changes to zooplankton become exceedingly small. As in the northern region, these increments are sensible considering impacts to grazing and uptake.

Unlike the northern zone where phytoplankton increments dominate changes to phytoplankton dynamics resulting from nutrient and zooplankton increments, the central coast region during the upwelling season exhibits changes in nearshore phytoplankton concentrations that are dominated by primary production (Figure 6). Because uptake depends on both phytoplankton and nutrient levels, both the negative phytoplankton increment and reduced nutrients resulting from PBDA contribute to lower uptake and lower phytoplankton biomass.

5.2.4. Budget changes

Increments in initial conditions cause total nitrogen within the model to be altered from cycle to cycle. As discussed, the coupled data assimilation system removes phytoplankton and nutrients along the central coast during the upwelling season (Figure 5). It is important to characterize the magnitude of these changes with respect to changes resulting from modeled biological dynamics. Here, we present terms in the budget for phytoplankton and nutrients pools averaged along the U.S. west coast during the upwelling season from April to September. The budget for phytoplankton, P, in the absence of data assimilation can be written

$$\frac{\partial P}{\partial t} + \nabla \cdot (\mathbf{u}P) \;\; = \;\; \nabla \cdot (\mathbf{K}\nabla P) + \text{Production} + \text{Grazing} + \text{Mortality}, \quad (6)$$

where the time-rate of change and advective flux divergences are balanced by diffusive flux divergences and biological sources and sinks. For phytoplankton, biological processes included in this model are phytoplankton production, grazing by zooplankton and phytoplankton mortality. A similar equation applies to the nutrient budget, but biological sources consist of unassimilated excretion and remineralization from detritus, phytoplankton and zooplankton, and the nutrient sink is uptake by phytoplankton. ROMS includes diagnostic tools to quantify each of these terms in a form consistent with the discretization, and the budget for the time and space average is shown in Figure 7. Time-mean increments for phytoplankton and nutrients, δP and δN are also shown.

In the phytoplankton budget, the largest term is biological production, and grazing has the next largest amplitude. The next most significant terms in order of their magnitude are phytoplankton mortality, the vertical diffusive flux, and the time-rate of change, ΔP. For comparison, we include the mean phytoplankton increment, denoted δP, produced by the assimilation, and observe that it is smaller than all the previously mentioned changes.

The nitrogen budget gives a similar impression. Although the assimilation-produced increment, δN is larger than the remineralization and excretion, it is smaller than the more dominant terms in the budget.

Finally, as a different measure of assimilation-induced budget changes, we calculate the time-mean of the absolute value of the ratio between the increment and the production in the phytoplankton budget and between the nutrient increment and uptake:

$$R_P = \overline{\frac{|\delta P|}{|\mathrm{Prod}|}} \tag{7}$$

$$R_N = \overline{\frac{|\delta N|}{|\mathrm{uptake}|}}. \tag{8}$$

We find $R_P = 6.4\%$ and $R_N = 7.4\%$ in our experiments. The assimilation procedure produces alterations to the state variables that are small compared to dominant biological processes in the respective biological budgets calculated by the NPZD model.

5.2.5. Subsurface, unassimilated data

Finally, we note that the RMSE computed using unassimilated data (chlorophyll and NO_3 in the upper 200 m) is also smallest in the PBDA experiment (Table 4). The error reduction in PBDA is quite small relative to FREE, but it shows potential for the assimilation system to spread some information vertically. Satellite estimates of chlorophyll can differ from in situ observations. Kahru et al. (2012) find that the chlorophyll estimation algorithm for SeaWiFS underestimates chlorophyll biomass concentrations in the California Current, indicating that assimilating SeaWiFS chlorophyll observation may not ensure a good fit to in situ observations. We find that the prior mean bias of chlorophyll for near-surface in situ observations is -0.37 mg m^{-3}, and it is reduced to -0.20 mg m^{-3} following assimilation. While this bias reduction near the surface is substantial, it is possible that the bias in satellite chlorophyll estimates limits the improvement in the posterior solution against in situ data.

As described in section 4, the background error covariance $\mathbf{B}$ contains the correlation matrix $\mathbf{C}$ whose vertical length scale is 7 m. Hence, most corrections for the initial biological conditions occur in the upper 20 m. Below that level, chlorophyll RMSEs are not very different between the three experiments. In deeper water, the background phytoplankton biomass is low, particularly below the euphotic zone. In contrast, the RMSEs of NO_3 differ from one another because of the different NO_3 fluxes associated with the ocean circulation (not shown). Overall, the reduction of RMSEs by PBDA is less than 13%, which is much smaller than the reduction for the surface chlorophyll. We note that while it is reassuring that subsurface changes are slightly improved by assimilation of surface information, subsurface RMSE would most likely benefit much more substantially from the availability of subsurface biogeochemical data.

6. Summary and Discussion

The theoretical development of the quadratic form for incremental, log-normal biogeochemical ocean data assimilation and the coupled physical and biogeochemical data assimilation (PBDA) approach are presented in companion papers (Song et al., 2016,), along with test cases using idealized model twin experiments. In this study, we applied the PBDA approach to a realistic problem by assimilating actual observations from the California Current System during the year 2000. PBDA was implemented using a simple four-component NPZD ecosystem model coupled to ROMS. Both physical observations from various platforms and SeaWiFS surface chlorophyll observations are used in PBDA to improve estimates of the physical and biological ocean states. We compared model results for a free run of the model, a run considering only physical data assimilation (PDA), and the PBDA solution.

Although PDA results in substantial improvements to the physical state, this procedure also yields ecosystem fields that on average are not improved over the free run. We find that PDA exhibits generally higher phytoplankton stock than the free run, consistent with results of Raghukumar et al. (2015) using a different biogeochemical model. In contrast, PBDA achieves dramatically smaller RMSEs for assimilated biological variables (in this case surface chlorophyll). PBDA also showed improvements in unassimilated subsurface biogeochemical data, but the reduction in RMSE was small compared to the free run (at most about 10-13%).

One intriguing result from Song et al. (2016) was that the lowest errors for physical observations resulted from PBDA and not PDA, suggesting that biological data can provide useful additional information to constrain physical fields. Here we find lower RMSE in PDA than PBDA, though the PBDA performance was only slightly worse than PDA relative to the improvement of both over the free run. In a model twin experiment of Song et al. (2016), the same model was used to produce observations and test the assimilation system. Thus in that configuration, the assimilation model is capable of reproducing the truth exactly. In a realistic configuration, as tested here, both physical and biological model components are inaccurate representations of nature for many reasons (e.g., model resolution, representation of subgridscale dynamics, parameterization of complex biological processes, specification of model and/or observational error statistics) and generally are not able to reproduce in a prognostic sense the natural environment exactly. As a result, we speculate that physical and biogeochemical model errors relative to nature are responsible for the slightly worse performance in terms of physical RMSE in this realistic configuration compared to the model twin experiment. Future studies will have to test whether improved models (physical, biological or both) could yield greater improvement in the physical variables through assimilation of biological information than through physical assimilation alone.

Examination of the temporal and spatial structure of the surface chlorophyll fields indicates that PBDA successfully adjusted the amplitude and timing of phytoplankton blooms in coastal waters to better match those observed. Such a result is to be expected if the assimilation system is functioning properly, but since this is the first demonstration of this technique using real data, we explored how the system achieved these changes. The assimilation model is free to adjust all elements of the control vector (in this case, model initial conditions at the start of each assimilation cycle) and the magnitude and relative proportion of those changes result from a combination of model dynamics, embodied by the nonlinear, tangent linear, and adjoint models, as well as prescribed observation and model uncertainties.

In regions where the free solution underestimated chlorophyll systematically (such as along the Pacific Northwest coast), the assimilation system adjusted phytoplankton, nutrient, and zooplankton levels such that each increment would contribute to an increase in phytoplankton stocks within the nonlinear model. We found that during the upwelling season, increments to the phytoplankton state variable contributed the most to the total change in phytoplankton concentrations. In other regions (e.g., along the central and northern California coast), the free solution overestimated chlorophyll levels. Here, we found that a reduction in phytoplankton growth, resulting from reductions in both phytoplankton stocks and nutrient levels by PBDA, was responsible for the largest decrease in phytoplankton concentrations.

We also noted several deficiencies of the unconstrained model that potentially limit agreement between the free run and observations. The model, for example, does not include high nitrate levels near the northern boundary that have been shown to result from tidal mixing within the Strait of Juan de Fuca (Davis et al., 2014). It is possible, even likely, that different, or spatially varying parameters for the NPZD model, different surface forcing or boundary conditions, or alternate biogeochemical or physical models altogether, may produce in a non-assimilative run ecosystem fields with greater fidelity than the one used in this study. However, although an alternate model may produce fields that are closer to the data available (and substantial effort to improve forward model calculations should be made), differences between observations and models are unavoidable. This study demonstrates that errors present in unconstrained model calculations can be adjusted sensibly through rigorous 4-dimensional data assimilation.

The improved spatial structure of surface chlorophyll produced by PBDA over the free run identifies a possible application of these model results. As is well known (and shown in Figure 1b), the coastal ocean undergoes frequent cloud cover that prevents direct satellite assessment of surface chlorophyll. We found that in the coastal strip defined here as the nearshore 100 km and delineated approximately by the blue line in Figure 1a, daily satellite surface chlorophyll estimates were unavailable about 70% of the year. More complete coverage can be attained by using temporal composite estimates (such as the 8-day composite shown in Figure 1d). However, such composites necessarily trade high frequency variability for temporal coverage. In contrast, the assimilative model produces a complete 4-dimensional estimate of the ocean state, regardless of cloud cover. It is data constrained during periods when observations are available, but uses model dynamics to extend assimilated fields through periods of low data availability. Hence, assimilative models can be thought of as sensible dynamical interpolators of sparse data.

Sensitivity studies (not shown) revealed that the assimilation system is quite robust, whereby small variations to a variety of assimilation-related parameters, such as assimilation window length and background error variances, did not substantively change ocean state estimates. We did use a smaller vertical decorrelation scale for phytoplankton (7 m) than for physical variables (30 m) because a larger vertical decorrelation scale in combination with the logarithm transform resulted in unrealistic enhancement of sub- surface mixed layer phytoplankton fields. The proposed data assimilation method applied to a completely different biogeochemical model, NEMURO, is also able to fit the satellite observations of surface chlorophyll well (Mattern et al., in prep). NEMURO includes phytoplankton and zooplankton community structure, and thus is arguably better suited to represent the different nutrient zones of the California Current System than is the presently applied NPZD model with single parameter values across the full domain.

Computational requirements for PBDA are increased over PDA by about the same factor as running a coupled biogeochemical model over only physics in a forward (nonlinear) run. In practice, 4-dimensional variational assimilation costs O(100) times the forward (nonlinear) model calculations because multiple iterations of tangent linear and adjoint models are required to approach the cost function minimum. The added cost of PBDA over PDA is the cost of running the biological tangent and adjoint models. For a 4-component NPZD model, coupled calculations require approximately twice the memory and processor time as a pure physics run. It is worth noting that ensemble Kalman Filter calculations are similarly more expensive than forward model calculations owing to the multiple runs of the forward model required to estimate the background covariance matrix (Edwards et al., 2015).

As mentioned, the 4DVar approach uses model dynamics, embodied in the tangent linear and adjoint models, to connect observations within each assimilation cycle to the model initial conditions, and the magnitudes of the initial condition increments are dependent on prescribed observation and model error statistics. In this study, we assumed univariate model errors, where the background error covariance matrix consists of variances on diagonal elements, and off-diagonal components are determined by the solution of a diffusion equation. Thus connections between SSH and velocity or SSH and phytoplankton, for example, are only attained through model dynamics. SSH of course can be related dynamically to phytoplankton concentrations through alterations of near surface velocity. Dynamics in this context distinguishes the 4DVar method from sequential methods, which rely purely on statistics to distribute observational information both locally and nonlocally. Multi-variate statistical approaches in sequential methods for coupled biogeochemical assimilation problems are beginning to emerge and have shown great promise (e.g., Shulman et al., 2013). Such developments suggest that 4DVar solutions can be further improved through alterations of background error covariances. For example, it may be possible to statistically relate local changes in density below the euphotic zone to changes in nutrients that should improve the nitrate density relationship relative to observations.

In addition, it has been demonstrated in physical systems that background error covariances can be partitioned into balanced and unbalanced parts (Derber and Bouttier, 1999, Weaver et al., 2005, Moore et al., 2011). This decomposition assumes that variables in the unbalanced part are uncorrelated. It is not clear how the imposition of a balance operator relating physical variables may impact coupled biogeochemical data assimilation. Furthermore, it may be possible to find analogous first order balanced relationships in coupled dynamical problems that could be included in physical and biogeochemical data assimilation. Investigating multivariate statistical relationships and sensitivities to balance operators remain subjects for future studies.

This investigation evaluates a fully coupled, physical and biogeochemical 4-dimensional variational data assimilation system in a realistic configuration of the U.S. west coast at 1/10 degree resolution and spanning a 1-year duration. We assimilate widely available physical and biological observations, and substantially reduce errors in a biological variable over an unconstrained model and a model that assimilates only physical observations. The approach is model independent, although the coding of the tangent linear and adjoint models is challenging and model-specific. While several improvements can be made to both the forward models and assimilative procedures to further improve estimates, this study demonstrates that implementation of 4DVar in this context is practical and potentially useful. Such methods should be of interest for historical reanalyses and regional ocean observing systems quite generally.

7. Acknowledgement

We are grateful for support from the Gordon and Betty Moore Foundation and from the National Oceanographic and Atmospheric Administration (NA10OAR4320156) that enabled this research. SST data was obtained from http://las.pfeg.noaa.gov/oceanWatch/oceanwatch.php and surface chlorophyll data from http://oceandata.sci.gsfc.nasa.gov/SeaWiFS/Mapped/Daily/9km/. Three anonymous reviewers significantly improved the manuscript with several helpful comments.

8. References

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., Rue, E., Smith, G., 2001. Iron and macronutrients in California coastal upwelling regimes: Implications for diatom blooms. Limnol. Oceanogr. 46, 1661–1674.

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

Campbell, J. W., Blaisdell, J. M., Darzi, M., 1995. Level-3 SeaWiFS data products: Spatial and temporal binning algorithms. In: Hooker, S., Firestone, E., Acker, J. (Eds.), NASA Tech. Memo. 104566. Vol. 32. NASA Goddard Space Flight Center, Greenbelt, Maryland.

Carr, M.-E., 2002. Estimation of potential productivity in Eastern Boundary Currents using remote sensing. Deep-Sea Res. Pt. II 49, 59–80.

Carton, J., Giese, B., 2008. A reanalysis of ocean climate using Simple Ocean Data Assimilation (SODA). Mon. Wea. Rev. 136, 2999–3017.

Chase, Z., Strutton, P. G., Hales, B., 2007. Iron links river runoff and shelf width to phytoplankton biomass along the U.S. West Coast. Geophys. Res. Lett. 34, L04607.

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.

Derber, J., Bouttier, F., 1999. A reformulation of the background error covariance in the ECMWF global data assimilation system. Tellus 51A, 195–221.

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.

Durazo, R., Baumgartner, T. R., Bograd, S. J., Collins, C. A., Campa, S. D. L., Garcia, J., Gaxiola-Castro, G., Huyer, A., Hyrenbach, K. D., Loya, D., Lynn, R. J., Schwing, F. B., Smith, R. L., Sydeman, W. J., Wheeler, P., 2001. The state of the california current, 20002001: a third straight la nina year. Calif. Coop. Ocean. Fish. Invest. Rep. 42, 29–60.

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.

Fiechter, J., Curchitser, E. N., Edwards, C. A., Chai, F., Goebel, N. L., Chavez, F. P., 2014. Air-sea CO2 fluxes in the California Current: Impacts of model resolution and coastal topography. Global Biogeochem. Cy 28, 371?385.

Firme, G., Rue, E., Weeks, D., Bruland, K., Hutchins, D., 2003. Spatial and temporal variability in phytoplankton iron limitation along the California coast and consequences for Si, N, and C biogeochemistry. Global Biogeochem. Cycles 17, 1016.

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.

Garcia, H. E., Locarnini, R. A., Boyer, T. P., Antonov, J. I., 2006. Volume 4: Nutrients (phosphate, nitrate, silicate). In: Levitus, S. (Ed.), World Ocean Atlas 2005. Vol. 64 of NOAA Atlas NESDIS. U.S. Government Printing Office, Washington, D.C., p. 396.

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, 221 – 241.

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

Haidvogel, D., Arango, H., Hedstrom, K., Beckmann, A., Malanotte-Rizzoli, P., Shchepetkin, A., 2000. Model evaluation experiments in the North Atlantic basin: Simulations in nonlinear terrain-following coordinates. Dynam. Atmos. Oceans 32, 239–281.

Haidvogel, D., H.G.Arango, Budgell, W., Cornuelle, B., Curchitser, E., Di Lorenzo, E., Fennel, K., Geyer, W., Hermann, A., Lanerolle, L., Levin, J., McWilliams, J., Miller, A., Moore, A., Powell, T., Shchepetkin, A., Sherwood, C., Signell, R., Warner, J., Wilkin, J., 2008. Ocean forecasting in terrain-following coordinates: Formulation and skill assessment of the Regional Ocean Modeling System. J. Comput. Phys. 227, 35953624.

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

Hodur, R. M., Pullen, J., Cummings, J., Hong, X., Doyle, J., Martin, P., Rennick, M., 2002. The coupled ocean/atmosphere mesoscale prediction system (COAMPS). Oceanography 15, 8898.

Hutchins, D., Bruland, K., 1998. Iron-limited diatom growth and Si:N uptake ratios in a coastal upwelling regime. Nature 393, 561–564.

Hutchins, D., DiTullio, G., Zhang, Y., Bruland, K., 1998. An iron limitation mosaic in the California upwelling regime. Limnol. Oceanogr. 43.

Ingleby, B., Huddleston, M., 2007. Quality control of ocean temperature and salinity profiles - Historical and real time data. J. Mar. Syst. 65, 158–175.

Kahru, M., Kudela, R. M., Manzano-Sarabia, M., Greg Mitchell, B., 2012.

Trends in the surface chlorophyll of the California Current: Merging data from multiple ocean color satellites. Deep-Sea Res. Pt. II 77–80, 89–98.

Kilpatrick, K. A., Podesta, G. P., Evans, R., 2001. Overview of the NOAA/NASA Advanced Very High Resolution Radiometer Pathfinder algorithm for sea surface temperature and associated matchup database. J. Geophys. Res.-Oceans 106, 9179–9197.

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.

Neveu, E., Moore, A. M., Edwards, C. A., Fiechter, J., Drake, P., Jacox, M. G., Nuss, E., 2015. An historical analysis of the California Current Circulation using ROMS 4D-Var. Part I: System configuration and diagnostics. Ocean Modell., submitted.

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.

Raghukumar, K., Edwards, C. A., Goebel, N. L., Broquet, G., Veneziani, M., Moore, A. M., Zehr, J. P., 2015. Impact of assimilating physical oceanographic data on modeled ecosystem dynamics in the California Current System. Prog. Oceanogr. 138 (0), 546–558.

Rio, M.-H., Schaeffer, P., Lemoine, J.-M., Hernandez, F., 2004. Estimation of the ocean mean dynamic topography through the combination of altimetric data, in situ measurements, and GRACE geoid: from global to regional studies. In: GOCINA International Workshop. Luxembourg.

Shcheptkin, A. F., McWilliams, J. C., 2004. The Regional Oeanic Modeling System (ROMS): A split explicit, free-surface, topography-following-coordinate oceanic model. Ocean Modell. 9, 347–404.

Shulman, I., Frolov, S., Anderson, S., Penta, B., Gould, R., Sakalaukus, P., Ladner, S., 2013. Impact of bio-optical data assimilation on short-term coupled physical, bio-optical model predictions. Journal of Geophysical Research: Oceans 118 (4), 2215–2230.
URL http://dx.doi.org/10.1002/jgrc.20177

Song, H., Edwards, C. A., Moore, A. M., Fiechter, J., 2012. Incremental 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., 2016a. 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. Ocean Modell. in press.

Song, H., Edwards, C. A., Moore, A. M., Fiechter, J., 2016b. Data assimilation in a coupled physical-biogeochemical model of the California Current System using an incremental lognormal 4-dimensional variational approach: Part 2, Joint physical and biological data assimilation twin experiments. Ocean Modell. submitted.

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.

Veneziani, M., Edwards, C. A., Doyle, J. D., Foley, D., 2009a. A central California coastal ocean modeling study: 1. Forward model and the influence of realistic versus climatological forcing. J. Geophys. Res. 114, C04015.

Veneziani, M., Edwards, C. A., Moore, A. M., 2009b. A central california coastal ocean modeling study: 2. adjoint sensitivities to local and remote forcing mechanisms. J. Geophys. Res. 114, C04020.

Weaver, A. T., Deltel, C., Machu, E., Ricci, S., Daget, N., 2005. A multivariate balance operator for variational ocean data assimilation. Quart. J. Roy. Meteorol. Soc. 131, 3605–3625.

Table 1: Observations used in physical and biological coupled data assimilation. Their sources, frequencies and numbers are listed.

Variable	Source	Frequency	number
Sea surface height	AVISO gridded product	7-day	772,856
Sea surface temperature	AVHRR/PathFinder	Daily	3,026,628
Sea surface chlorophyll	SeaWiFS	Daily	1,029,735
in situ T	EN3 (Ingleby and Huddleston, 2007)		24,526
in situ S	EN3 (Ingleby and Huddleston, 2007)		9,669

Table 2: Parameter names, values and units for the NPZD model

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

Table 3: The mean RMSEs for SSH, SST, $T_{in\ situ}$, $S_{in\ situ}$ and surface chlorophyll (SChl) are computed using assimilated observations. The chlorophyll RMSE was computed without logarithm transformation.

	SSH (cm)	SST (°C)	$T_{in\ situ}$ (°C)	$S_{in\ situ}$ (psu)	SChl (mg m^{-3})
Free	9.26	1.11	1.31	0.29	0.74
PDA	**3.16**	**0.58**	**0.82**	**0.17**	0.78
PBDA	3.94	0.59	0.89	0.20	**0.45**

Table 4: The mean RMSEs for subsurface chlorophyll (Chl) and NO_3 are also computed using the unassimilated in situ observations from the GLOBEC-LTOP and CalCOFI stations as marked in black and blue in Figure 1a, respectively.

	GLOBEC-LTOP		CalCOFI	
	Chl (mg m^{-3})	NO_3 (mmol N m^{-3})	Chl (mg m^{-3})	NO_3 (mmol N m^{-3})
Free	1.41	5.61	0.71	4.01
PDA	1.44	5.43	0.70	4.10
PBDA	**1.39**	**4.88**	**0.65**	**3.96**

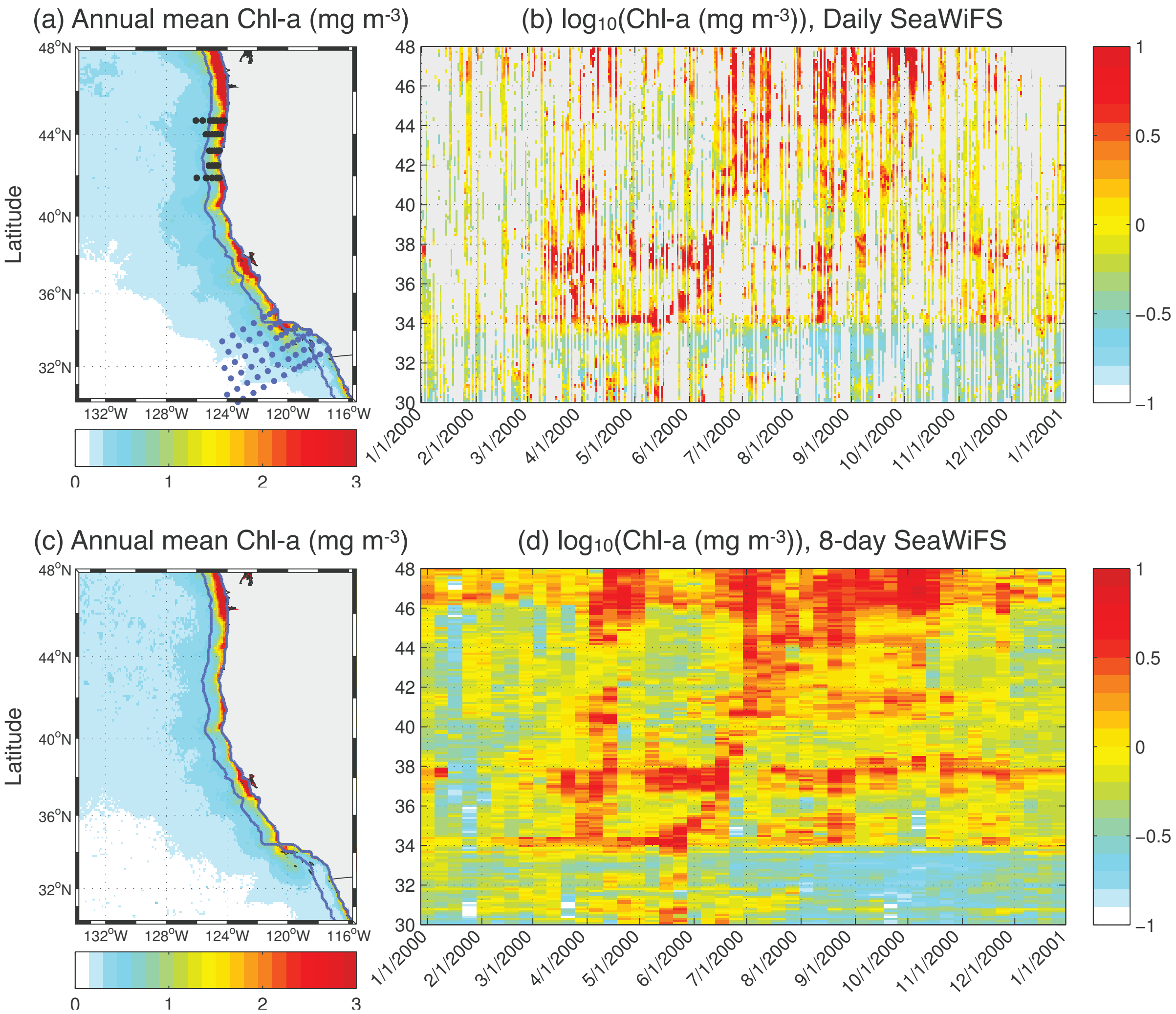

Figure 1: Annual mean surface chlorophyll (left) and Hovmöller diagrams of log10-transformed surface chlorophyll at the coast (right). (a,b) and (c,d) represent the daily and 8-day composite SeaWiFS chlorophyll data products, respectively. Surface chlorophyll within the blue contours (roughly 100 km wide) on the left column plots are averaged for the Hovmöller diagrams on the right column. Black and blue dots in (a) represent the GLOBEC-LTOP and CalCOFI stations, respectively.

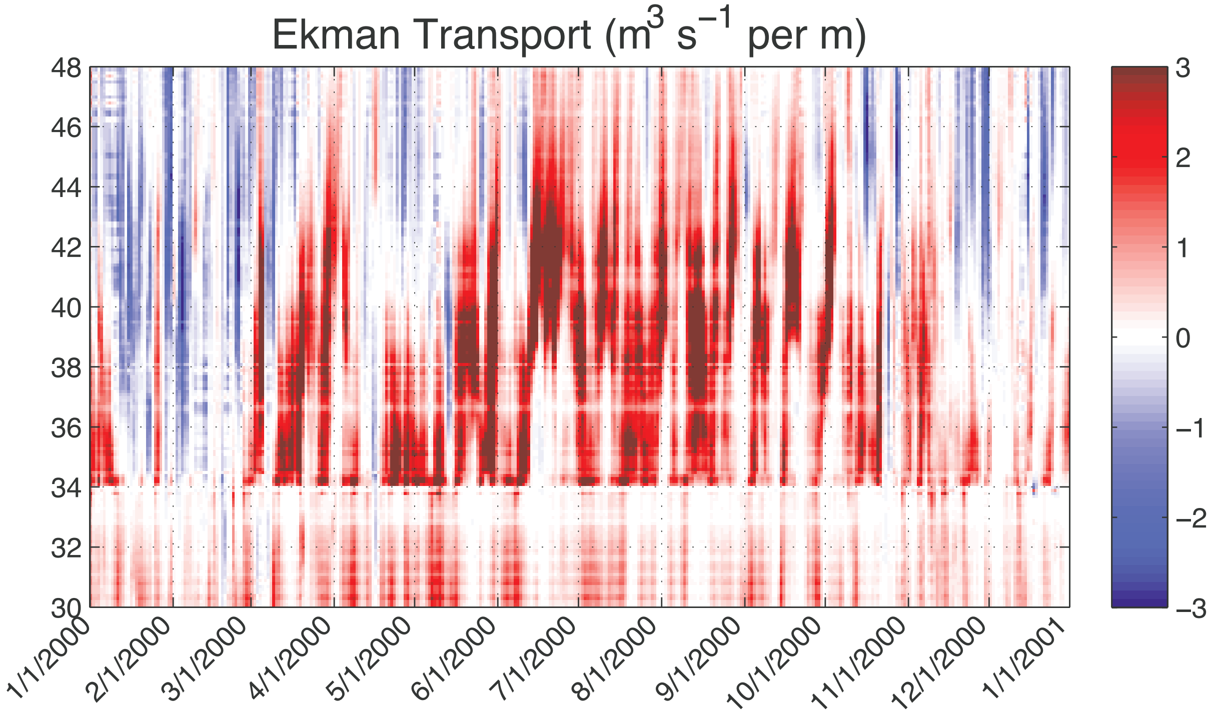

Figure 2: Hovmöller diagram of zonally averaged surface Ekman transport within 100 km of the coast (blue contour in Fig 1(a)).

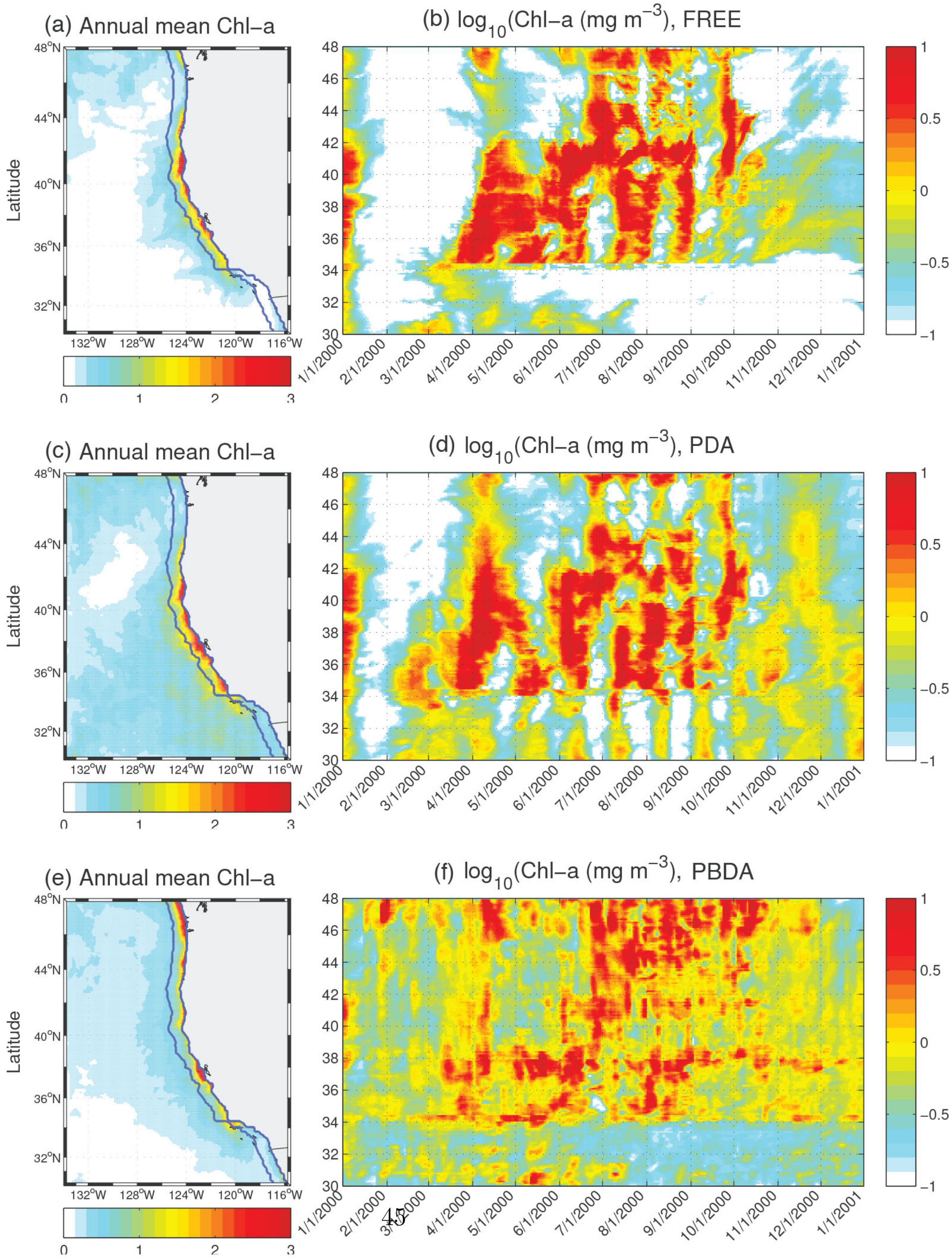

Figure 3: Same as Fig 1 with the data from (a, b) free forward simulation, (c, d) PDA and (e, f) PBDA state estimation.

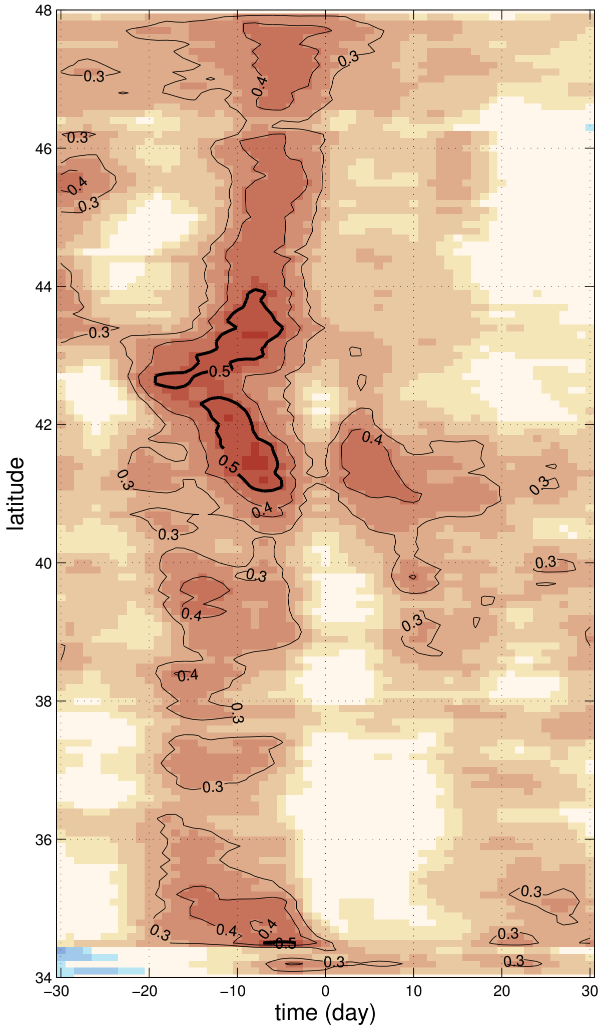

Figure 4: Lagged cross correlation between the Ekman transport (Figure 2) and phytoplankton biomass for the FREE run (Figure 3(b)). Negative time means that the Ekman transport precedes the growth of phytoplankton biomass.

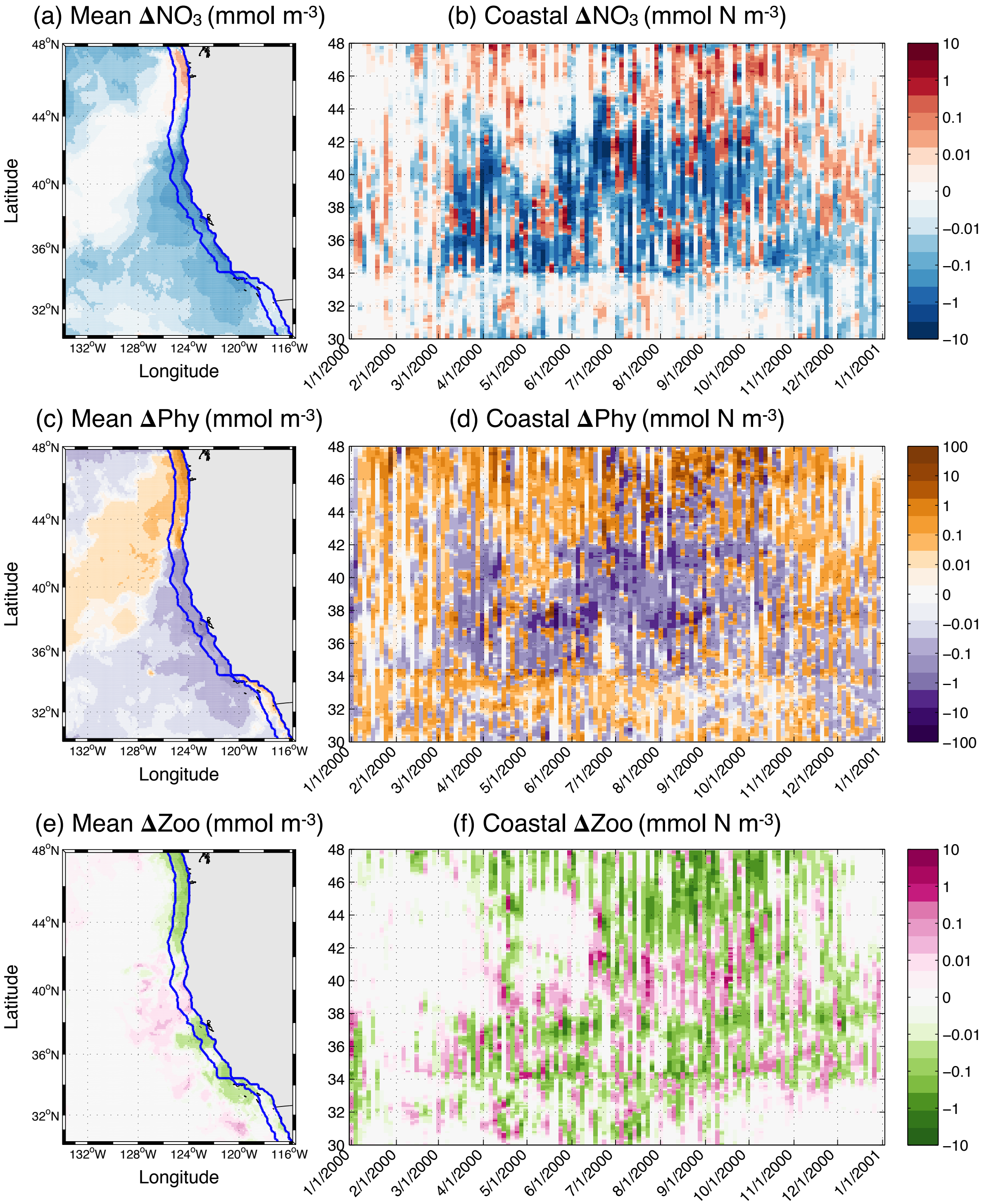

Figure 5: Spatial map and Hovmöller diagrams of initial increments by PBDA for (a,b) nitrate, (c,d) phytoplankton and (e,f) zooplankton. The increments are averaged in time (left column) or in space at the coastal regions (right column). It is noted that the scale for phytoplankton in (c,d) is greater than the other two variables.

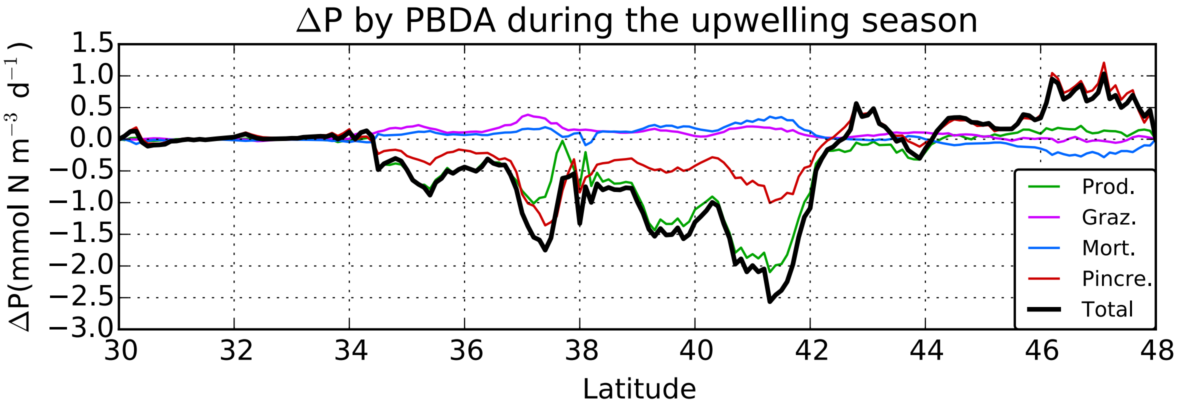

Figure 6: Changes in phytoplankton by assimilating physical and biological observations averaged over the upwelling season (from April to September). Total changes (black) are partitioned by the production (green), grazing (purple), mortality (blue) and increment in phytoplankton biomass (red).

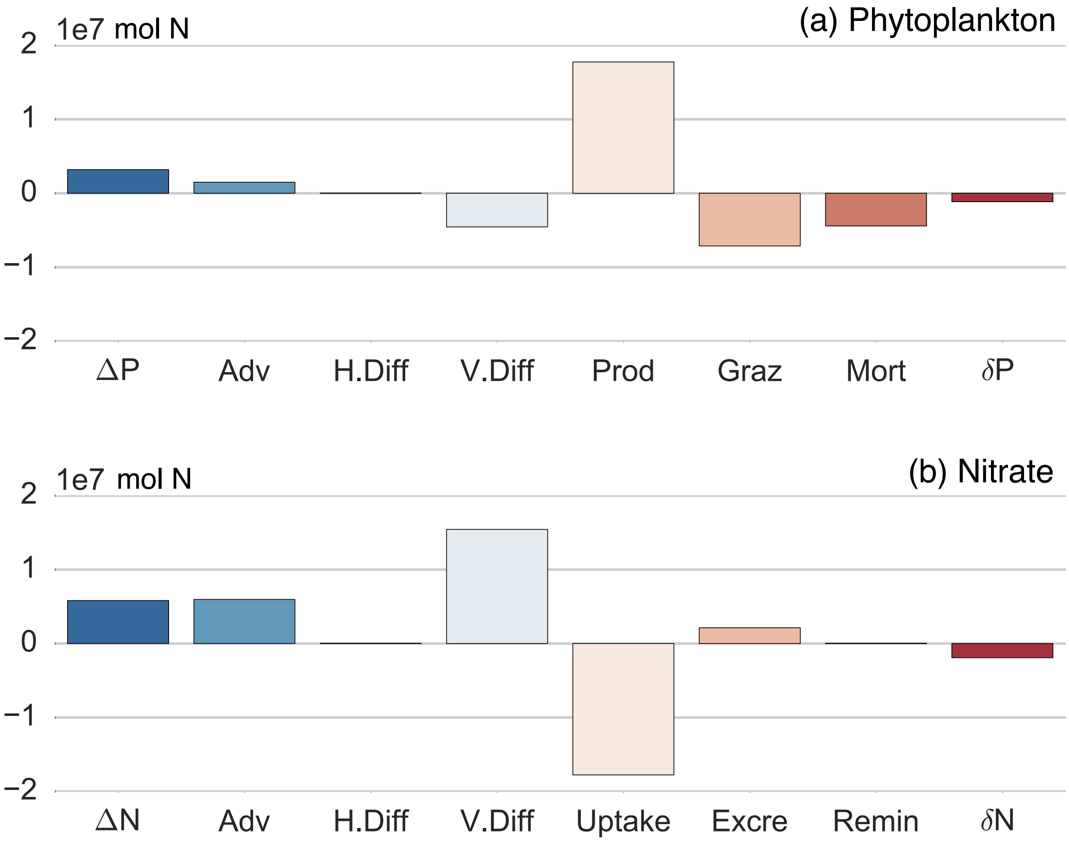

Figure 7: Terms in the phytoplankton (a) and nutrient (b) budgets averaged over the coastal area during the upwelling season (from April to September) from the posterior solution (after data assimilation). The time-rate of change in phytoplankton (ΔP) is partitioned into an advective flux divergence (Adv), horizontal diffusive (H.Diff) and vertical (V.Diff) flux divergence, primary productivity (Prod), grazing (Graz) and mortality (Mort). The phytoplankton increment produced by assimilation is denoted δP. In the NO_3 budget, biological sources consist of remineralization (Remin) and addition by excretion (Excre). Uptake is the sink, and the NO_3 increment is labeled δN.