Version of Record: https://www.sciencedirect.com/science/article/pii/S1463500316300919
Manuscript_bc95bcd01ec2b6d706d337550f5c22fe

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

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

Coupled physical and biological data assimilation is performed within the California Current System using model twin experiments. The initial condition of physical and biological variables is estimated using the four-dimensional variational (4DVar) method under the Gaussian and lognormal error distributions assumption, respectively. Errors are assumed to be independent, yet variables are coupled by assimilation through model dynamics. Using a nutrient-phytoplankton-zooplankton-detritus (NPZD) model coupled to an ocean circulation model (the Regional Ocean Modeling System, ROMS), the coupled data assimilation procedure is evaluated by comparing results to experiments with no assimilation and with assimilation of physical data and

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

© 2016 published by Elsevier. This manuscript is made available under the Elsevier user license
https://www.elsevier.com/open-access/userlicense/1.0/

biological data separately. Independent assimilation of physical (biological) data reduces the root-mean-squared error (RMSE) of physical (biological) state variables by more than 56% (43%). However, the improvement in biological (physical) state variables is less than 7% (13%). In contrast, coupled data assimilation improves both physical and biological components by 57% and 49%, respectively. Coupled data assimilation shows robust performance with varied observational errors, resulting in significantly smaller RMSEs compared to the free run. It still produces the estimation of observed variables better than that from the free run even with the physical and biological model error, but leads to higher RMSEs for unobserved variables. A series of twin experiments illustrates that coupled physical and biological 4DVar assimilation is computationally efficient and practical, capable of providing the reliable estimation of the coupled system with the same and ready to be examined in a realistic configuration.

Keywords: Coupled data assimilation, Biogeochemical model, 4DVAR, California Current System Keyword

1. Introduction

Marine ecosystem and biogeochemical models coupled to realistic ocean circulation models are applied routinely today for an extensive range of studies, such as primary production (Franks and Chen, 2001), ecosystem phenology (Chenillat et al., 2013), biogeography (Follows et al., 2007, Goebel et al., 2010), nutrient cycling (Fennel, 2010), air-sea carbon exchange (Chai et al., 2009) and climate change (Cox et al., 2000, Behrenfeld et al., 2006, Stock et al., 2011). Despite real advances in the representation of complex biolog-

ical interactions and improvements in physical circulation modeling, many factors still contribute to errors in model output, including imperfect parameterization of biological and physical processes at both resolved and subgrid scales. In realistic applications in which ocean state estimates at particular times are sought, for example as part of an ocean observing system, additional errors result from uncertainties in ocean model initial conditions and applied forcing.

One approach to improving model fidelity for ocean state estimation is through data assimilation in which model estimates are constrained through adjustment of control variables to better match available observations. Developments in data assimilation in physical oceanography for over two decades now provide many estimates of the physical ocean state at global (Behringer et al., 1998, Bell et al., 2000, Chassignet et al., 2007, Köhl et al., 2007, Balmaseda et al., 2008, Carton and Giese, 2008) and regional (Oke et al., 2008, Cummings et al., 2009, Broquet et al., 2009, Shulman et al., 2009, Kurapov et al., 2011, Matthews et al., 2012, Sakov et al., 2012) scales. Data assimilation techniques have been developed to a lesser extent in biological oceanography, though their application has been used for both the determination of poorly known model parameters (e.g., Matear, 1995, Spitz et al., 1998, Losa et al., 2004, Mattern et al., 2012, Roy et al., 2012, Doron et al., 2013, Simon et al., 2015) as well as quantitative improvement in modeled biological fields (see Gregg (2008) and Edwards et al. (2015) for reviews).

Data assimilation in physical-biological coupled systems has focused generally on either physical or biological data assimilation in isolation. Better representation of the ocean circulation by physical data assimilation is ex-

pected to improve the distribution of biological variables. Studies focusing on the impact of physical data assimilation on biological fields include Oschlies and Garçon (1998), Miller et al. (2000), Berline et al. (2007) and Fiechter et al. (2011). For example, Oschlies and Garçon (1998) assimilate satellite estimates of sea surface height (SSH) over the North Atlantic to improve the eddy representation in the physical model, which provides currents for a coupled biological model. They report that the nitrate flux into the euphotic zone is increased by improving the underestimated mesoscale eddy activity in the free simulation. Raghukumar et al. (2015) present a counter example in which physical data assimilation alone can drive spurious nutrient fluxes into the euphotic zone degrading ecosystem model performance. Improvements in biological fields have also resulted from assimilation of biological fields alone, where physical fields have been assumed a priori to be sufficiently accurate or already modified through physical data assimilation (e.g., Friedrichs, 2001, Garcia-Gorriz et al., 2003, Natvik and Evensen, 2003, Hoteit et al., 2005, Triantafyllou et al., 2007, Ciavatta et al., 2011, Rousseaux and Gregg, 2012, Ford et al., 2012, Hu et al., 2012). For example, Garcia-Gorriz et al. (2003) assimilate the Sea-viewing Wide Field-of-view Sensor (SeaWiFS) data into an NPZ model coupled to a 3D ocean model over the Adriatic Sea, improving ecosystem parameters to reduce misfits between observations and the model output. Natvik and Evensen (2003) use SeaWiFS data to fit an 11-component biochemical model coupled to a 3D ocean circulation model configured over the North Atlantic. Using an ensemble Kalman filter approach, they adjust state variables and show that multivariate biochemical data assimilation can not only improve the representation of an observed variable but also reduce

the variance of unobserved variables.

A few studies have addressed the assimilation of both physical and biological data into coupled models and its advantages over the fit to either physical or biological data alone. Using an optimal interpolation approach applied to the Gulf Stream region, Anderson et al. (2000) find that dynamically consistent physical and biological fields created through joint physical and biological assimilation are superior to those obtained through either physical or biological data assimilation alone. In a sequential data assimilation study of the North Atlantic, Ourmières et al. (2009) find that ecosystem state estimates are improved through assimilation of both physical (sea surface temperature (SST), SSH and climatological temperature (T) and salinity (S)) and biological (nitrate climatology) data more than through physical data assimilation alone. Indeed in their results, physical data assimilation alone may degrade ecosystem estimates, depending on the accuracy of the nitrate background state. In a study of Monterey Bay, Shulman et al. (2013) report that substantial improvement in biological estimates did not result from physical assimilation alone but required assimilation of biological fields or updates to biological fields (nitrate) through statistical relations. Simon et al. (2015) perform two steps of assimilation using an ensemble Kalman filter in coupled physical and biological state estimation. After fitting the model to SST, along-track sea level anomalies and ice concentration observations, they assimilate 8-day composite chlorophyll data to estimate the biological states and parameters in log-transformed space. A stronger error reduction results from assimilating all observations, but physical and biological components are independent during assimilation.

To date biological assimilation efforts on adjusting state variables have largely used sequential methods, based on optimal interpolation or Kalman filter approaches. In particular, Bertino et al. (2003) applied Gaussian anamorphosis to biogeochemical variables with the non-Gaussian distributions in the ensemble Kalman filter. This transformation satisfies the assumption of Gaussian error distribution and they show promising results in fitting 1-D numerical ecosystem model to observations. Recently, a four-dimensional variational assimilation method appropriate for ocean ecosystem variables was studied in an idealized 1-dimensional context (Song et al., 2012). This method accounts for the non-Gaussian statistics of ecosystem variables by assuming lognormal statistics following Fletcher and Zupanski (2006). In Song et al. (2016a), this approach is modified and implemented within a realistic ocean circulation model (ROMS; the Regional Ocean Modeling System). The modification includes a linearization of the log-transformation function to enable efficient searching for the cost function minimum. Although the linearization requires the exclusion of observations whose values substantially exceed the background state, the modified log-transformed 4DVar outperforms the conventional 4DVar (which assumes Gaussian error distributions) both in terms of RMSE and non-negativity in a series of model twin experiments configured for the California Current System (Song et al., 2016a).

In this study, we extend that work by developing the ability to jointly assimilate physical and biological data within ROMS. The four-dimensional variational (4DVar) method provides dynamically consistent state estimates within each assimilation cycle. Coupled dynamics within the tangent linear and adjoint models of the 4DVar system have the potential to enable both

physical data to improve biological fields and biological data to improve physical fields. In most sequential data assimilation methods the latter is possible only through statistical adjustments and not directly through model dynamics. We investigate the advantage of coupled data assimilation using both physical and biological data by comparing results from multiple runs: no data assimilation, physical data assimilation alone, biological data assimilation alone, and joint physical and biological assimilation. Using model twin experiments, we fit the coupled model to pseudo observations of SST, SSH, in situ temperature and salinity, and surface chlorophyll data. The statistical analyses highlight the advantage of the coupled data assimilation in the present model configuration.

The organization of this paper is as follows. A brief introduction to physical and biological variational data assimilation method is given in section 2. Section 3 describes the twin experiment design to evaluate the coupled data assimilation system and its performance is presented in section 4. We summarize results and provide a discussion in section 5 to end the paper.

2. Coupled variational data assimilation

It is desirable to perform physical and biological data assimilation simultaneously rather than independently because (a) it is computationally more efficient to carry out a single assimilation procedure with a larger model than to perform two sequential but smaller assimilation operations and (b) the physical and biological fields are coupled through model dynamics. This coupling enables biological observations to constrain physical fields and vice versa, in principle leading to improved state estimates over results from in-

dependent (i.e., uncoupled) physical and biological assimilation. In practice, however, the beneficial results of coupled assimilation must be demonstrated, and issues such as model error and observational uncertainty may limit the ultimate improvements obtained through coupled state estimation.

Fundamentally, the basic error assumptions for physical and biological data assimilation are different. Errors in physical variables are assumed to be Gaussian distributed, whereas errors in biological variables are better represented by lognormal distributions (Campbell, 1995, Simon and Bertino, 2009); as a result, the assumption of Gaussian-distributed errors is appropriate for physical but not biological variables. Here, we present a method for combining the two different 4DVar approaches by following Fletcher and Zupanski (2006), Fletcher (2010) and Fletcher and Jones (2014). Although the state vector could include initial conditions, boundary conditions, surface forcing fields, and biological parameters, we consider here for simplicity a control vector consisting only of the initial state $\mathbf{x}_0$.

We first adapt the incremental form of 4DVar following Song et al. (2016a). If the posterior initial condition $\mathbf{x}_0$ is written as the sum of the background state $\mathbf{x}_{b,0}$ and a (small) increment $\delta\mathbf{x}_0$, then $\mathbf{x}_i^o = \mathcal{H}_i(\mathcal{M}_{i,0}(\mathbf{x}_{b,0} + \delta\mathbf{x}_0)) \approx \mathcal{H}_i(\mathcal{M}_{i,0}(\mathbf{x}_{b,0})) + \mathbf{H}_i\mathbf{M}_{i,0}\delta\mathbf{x}_0$, where the nonlinear model $\mathcal{M}_{i,0}$ integrates the state vector from $t = t_0$ to $t = t_i$, and the observation operator $\mathcal{H}_i$ maps model states to observation space and $\mathbf{x}_{b,0}$ is the background state vector. The matrices $\mathbf{H}_i$ and $\mathbf{M}_{i,0}$ are the tangent linear forms of $\mathcal{H}_i$ and $\mathcal{M}_{i,0}$, respectively. In this case, the cost function J_G appropriate for Gaussian-distributed

variables becomes

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

where $\mathbf{d}_{g,i} = \mathbf{y}_i - \mathbf{x}_{b,i}^o = \mathbf{y}_i - \mathcal{H}_i(\mathcal{M}_{i,0}(\mathbf{x}_{b,0}))$ define the innovations.

Similarly, the cost function J_L for lognormally-distributed variables is expressed in terms of increments $\delta\mathbf{g}_0 = \ln\mathbf{x}_0 - \ln\mathbf{x}_{b,0}$ as

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

where $\mathbf{d}_{l,i} = \ln\mathbf{y}_i - \ln\mathbf{x}_{b,i}^o = \ln\mathbf{y}_i - \ln(\mathcal{H}_i(\mathcal{M}_{i,0}(\mathbf{x}_{b,0})))$. The diagonal matrices $\mathbf{O}_{L,i}$ and $\mathbf{X}_L$ are introduced during the linearization of ln and exp function (Song et al., 2016a) where specifically, $\mathbf{O}_{L,i} = diag[(\mathbf{x}_{b,i}^o)_1,\ (\mathbf{x}_{b,i}^o)_2,\ \ldots,\ (\mathbf{x}_{b,i}^o)_{m_i}]^{-1}$ and $\mathbf{X}_L = diag[(\mathbf{x}_{b,0})_1,\ (\mathbf{x}_{b,0})_2,\ \ldots,\ (\mathbf{x}_{b,0})_{n_l}]$. A total of m_i lognormally-distributed observations exist at $t = t_i$, and a total of n_l lognormal variables exist within the model.

The state vector increment in the coupled, physical and biological system is defined as $\delta\mathbf{z} = [\delta\mathbf{x}_0^T\ \delta\mathbf{g}_0^T]^T$. The state vectors for physical and biological variables have dimensions $(n_g \times 1)$ and $(n_l \times 1)$, respectively. Hence the size of $\delta\mathbf{z}_0$ is simply $(n \times 1)$, where $n = n_g + n_l$. Then a compact form of the cost

function can be written as

$$
\begin{aligned}
J(\delta \mathbf{z}_0) &= \frac{1}{2}\delta \mathbf{z}_0^T \mathbf{B}^{-1} \delta \mathbf{z}_0 \\
&\quad + \frac{1}{2}\sum_{i=1}^{N_o} \left(\mathbf{d}_i - \mathbf{O}_i^{-1}\mathbf{H}_i \mathbf{M}_{i,0} \mathbf{X} \delta \mathbf{z}_0\right)^T \mathbf{R}_i^{-1} \\
&\quad \left(\mathbf{d}_i - \mathbf{O}_i^{-1}\mathbf{H}_i \mathbf{M}_{i,0} \mathbf{X} \delta \mathbf{z}_0\right), \qquad (3)
\end{aligned}
$$

where $\mathbf{d}_i^T = [\mathbf{d}_{g,i}^T \ \mathbf{d}_{l,i}^T]$, $\mathbf{O}_i = diag[1,\ 1,\ \ldots,\ 1,\ (\mathbf{x}_{b,i}^o)_1,\ (\mathbf{x}_{b,i}^o)_2,\ \ldots,\ (\mathbf{x}_{b,i}^o)_{m_i}]$ and $\mathbf{X} = diag[1,\ 1,\ \ldots,\ 1,\ (\mathbf{x}_{b,0})_1,\ (\mathbf{x}_{b,0})_2,\ \ldots,\ (\mathbf{x}_{b,0})_{n_l}]$. Error covariances $\mathbf{B}$ and $\mathbf{R}_i$ consist of error covariances for physical and biological components and their cross covariances.

The gradient of $J(\delta \mathbf{z}_0)$ with respect to $\delta \mathbf{z}_0$ is given by

$$
\frac{\partial J}{\partial \delta \mathbf{z}_0} = \mathbf{B}^{-1}\delta \mathbf{z}_0 - \mathbf{X}^T \sum_{i=1}^{N_o} \mathbf{M}_{0,i}^T \mathbf{H}_i^T \mathbf{O}_i^{-T} \mathbf{R}_i^{-1} \left(\mathbf{d}_i - \mathbf{O}_i^{-1}\mathbf{H}_i \mathbf{M}_{i,0} \mathbf{X} \delta \mathbf{z}_0\right), \qquad (4)
$$

and we seek a solution $\delta \mathbf{z}_0$ that satisfies $\partial J / \partial \delta \mathbf{z}_0 = 0$. The optimal $\delta \mathbf{z}_0$ is identified iteratively by applying a conjugate gradient descent algorithm using the Lanczos formulation (Moore et al., 2011c).

3. Experiment design for the coupled data assimilation system evaluation

We evaluate the performance of the new system by comparing results from multiple data assimilation experiments with a free run. For clarity of description, we distinguish between the coupled nonlinear model and the coupled data assimilation system by referring to the former as the forward model with no data assimilation. The free run results exclusively from the integration of the forward model.

We perform three data assimilation runs: the physical data assimilation (PDA) run, the biological data assimilation (BDA) run and the coupled data assimilation (PBDA) run. In the PDA run, physical data are used to constrain both physical and biological variables in the forward model. In the BDA run, only biological data are used to constrain physical and biological variables in the forward model. Third, both physical and biological data are used to constrain the physical and biological variables in the coupled model, and is referred to as the PBDA run. The PDA (BDA) applies Gaussian (Lognormal) 4DVar to fit the coupled model to the data. The PBDA fits both physical and biological data into the coupled model using a hybrid Gaussian and lognormal 4DVar approach.

Some additional experiments were performed to better evaluate the impact of model dynamics within the coupled assimilative system. Specifically, we modify the BDA experiment by reducing the control vector to include only biological variables alone (BDAb) or physical variables only (BDAp). Similarly, we consider the PDA experiment, but with adjustments to only the physical (PDAp) or biological (PDAb) variables. While physical state variables in the forward model clearly influence biological variables (e.g., through transport and mixing), biological variables generally do not alter physical variables in the forward model. In nature, chlorophyll pigments quantitatively impact light absorption and thus heat flux within the water column (Morel, 1988, Lewis et al., 1990, Frouin and Iacobellis, 2002, Murtugudde et al., 2002, Park et al., 2015), but this feedback is not included in the present model implementation. As a result, any misfit with respect to physical data cannot be reduced by adjusting the initial conditions of biolog-

ical variables. Therefore, the physical data misfit is identical for PDAp and PDA, and for PDAb and the free run.

3.1. Model

We use a Nutrient-Phytoplankton-Zooplankton-Detritus (NPZD) biological model coupled to ROMS, a 3-dimensional ocean circulation model, configured for the California Current System. This implementation has been applied repeatedly as a useful testbed for various developments of the ROMS 4DVar system (Broquet et al., 2009, 2011, Moore et al., 2011b,a). The model domain extends from the middle of the Baja Peninsula to the Washington coast and offshore to 137W. The horizontal model resolution is 1/30 degree, and it includes 30 terrain-following levels in the vertical. The configuration used here is identical to that testing the lognormal 4DVar in isolation and described in Song et al. (2016a), which provides additional details of the configuration, including the parameters used for the NPZD model.

Model twin experiments are an excellent way to evaluate the performance of data assimilation schemes because the true state is known exactly and the error statistics can be controlled. A 4-year forward simulation, begun on January 1^{st}, 2001, represents the "true" ocean state. A data assimilated run described by Broquet et al. (2009) provides the physical initial conditions. Biological initial conditions were obtained from the final state of a 45-year forward spin-up run described in Song et al. (2016a). Surface forcing and boundary conditions were derived from the output of the Coupled Ocean Atmosphere Mesoscale Prediction System (COAMPS) (Doyle et al., 2009) and the Simple Ocean Data Assimilation (SODA) (Carton and Giese, 2008) product, respectively.

Although real data is not used in the present experiments, we present an evaluation of the forward simulation using satellite derived estimates of sea surface temperature (AVHRR Pathfinder V5 SST, 0.44 degree resolution) and surface chlorophyll-a (SeaWiFS, 0.036 degree resolution) obtained from http://las.pfeg.noaa.gov/oceanWatch. Model chlorophyll is estimated using a constant carbon to chlorophyll ratio of 50 g C (g chl)$^{-1}$ and a Redfield ratio to convert model units of nitrogen to carbon. The model yields phytoplankton bloom-like patterns, intensity and spatial distribution comparable to satellite data (Song et al., 2016a). Monthly average fields are used to calculate the bias (model minus data), normalized by the standard deviation, and correlation coefficient (Figure 1). The standard deviation is estimated at each grid cell using the output from the model spin-up. Although SST in the simulation has a warm bias overall (1 °C), the correlation coefficient (r) is very high ($\bar{r} = 0.92$), indicating a good representation of the annual cycle in the model. Surface chlorophyll-a is biased low offshore and very near the coast north of 44°N, and biased high along the northern and central California coast out into the coastal transition zone (Brink and Cowles, 1991). On average, the model is biased low by approximately 0.5 mg m^{-3}. The correlation coefficient for chlorophyll-a reveals generally positive values over the whole domain.

3.2. Data

Physical and biological data are sampled from the true state. SSH and SST are observed at all grid points once per day (we assume no data dropouts due to cloud cover). In situ temperature and salinity profiles are obtained at times and locations based on the EN3 data set (Ingleby and Huddleston,

2007), which includes observations from the California Cooperative Fisheries Investigations (CalCOFI) surveys, as well as Argo and glider data within our model domain. For biological assimilation, only surface phytoplankton is used, analogous to what might be obtained under cloud-free conditions from satellite ocean color data. We note that our data collection for surface fields is larger than occurs in nature (approximately 13% data coverage on our model domain in the year 2000), but allows investigation of a best-case, data-rich scenario. Assimilation of real data is performed in Song et al. (2016b)

Observation errors are added to the sampled data. Errors for physical variables are drawn from normal random distributions ($\mathcal{N}(0, 0.1^2)$, $\mathcal{N}(0, 0.01^2)$, $\mathcal{N}(0, 0.02^2)$ and $\mathcal{N}(0, 0.1^2)$ for in situ temperature, salinity, SSH and SST, respectively). The observational error levels for in situ temperature, salinity and sea surface height were adopted from Broquet et al. (2009), where the same data assimilation system was used to fit the data in the same domain. The observational error level for SST is close to the global standard deviation of errors (0.13K for AVHRR) (O'Carroll et al., 2012). Errors in phytoplankton biomass data were drawn randomly from $\mathcal{N}(0, 0.2^2)$ and added in log-transformed space. This distribution approximately corresponds to a 20% multiplicative error which is lower than uncertainty estimates for global chlorophyll data (Gregg and Casey, 2004, Moore et al., 2009). We also consider sensitivity experiments in which the observational error for SST is elevated to 0.4°C and for phytoplankton is increased to 35% and 50%.

3.3. Assimilation setup

Following the method presented in Weaver and Courtier (2001), the background error covariance is factorized as $\mathbf{B} = \mathbf{\Sigma C \Sigma}^T$, where $\mathbf{\Sigma}$ is a diagonal matrix whose diagonal elements are model standard deviations and $\mathbf{C}$ is a correlation matrix. Standard deviations are computed from the 4-year forward simulation. The background error covariance $\mathbf{B}_L$ is for $\ln \mathbf{x}$, and therefore biological variables should be log-transformed before computing the standard deviation. The correlation matrix $\mathbf{C}$ is obtained through solution of a diffusion equation (Weaver and Courtier, 2001), and we apply horizontal and vertical length scales of 50 km and 30 m, respectively. It is reasonable to expect that in general physical and biological variables have different decorrelation length scales; for example, Lagrangian measurements in offshore portions of the CCS reveal different decorrelation time-scales for chlorophyll-a and temperature by Abbott and Letelier (1998). In this study, we assume that the length scales are identical. Song et al. (2016b) discuss the requirements of a smaller vertical decorrelation length scale for biological variables than physical variables in the fully realistic assimilation scenario.

The set of experiments proceeds in sequences of 5-day assimilation cycles. Although Veneziani et al. (2009) has shown that the tangent linear assumption in the physical model is reasonable over a time-scale of 14 days, a shorter time-scale is required for biological models due to the inherent nonlinearities of the biological interactions. Song et al. (2016a) find that a time-scale of 5 days is reasonable for the NPZD model and the California Current System implementation.

We examine the coupled assimilative system over the 4-year period 2001–

2004, divided into 48, 30-day experiments. Every 30-day experiment consists of 6 sequential assimilation cycles, each extending for 5-days. The initial condition on the first day of each experiment is the 4-year mean state obtained for that particular day obtained from the true run. Within each experiment, the state estimate at the end of one cycle is used to initialize the background estimate for the next 5-day cycle. In our analysis, the first cycle of each experiment is treated as a spin-up cycle when the linear approximation is the least accurate (Song et al., 2016a) and not included in the statistical results.

Although we recognize that cross-covariances between model variables exist, we calculate univariate correlations only, and we assume that observation errors are independent and uncorrelated. These assumptions simplify the construction of error covariances $\mathbf{B}$ and $\mathbf{R}_i$, with

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

and

$$\mathbf{R}_i = \begin{bmatrix} \mathbf{R}_{G,i} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_{L,i} \end{bmatrix}. \tag{6}$$

For simplicity, we take this highly simplified approach but acknowledge that accurate cross-correlations between model variables should yield additional improvements in coupled biological and physical state estimation.

The reformulation of lognormal 4DVar to the quadratic cost function creates an additional linearization approximation. The logarithmic transformation of the model state in observation space is linearized using Taylor series whose necessary condition is that the difference between the model state and observations is small. In order to satisfy this condition, we filter observations

such that those more than twice the modeled state are excluded from the assimilation procedure (Song et al., 2016a). While this procedure reduces the total number of assimilated observations it helps ensure the stability of the data assimilation calculations.

We also perform sensitivity experiments for the year 2001 to understand how increased observational error and the ability of the model to represent truth impact results. We consider experiments in which either or both SST and phytoplankton observational error are increased from their original values. In addition, the coupled data assimilation system is examined after purposely introducing model error. Physical model error is introduced by applying surface forcing from the year 2002, and biological model error is obtained by applying different biological parameter values than the reference run. Those parameters are listed in Table 1. Physical and biological model errors are included in the assimilative runs either separately or jointly.

4. Performance of the coupled data assimilation system

4.1. The improvement in the Root-mean-squared error

Performance is first evaluated using the root-mean-squared error (RMSE). RMSEs for surface physical variables – zonal velocity (u), meridional velocity (v), SST, and SSH – are significantly improved by coupled assimilation of physical variables (PDA) (Figure 2). Using the forward model provides better prediction skill than persistence, but assimilating physical observations further decreases RMSE. The mean error reduction from the free run for physical variables is approximately 56% with improvement in both observed variables (SST and SSH) and unobserved variables (u and v). RMSE

reduction is larger for observed variables than in unobserved variables, as expected. Coupled assimilation of biological data (BDA) also decreases RMSE for physical variables by about 13% from the free run on average, indicating that surface phytoplankton observations can improve initial conditions for current, SSH and SST entirely through the dynamics of the coupled system. Tracers are influenced by advection and diffusion, and because the adjoint model includes these coupled dynamics, cost function reduction can be achieved through alteration of physical as well as biological variables.

RMSE reduction is greatest, approximately 57% on average with respect to the free run, when assimilating both physical and biological data (PBDA). The modest improvement over PDA indicates that the physical fields are most constrained by the physical data provided in this experiment. It is noted that PBDA does not always lead to the improvement over PDA. The RMSE of PDA for SSH is slightly lower than what PBDA offers, but within the error bars (0.980 ± 0.017 cm versus 0.985 ± 0.014 cm). In a case of SST, PBDA has lower RMSE (0.310 ± 0.006 °C) than PDA (0.314 ± 0.006 °C), but again within error bars. Although the improvement by PBDA over PDA is statistically marginal for u and v, it is negligible compared with the RMSE reduction from the free run to PDA.

PDA and PBDA reduce the RMSEs for the observed variables (SST and SSH) more than half almost everywhere (Figure 3(c,d,k,l)). Although the reduction of RMSE by BDA is less than 20%, BDA reduces the RMSE of SST and SSH over the entire region (Figure 3(g,h)). Surface currents are not assimilated, and RMSE reduction in those variables is smaller than for SST or SSH. There are regions even with the increased RMSE. For instance,

PDA shows approximately 20% higher RMSE for u near the coast between 40°N and 45°N (Figure 3(a)). Although the RMSE reduction for u and v by BDA is smaller than for PDA, it occurs over the all model domain (Figure 3(e,f)). PBDA yields the best estimate of surface currents, resulting in the smallest RMSE. Interestingly, the increased RMSE for u by PDA in Figure 3(a) became less obvious when assimilating both physical and biological observations.

Forward simulation of the NPZD model improves the estimation of biological variables as the RMSEs of the free run are smaller than the result from one-month persistence. Assimilating data provides a better estimate for biological variables, further reducing the RMSEs for biological variables with respect to the free run. In BDA, incorporating surface phytoplankton data improves the estimate of not only phytoplankton but also other unobserved biological variables. RMSE reduction of biological variables is approximately 43% on average with respect to that by the free run. PDA also improves the estimate of biological variables by approximately 7%, and this improvement results not from the adjustment of biological initial conditions but from the improved representation of the circulation fields. As described above, biological variables are passively coupled to the currents, and therefore the sensitivity of the misfits in physical variables to variations in the biological variables is zero. Improvements in T, S and SSH cannot be reduced by changes in biological variables at the initial time. As in the evaluation of physical variables, greatest RMSE reduction ($\sim$49% on average) for biological variables occurs through coupled assimilation of both physical and biological data (PBDA), and this reduction is statistically significant.

Assimilating physical variables leads to mixed effects on the biological state estimation at the surface. Figure 5(a-d) show that PDA has a positive effect near the coastal regions but generally degrades the biological estimation offshore. Changes in RMSEs by PDA are similar in overall magnitude for all four biological variables. BDA and PBDA result in comparable RMSE reduction (Figure 5(e-l)). Largest RMSE improvement using these two methods occurs for phytoplankton, the observed variable. The second largest reduction in RMSE is seen in detritus. RMSE reduction for P and D occurs throughout the model domain. The improvements in zooplankton by BDA and PBDA occur mainly near the coast (Figure 5(f,j)). Least improvement is found in the nutrient estimation, and it is visually similar to the improvement by PDA (Figure 5(c,g,k)).

We note that although overall error decreases in all variables, there are limited regions where the RMSE increases after assimilating surface phytoplankton, even in PBDA. Such error increases occur most visibly in Z and N, but also at one location in D (Figure 5(f,g,j,k)). The NPZD model is a simple but highly nonlinear system, sometimes stretching the linear approximation used in 4DVar systems. In such cases, the increments can degrade the posterior estimate. Although this is not limited to our assimilation system, it is possible that degrading increments can be amplified due to the transformation back to the original space using the exponential function. We note that the limited areas of degradation occur for unobserved variables only, indicating that the system improves the phytoplankton estimates through occasionally unrealistic changes to variables for which we have no information other than background error statistics. Such performance is not

surprising in an estimation system, and generally could be improved through observation of other ecosystem elements.

Two additional experiments outlined above better illustrate how modeled dynamics in coupled data assimilation influence the final state estimate. In BDAb, biological data is assimilated but only biological variables are adjusted. In this case, RMSE in biological fields is reduced by approximately 35% on average (not shown), less than the reduction by BDA with adjustments to all variables. In BDA, the coupled data assimilation system partitions improvements to both physical and biological variables. Adjusting only biological fields limits the quantitative improvement in biological fields over the full assimilation cycle relative to what can be achieved through adjustment also of physical fields. RMSE reduction in physical variables is zero in BDAb because adjusted biological variables do not affect the physical fields.

In BDAp physical variables only are adjusted, and misfits in biological variables can be reduced only through improvement in circulation and mixing. Although the phytoplankton biomass RMSE is reduced in this case, assimilation generally degrades the estimates of other variables (not shown). In particular, physical variables are adjusted such that their RMSEs are larger than in the free run. While the misfit in observed variables can be reduced through modification of various fields through coupled dynamics, not all adjustments result in a better overall estimate of the unobserved variables.

4.2. The improvement in the statistical states

Model performance can be visualized also through a Taylor diagram which summarizes the model variability relative to the truth, specifically the correlation coefficient and standard deviation (Taylor, 2001). PDA (square termi-

nator) and PBDA (circular terminator) show substantial improvements in all three statistics relative to the prior, especially in observed variables (SST and SSH) (Figure 6). Unobserved u and v are also statistically improved when physical data are assimilated. For both PDA and PBDA, physical variables show about the same amount of variability as the true state, and posterior correlation coefficients are greater than 0.8. Statistically, the improvements realized by the PDA and PBDA are comparable, implying that physical observations in this case provide sufficient information for the optimal physical solution. In BDA (triangular terminator), the statistics of the physical variables are also improved, with the posterior located closer to the reference point than the line origins for u, v and SSH, although the improvements are not as large as for PDA and PBDA. Chlorophyll and ocean currents are strongly coupled in the advection/diffusion equation while temperature does not appear in the equations for NPZD model. Hence, chlorophyll observations have a more substantial impact on u,v and SSH.

In the right panel in Figure 6, results from BDA and PBDA indicate improvements in the statistics for all biological variables. As expected, phytoplankton, the observed variable, exhibits the greatest improvement, showing a normalized standard deviation close to 1 and correlation coefficient greater than 0.9. The statistics of the unobserved variables are also improved, although not as much as for P. Improvements in the biological error statistics from BPDA are greater than those for BDA, and more so than improvements in the physical error statistics by PBDA over PDA; this result indicates again the significant role that the physical state has on biological fields but not the reverse. Results from PDA show the least improvement in biological vari-

480 ables.

481 *4.3. Sensitivity to observational errors*

482 The performance of data assimilation is dependent on the error levels. We
483 conduct a sensitivity test with varied observational error levels for SST and
484 phytoplankton. When observational error for SST is increased from 0.1°C
485 to 0.4°C, the RMSE for SST is elevated by approximately 50% (PDA(0.1,
486 N/A) versus PDA(0.4, N/A) in Figure 7). Higher SST observational error
487 also influences the estimation of surface currents and SSH, increasing RM-
488 SEs more than 20%. The RMSEs for physical variable are less sensitive to
489 the phytoplankton observational error (BDA(N/A, 20%) versus BDA(N/A,
490 35%) versus BDA(N/A, 50%) in Figure 7). BDA reduces the RMSEs for
491 physical variables with respect to the free run even when the phytoplankton
492 observational error level is 50%. The sensitivity of PBDA to the observa-
493 tional error can be considered as the mixed response of PDA and BDA to
494 the changes in observational error for SST and phytoplankton, respectively
495 (PBDA(0.1, 20%) versus PBDA(0.4, 35%) versus PBDA(0.4, 50%) in Figure
496 7).

497 Changing observational error for SST does not provide a statistically
498 significant impact on the RMESs for biological variables (PDA(0.1, N/A)
499 versus PDA(0.4, N/A) in Figure 8). Higher phytoplankton observational
500 error degrades the estimation of phytoplankton by elevating the RMSE by
501 more than 15%, which is statistically significant (BDA(N/A, 20%) versus
502 BDA(N/A, 35%) versus BDA(N/A, 50%) in Figure 8). Other biological
503 variables do not have the influence of higher phytoplankton observational
504 error as much as phytoplankton, showing less than 10% RMSE increase. As

for physical variables, the sensitivity of PBDA to the observational error can also be viewed as the mixed response of PDA and BDA to the changes in observational error for SST and phytoplankton (PBDA(0.1, 20%) versus PBDA(0.4, 35%) versus PBDA(0.4, 50%) in Figure 8). It is noted that higher phytoplankton observational error does not always degrade the estimation for the biological variables. For instance, BDA(N/A, 50%) shows smaller RMSE for P than BDA(N/A, 35%). We attribute this to the observation filtering process discussed in subsection 3.3. More outlying observations are rejected in BDA(N/A, 50%) than in BDA(N/A, 35%), and eventually the filter helps the data assimilation run fit observations better, leading to smaller RMSE.

4.4. Sensitivity to the model errors

The performance of coupled physical and biological data assimilation is also evaluated under the presence of model errors. In this sensitivity test, the control run is PBDA with 0.4°C and 35% observational errors for SST and phytoplankton, respectively. We refer to EF as error free, EP as error in physics (in which the wrong year's surface forcing has been introduced), EB as error in biology (with different NPZD model parameters than the reference run), and EPB as error in physics and biology. PBDA results in higher RMSEs in the observed variables, SST and SSH, when the surface forcing of the year 2002 is used in the assimilative run for the year 2001, but RMSEs are still considerably lower than that of the free run in which no model error is included (PBDA, EF versus PBDA, EP in Figure 9). However, the introduced physical model error degrades the estimation of surface currents: RMSEs of PBDA, EP for u and v are greater than that of the free run with no model error. The impact of biological model error to physical variables

is not statistically significant. Using incorrect biological parameter values in PBDA does not change RMSEs for all physical variables (PBDA, EP versus PBDA, EPB in Figure 9).

Estimating the biological state can be influenced by both physical and biological model errors. Using the wrong forcing degrades the estimate of biological variables, leading higher RMSEs (PBDA, EF versus PBDA, EP in Figure 10). The degradation is particularly strong for N and D, resulting in higher RMSEs than those from the free run. The introduced biological model error also has statistically significant impact on the estimation of biological variables. Using wrong parameter values and surface forcing increased the RMSE for phytoplankton, although it is still slightly smaller than that from the free run with no model error (PBDA, EF versus PBDA, EP versus PBDA, EPB in Figure 10). However, our model errors result in higher RMSEs for unobserved biological variables. Including both physical and biological errors makes the RMSEs for Z, N and D greater with respect to the free run. The most substantial impact is observed in Z and is perhaps results from the fact that two of four modified parameters are associated with the zooplankton dynamics.

The incremental form of 4DVar used here does not allow easily for the accurate computation of posterior error estimates. The posterior error covariance is equivalent to the inverse of Hessian matrix (Moore et al., 2012). Here, the inverse of Hessian matrix is estimated as $\mathbf{V}\mathbf{T}^{-1}\mathbf{V}^{T}$, where $\mathbf{V}$ is the orthogonal matrix with Lanczos vectors, and $\mathbf{T}$ is the symmetric tridiagonal matrix that contains coefficients in Lanczos recurrence relation (Song et al., 2016a). The Hessian matrix is approximated by a tridiagonal factorization

using the Lanczos vectors, and represents a reduced approximation. The leading eigenvectors of the Hessian matrix can be estimated from the Lanczos vectors but these would represent the smallest eigenvectors of the analysis error covariance. Therefore, this calculation provides a poor representation of the analysis error covariance matrix. One can consider other effective methods (e.g. Daescu and Navon, 2007, a reduced second order adjoint model) for the estimation of Hessian, but the inverse of the Hessian matrix is still approximated with the least important orthogonal vectors. We note that that ROMS does provide options to estimate the posterior error covariance in dual form (Moore et al., 2012), but that is not the form used in this study.

5. Summary and Discussion

We have developed and investigated combined physical and biological 4-dimensional variational data assimilation in an ocean model. Biological data assimilation benefits from a unique approach because of the non-Gaussian statistics of biological variables and their errors. We have assumed lognormal statistics for these variables and applied the quadratic formulation of the incremental approximation developed by Song et al. (2016a). Assimilation of variables having different error statistics is required for combined physical and biological assimilation and proceeds here following the approach of Fletcher (2010) and Fletcher and Jones (2014).

In model twin experiments using ROMS and a 4-component, NPZD ecosystem model configured for the realistic California Current System, we investigated how coupled biological and physical data assimilation improves overall estimates of the combined physical and biological ocean state. Ob-

servations were drawn from a forward model run which represented the true ocean state with errors drawn from distributions with known statistics added. Three observation sampling strategies were chosen: (1) biological observations, (2) physical observations, and (3) both biological and physical observations. Then we altered the model state through adjustment of initial conditions in physical and/or biological variables using those observations. Statistics of RMSE, correlation coefficients, state variability were analyzed from a total of 48 sequences of six 5-day assimilation cycles.

We found that assimilation of physical data (PDA) improves model-data misfit of physical variables, and assimilation of biological data (BDA) reduces model error for biological variables. Such results should be expected. In addition, PDA resulted in biological error reduction, and BDA yielded improvements in the physical variable misfit. Even though PDA has no effect on biological initial conditions, it does influence biological variables over the entire assimilation cycle through improvements to the physical fields which then feed back on biological variables through tangent linear (and forward) model advection and diffusion. In contrast, a change in biological initial conditions, for example resulting from BDA, has no influence on physical variables through the forward model directly. However, BDA influences physical variable initial conditions through coupled dynamics included in the adjoint model. The coupled data assimilation system partitions changes to all control vector elements that can reduce error in the biological data misfit, and this partitioning extends to both biological and physical variables. This conceptual division of influence is drawn schematically in Figure 11.

Overall, the greatest performance in both physical and biological fields

as quantified by various measures resulted from the combined assimilation of physical and biological data (PBDA), further supporting the interpretation of the PDA and BDA results. While physical observations provide the most effective constraint for physical variables, and biological observations most constrain biological variables, additional improvement in physical variables derived from biological information through model adjoint dynamics and biological errors can be reduced through physical observations via the tangent linear (and forward) model. Higher observational errors in SST and phytoplankton increase the RMSEs of PBDA, but the increase is smaller than the combined increments in RMSE of PDA with higher SST observational error and of BDA with higher phytoplankton observational error. Introducing model errors (through the application of incorrect surface forcing or altered biological parameters) also degrades the performance of PBDA, but its impact on PBDA is not particularly different from that for PDA and BDA on a monthly time scale.

More generally, variables in a coupled 4DVar system can be influenced in two ways, dynamically through the adjoint and tangent linear models and statistically through covariances of the background error covariance matrix. In this study, univariate spatial correlations in fields were assumed through the integration of a diffusion equation; however, no multivariate correlations were represented, and therefore all improvements resulting from observations of coupled variables resulted exclusively from adjoint and tangent linear model dynamics. These coupling dynamics are not included in alternate data assimilation approaches based on statistical estimation alone, and statistical correlations between physical and biological variables provide

the only way to transfer this critical information. These correlations are not well-known in nature, though some groups have reported such information (Behrenfeld et al., 2006), and estimates that are consistent with the ocean circulation and in principle ecosystem models can be calculated from forward model calculations (Shulman et al., 2013). Ensemble-based data assimilation approaches use the ensemble to estimate time-dependent correlations (Simon et al., 2015), usually with an inflation factor and localization to compensate the effects of having a small ensemble. We would expect that the present 4DVar assimilation approach would further benefit from better background error covariance estimates, a subject for future study.

The variational approach to coupled dynamics with mixed statistics presented here is conceptually straightforward to implement within any existing coupled system equipped with tangent and adjoint models and assuming no multivariate correlations between physical and biological variables. The computational cost of the combined physical and biological system is comparable to the cost of either the physical or biological system in isolation. Results from this study suggest that coupled assimilation using 4DVar is practical and realizable. However, the twin experiment framework used here represents an idealized setting in which the data is unencumbered by cloud cover and the model surface forcing and boundary conditions are error free.

Our conclusions are drawn based on the ensemble of 30-day assimilation (6 cycles). It is possible that the model may drift from the truth if biases are introduced by assimilation and accumulate over time scales longer than one month. In our ideal twin experiment, error is introduced in the initial conditions only and estimating accurate initial conditions always improves

654 the model bias. Indeed, the model bias does not increase for more than 6
655 months when we tested the data assimilation system using 48 cycles (not
656 shown). However, in realistic scenarios in which errors in surface forcing,
657 boundary forcing, and model construction exist, it is possible that model
658 bias develops more rapidly. Such an issue does not appear in Song et al.
659 (2016b), but further studies of model bias in realistic scenarios over long
660 periods of time is warranted.

661 When assimilating real observations, the presented assimilation system
662 may encounter obstacles. For instance, the model dynamics inevitably mis-
663 represent or entirely miss important processes in nature. Under these circum-
664 stances, adjustments to the initial conditions determined by model dynamics
665 are of limited value in matching observations. Large differences between the
666 observed and prior state can also create an issue because they violate the
667 linearization of log-transformation function and may prevent solution con-
668 vergence. In order to prevent this outcome, the coupled assimilation system
669 requires a filtering process that excludes observations far from the prior. The
670 filtering process may reduce the number of observations, but it stabilizes the
671 assimilation system and may lead to a better posterior solution as shown
672 in subsection 4.3. We note that the filtering procedure is reevaluated during
673 each outer loop of the assimilation system, and observations that are rejected
674 initially may be included in the final outer loop.

675 In a companion paper Song et al. (2016b), we investigate the assimila-
676 tion system's performance in a more realistic system in which real remotely
677 sensed and in situ physical and ecosystem data are assimilated. In that real-
678 istic setting, improvements to RMSE for physical variables is not improved

by assimilating real chlorophyll observations. Several factors may account for this result that stands in contrast to that offered by the twin experiments here. Data availability is reduced relative to this study as frequent cloud cover prevents collection of SST and chlorophyll data over much of our domain. In addition, physical and biological model error are likely greater than that considered in the present study. Even though the real assimilation experiment is carried out on a higher resolution grid that better resolves the CCS mesoscale circulation, the physical model is still imperfect relative to nature. The NPZD ecosystem model used is advantageous for its relative simplicity, and has been applied to multiple realistic studies of ocean biogeochemistry, including in the CCS (Powell et al., 2006). However, with only one phytoplankton functional group, it is less than ideal in representing the multiple phytoplankton communities in different geographical regions of the CCS. Assimilation improvements may result from application of a more complex biogeochemical model. At this time, we do not know which of these elements is responsible for the differences between the two studies, but the present study shows that under excellent conditions in which a model is nearly able to represent truth and observations are abundant, the lowest RMSE for physical and biological variables results from assimilation of both biological and physical variables into the coupled system.

6. Acknowledgement

We are grateful for grants from the Gordon and Betty Moore Foundation and from the National Oceanographic and Atmospheric Administration office of Oceanic and Atmospheric Research (award number NA10OAR4320156)

703 that supported this research. The authors would like to thank four anony-
704 mous reviewers for valuable comments and suggestions, which significantly
705 improved the manuscript.

706 Abbott, M. R., Letelier, R. M., 1998. Decorrelation scales of chlorophyll
707 as observed from bio-optical drifters in the california current. Deep Sea
708 Research Part II: Topical Studies in Oceanography 45 (89), 1639 – 1667.

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

712 Balmaseda, M. A., Vidard, A., Anderson, D. L. T., 2008. The ECMWF
713 Ocean Analysis System: ORA-S3. Mon. Wea. Rev. 136, 3018–3034.

714 Behrenfeld, M. J., O'Malley, R. T., Siegel, D. A., McClain, C. R., Sarmiento,
715 J. L., Feldman, G. C., Milligan, A. J., Falkowski, P. G., Letelier, R. M.,
716 Boss, E. S., 2006. Climate-driven trends in contemporary ocean produc-
717 tivity. Nature 444, 752–755.

718 Behringer, D. W., Ji, M., Leetmaa, A., 1998. An improved coupled model
719 for ENSO prediction and implications for ocean initialization. Part I: The
720 ocean data assimilation system. Mon. Wea. Rev. 126, 1013–1021.

721 Bell, M. J., Forbes, R. M., Hines, A., 2000. Assessment of the FOAM global
722 data assimilation system for real-time operational ocean forecasting. J.
723 Mar. Sys. 25, 1–22.

724 Berline, L., Brankart, J. M., Brasseur, P., Ourmières, Y., Verron, J., 2007.
725 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.

Bertino, L., Evensen, G., Wackernagel, H., 2003. Sequential data assimilation techniques in oceanography. Int. Statist. Rev. 71, 223–241.

Brink, K. H., Cowles, T. J., 1991. The coastal transition zone program. Journal of Geophysical Research: Oceans 96 (C8), 14637–14647.

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.

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, 2999–3017.

Chai, F., Liu, G., Xue, H., Shi, L., Chao, Y., Tseng, C.-M., Chou, W.-C., Liu, K.-K., 2009. Seasonal and interannual variability of carbon cycle in South China Sea: A three-dimensional physical-biogeochemical modeling study. Journal of Oceanography 65 (5), 703–720.

Chassignet, E. P., Hurlburt, H. E., Smedstad, O. M., Halliwell, G. R., Hogan, P. J., Wallcraft, A. J., Baraille, R., Bleck, R., 2007. The HYCOM (HYbrid

748 Coordinate Ocean Model) data assimilative system. J. Marine Syst. 65,
749 60–83.

750 Chenillat, F., Rivière, P., Capet, X., Franks, P. J. S., Blanke, B., 2013.
751 California coastal upwelling onset variability: Cross-shore and bottom-up
752 propagation in the planktonic ecosystem. PLoS ONE 8 (5).

753 Ciavatta, S., Torres, R., Saux-Picart, S., Allen, J. I., 2011. Can ocean color
754 assimilation improve biogeochemical hindcasts in shelf seas? J. of Geophys.
755 Res.: Oceans 116 (C12), n/a–n/a.

756 Cox, P. M., Betts, R. A., Spall, S. A., Totterdell, I. J., 2000. Acceleration of
757 global warming due to carbon-cycle feedbacks in a coupled climate model.
758 Nature 408, 184–187.

759 Cummings, J., Bertino, L., Brasseur, P., Fukumori, I., Kamachi, M., Martin,
760 M., Mogensen, K., Oke, P., Testut, C., Verron, J., Weaver, A., 2009. Ocean
761 data assimilation systems for GODAE. Oceanography 22 (3), 96–109.

762 Daescu, D., Navon, I., 2007. Efficiency of a POD-based reduced second order
763 adjoint model in 4D-Var data assimilation. Int. J. Numer. Methods Fluids
764 53, 985–1004.

765 Doron, M., Brasseur, P., Brankart, J.-M., Losa, S., Mellet, A., 2013.
766 Stochastic estimation of biogeochemical parameters from Globcolour ocean
767 colour satellite data in a North Atlantic 3D ocean coupled physical-
768 biogeochemical model. J. Marine Syst. 117–118, 81–95.

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

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

774 Fennel, K., 2010. The role of continental shelves in nitrogen and carbon
775 cycling: Northwestern North Atlantic case study. Ocean Sci. 6, 539–548.

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

779 Fletcher, S. J., 2010. Mixed Gaussian-lognormal four-dimensional data as-
780 similation. Tellus A 62, 266–287.

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

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

786 Follows, M. J., Dutkiewicz, S., Grant, S., Chisholm, S. W., 2007. Emer-
787 gent biogeography of microbial communities in a model ocean. Science
788 315 (5820), 1843–1846.

789 Ford, D. A., Edwards, K. P., Lea, D., Barciela, R. M., Martin, M. J., De-

790 maria, J., 2012. Assimilating globcolour ocean colour data into a pre-
791 operational physical-biogeochemical model. Ocean Science 8 (5), 751–771.

792 Franks, P. J. S., Chen, C., 2001. A 3-D prognostic numerical model study of
793 the Georges bank ecosystem. Part ii: biologicalphysical model. Deep-Sea
794 Res. II 48, 457–482.

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

798 Frouin, R., Iacobellis, S. F., 2002. Influence of phytoplankton on the global
799 radiation budget. J. Geophys. Res.: Atmospheres 107 (D19), ACL 5–1–
800 ACL 5–10, 4377.
801 URL http://dx.doi.org/10.1029/2001JD000562

802 Garcia-Gorriz, E., Hoepffner, N., Ouberdous, M., 2003. Assimilation of Sea-
803 WiFS data in a coupled physical-biological model of the Adriatic Sea. J.
804 Marine Syst. 40-41, 233–252.

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

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

810 Gregg, W. W., Casey, N. W., 2004. Global and regional evaluation of the
811 SeaWiFS chlorophyll data set. Remote Sens. Environ. 93, 463 – 479.

812 Hoteit, I., Triantafyllou, G., Petihakis, G., 2005. Efficient data assimilation
813 into a complex, 3-d physical-biogeochemical model using partially-local
814 Kalman filters. Ann. Geophys. 23, 3171–3185.

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

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

820 Köhl, A., Stammer, D., Cornuelle, B. D., 2007. Interannual to decadal
821 changes in the ECCO global synthesis. J. Phys. Oceanogr. 37, 313–337.

822 Kurapov, A. L., Foley, D., Strub, P. T., Egbert, G. D., Allen, J. S., 2011.
823 Variational assimilation of satellite observations in a coastal ocean model
824 off Oregon. J. Geophys. Res.: Oceans 116 (C5).

825 Lewis, M., Carr, M., Feldman, G., Esaias, W., McClain, C., 1990. Influence
826 of penetrating solar radiation on the heat budget of the equatorial Pacific
827 Ocean. Nature 347 (6293), 543–545.

828 Losa, S., Kivman, G., Ryabchenko, V., 2004. Weak constraint parameter
829 estimation for a simple ocean ecosystem model: What can we learn about
830 the model and data? J. Marine Syst. 45, 1–20.

831 Matear, R. J., 1995. Parameter optimization and analysis of ecosystem mod-
832 els using simulated annealing: A case study at Station P. J. Mar. Res.
833 53 (4), 571–607.

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

837 Matthews, D., Powell, B. S., Janekovi, I., 2012. Analysis of four-dimensional
838 variational state estimation of the hawaiian waters. J. of Geophys. Res.:
839 Oceans 117 (C3).

840 Miller, A. J., Di Lorenzo, E., Neilson, D. J., Cornuelle, B. D., Moisan, J. R.,
841 2000. Modeling CalCOFI observations during El Niño: Fitting physics and
842 biology. Calif. Coop. Ocean. Fish. Invest. Rep. 41, 87–97.

843 Moore, A., Arango, H., Broquet, G., Edwards, C. A., Veneziani, M., Pow-
844 ell, B., Foley, D., Doyle, J., Costa, D., Robinson, P., 2011a. The Regional
845 Ocean Modeling System (ROMS) 4-dimensional variational data assimila-
846 tion systems, Part III: Observation impact and observation sensitivity in
847 the California Current System. Prog. Oceanogr. 91, 74–94.

848 Moore, A. M., Arango, H. G., Broquet, G., 2012. Estimates of analysis and
849 forecast error variances derived from the adjoint of 4D-Var. Mon. Wea.
850 Rev. 140, 3183–3203.

851 Moore, A. M., Arango, H. G., Broquet, G., Edwards, C. A., Veneziani, M.,
852 Powell, B. S., Foley, D., Doyle, J., Costa, D., Robinson, P., 2011b. The
853 Regional Ocean Modeling System (ROMS) 4-dimensional variational data
854 assimilation systems, Part II: Performance and application to the Califor-
855 nia Current System. Prog. Oceanogr. 91, 50–73.

Moore, A. M., Arango, H. G., Broquet, G., Powell, B. S., Zavala-Garay, J., Weaver, A. T., 2011c. 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.

Morel, A., 1988. Optical modeling of the upper ocean in relation to its biogenous matter content (case i waters). J. Geophys. Res.: Oceans 93 (C9), 10749–10768.
URL http://dx.doi.org/10.1029/JC093iC09p10749

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., 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.

O'Carroll, A. G., August, T., Borgne, P. L., Marsouin, A., 2012. The accuracy of SST retrievals from Metop-A IASI and AVHRR using the EUMETSAT OSI-SAF matchup dataset. Remote Sens. Environ. 126, 184–194.

Oke, P., Brassington, G., Griffin, D., Schiller, A., 2008. The Bluelink ocean data assimilation system (BODAS). Ocean Modell. 21, 46–70.

878 Oschlies, A., Garçon, V., 1998. Eddy enhancement of primary production in
879 a model of the North Atlantic Ocean. Nature 394, 266–269.

880 Ourmières, Y., Brasseur, P., Lévy, M., Brankart, J.-M., Verron, J., 2009. On
881 the key role of nutrient data to constrain a coupled physicalbiogeochemical
882 assimilative model of the North Atlantic Ocean. J. Marine Syst. 75, 100–
883 115.

884 Park, J.-Y., Kug, J.-S., Bader, J., Rolph, R., Kwon, M., 2015. Amplified arc-
885 tic warming by phytoplankton under greenhouse warming. PNAS 112 (19),
886 5921–5926.

887 Powell, T., Lewis, C., Curchitser, E., Haidvogel, D., Hermann, A., Dob-
888 bins, E., 2006. Results from a three-dimensional, nested, biologicalphysi-
889 cal model of the california current system and comparisons with statistics
890 from satellite imagery. J. Geophys. Res. 111, C07018.

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

895 Rousseaux, C. S., Gregg, W. W., 2012. Climate variability and phyto-
896 plankton composition in the Pacific Ocean. J. of Geophys. Res.: Oceans
897 117 (C10), n/a–n/a.

898 Roy, S., Broomhead, D., Platt, T., Sathyendranath, S., Ciavatta, S., 2012.
899 Sequential variations of phytoplankton growth and mortality in an NPZ
900 model: A remote-sensing-based assessment. J. Marine Syst. 92, 16–29.

Sakov, P., Counillon, F., Bertino, L., Lister, K., Oke, P., Korablev, A., 2012. TOPAZ4: an ocean-sea ice data assimilation system for the North Atlantic and Arctic. Ocean Science 8 (4), 633–656.

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. J. of Geophys. Res.: Oceans 118 (4), 2215–2230.

Shulman, I., Rowley, C., Anderson, S., DeRada, S., Kindle, J., Martin, P., Doyle, J., Cummings, J., Ramp, S., Chavez, F., Fratantoni, D., Davis, R., 2009. Impact of glider data assimilation on the monterey bay model. Deep-Sea Res. II 56 (35), 188 – 198.

Simon, E., Bertino, L., 2009. Application of the Gaussian anamorphosis to assimilation in a 3-D coupled physical-ecosystem model of the North Atlantic with the EnKF: a twin experiment. Ocean Sci. 5, 495–510.

Simon, E., Samuelsen, A., Bertino, L., Mouysset, S., 2015. Experiences in multiyear combined state-parameter estimation with an ecosystem model of the North Atlantic and Arctic Oceans using the Ensemble Kalman Filter. J. Marine Syst. 152, 1–17.

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 Cur-

rent 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 3, Assimilation in a realistic context using satellite and in situ observations. Ocean Modell. In Press.

Spitz, Y., Moisan, J., Abbott, M., Richman, J., 1998. Data assimilation and a pelagic ecosystem model: parameterization using time series observations. J. Mar. Systems 16 (12), 51 – 68.

Stock, C. A., Alexander, M. A., Bond, N. A., Brander, K. M., Cheung, W. W., Curchitser, E. N., Delworth, T. L., Dunne, J. P., Griffies, S. M., Haltuch, M. A., Hare, J. A., Hollowed, A. B., Lehodey, P., Levin, S. A., Link, J. S., Rose, K. A., Rykaczewski, R. R., Sarmiento, J. L., Stouffer, R. J., Schwing, F. B., Vecchi, G. A., Werner, F. E., 2011. On the use of IPCC-class models to assess the impact of climate on Living Marine Resources. Prog. Oceanogr. 88, 1–27.

Taylor, K. E., 2001. Summarizing multiple aspects of model performance in a single diagram. J. Geophys. Res. 106, 7183–7192.

Triantafyllou, G., Korres, G., Hoteit, I., Petihakis, G., Banks, A., 2007. Assimilation of ocean colour data into a biogeochemical flux model of the eastern Mediterranean Sea. Ocean Sci. 3, 397–410.

Table 1: A list of parameters for the NPZD model that is chosen differently in the assimilation runs to include the biological model error. The columns from the left to the right represent names, units, values in the true run and values for biological model error, respectively. The parameter values for the biological model error are from the study for Gulf of Alaska in Fiechter et al. (2011).

Parameter name	unit	Value, True	Value, EB
Uptake rate for nitrate	day^{-1}	1.0	0.8
Zooplankton grazing rate	day^{-1}	0.65	0.4
Ivlev constant	Dimensionless	0.4	0.84
Detritus remineralization rate	Dimensionless	0.1	0.2

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

950 Weaver, A., Courtier, P., 2001. Correlation modelling on the sphere using a
951 generalized diffusion equation. Quart. J. Roy. Meteorol. Soc. 127, 1815–
952 1846.

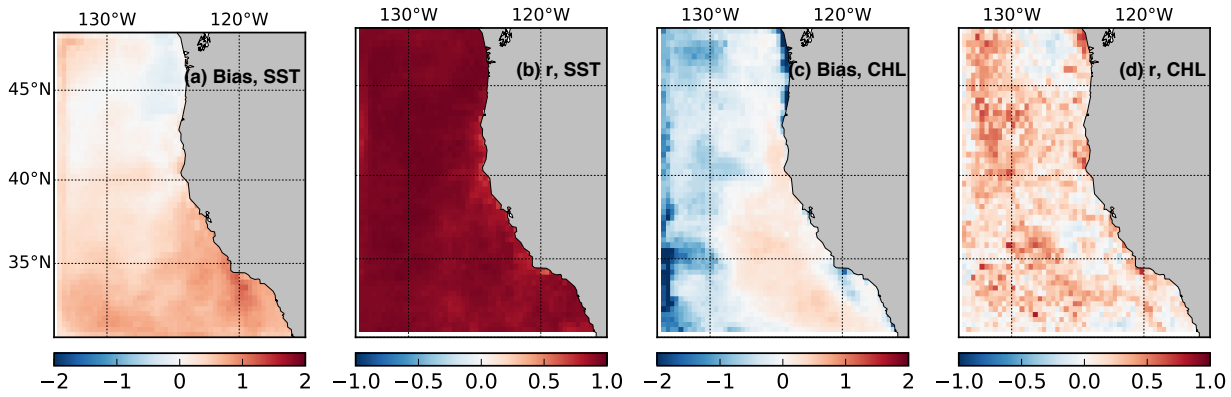

Figure 1: Normalized bias of the simulated (a) SST and (c) surface chlorophyll that are assumed as the truth in a twin experiment, and their correlation coefficients (b, d). Monthly mean values from the model and the satellite observation AVHRR (SeaWiFS), as the reference states for SST (surface chlorophyll), are considered in the calculation.

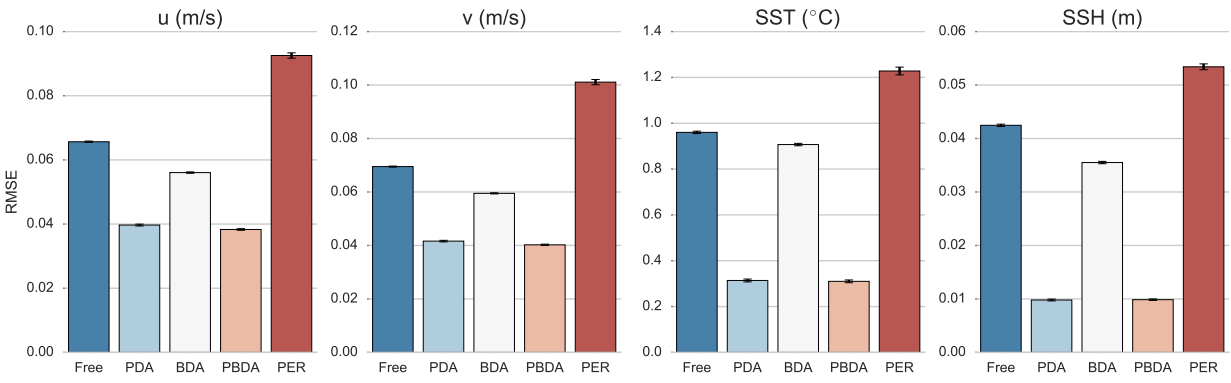

Figure 2: Root-mean-squared error (RMSE) of u, v, SST and SSH at the surface in the four different simulation: Free run (blue), analysis by the PDA (light blue), analysis by the BDA (white), analysis by the PBDA (light red). Red bar represents the one-month persistence RMSE. Error bars represent standard error from 1200 days (25 days $\times$ 12 months $\times$ 4 years).

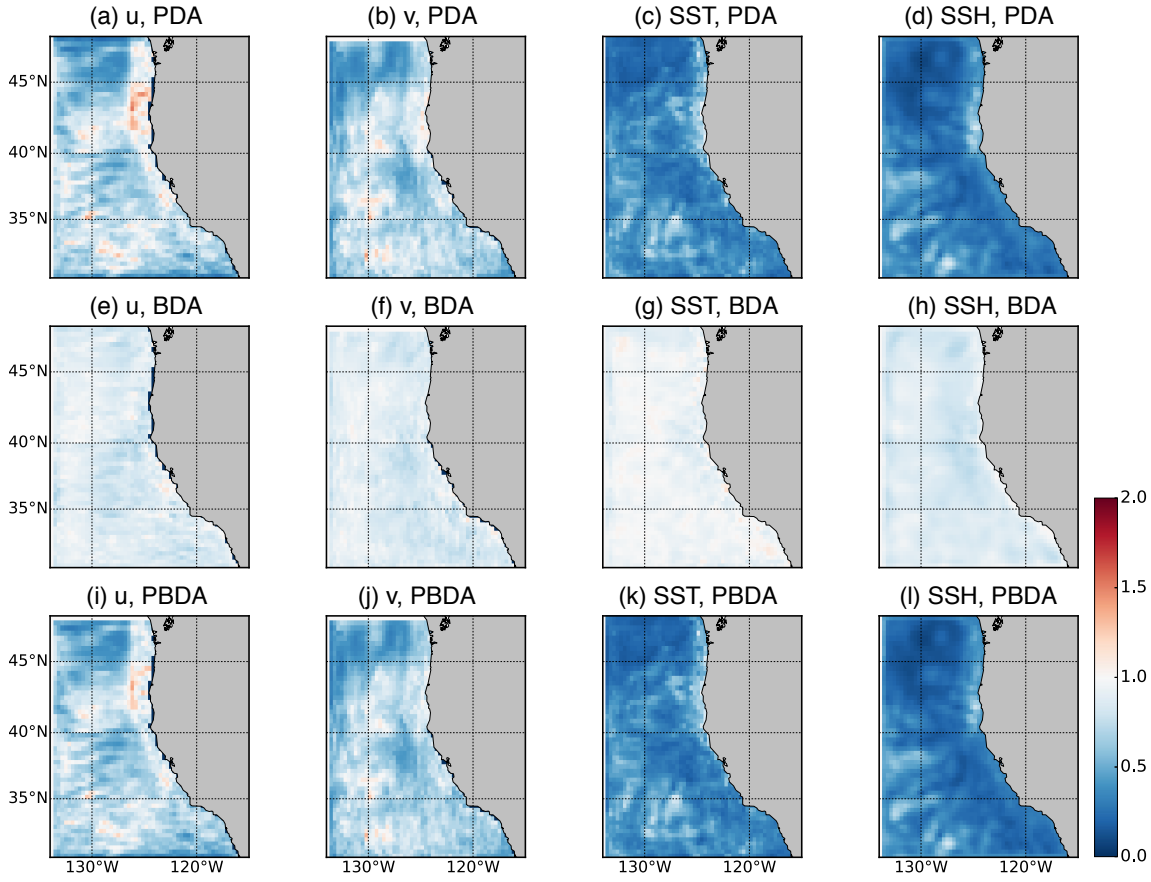

Figure 3: The ratio of the physical variables' RMSEs between data assimilation runs and free run. Smaller than 1 (cold colors) represents the reduction of the RMSE while larger than 1 (warm colors) corresponds to the increased RMSE. White areas with the value 1 mean no change in the RMSE. Top, middle and bottom rows are for PDA, BDA and PBDA, respectively, and the columns represent u, v, SST and SSH from the left to the right.

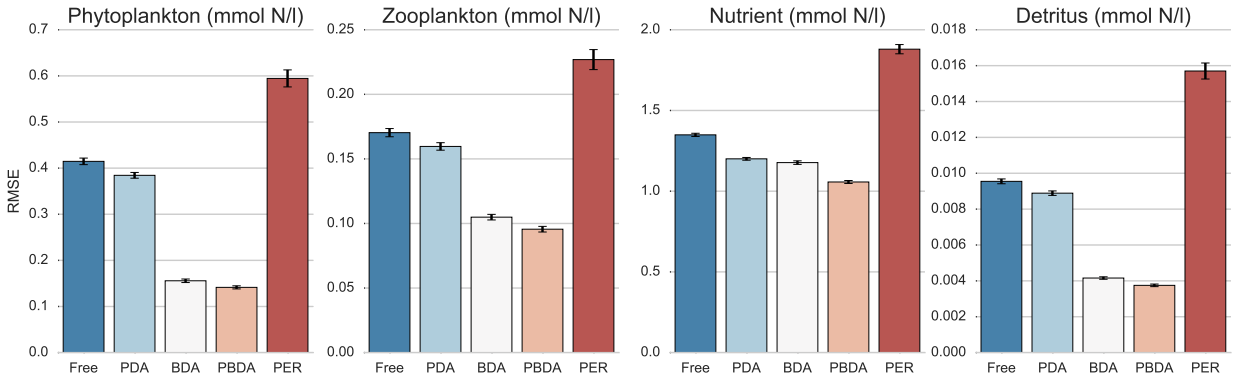

Figure 4: Same as Figure 2, but RMSE of P, Z, N and D at the surface.

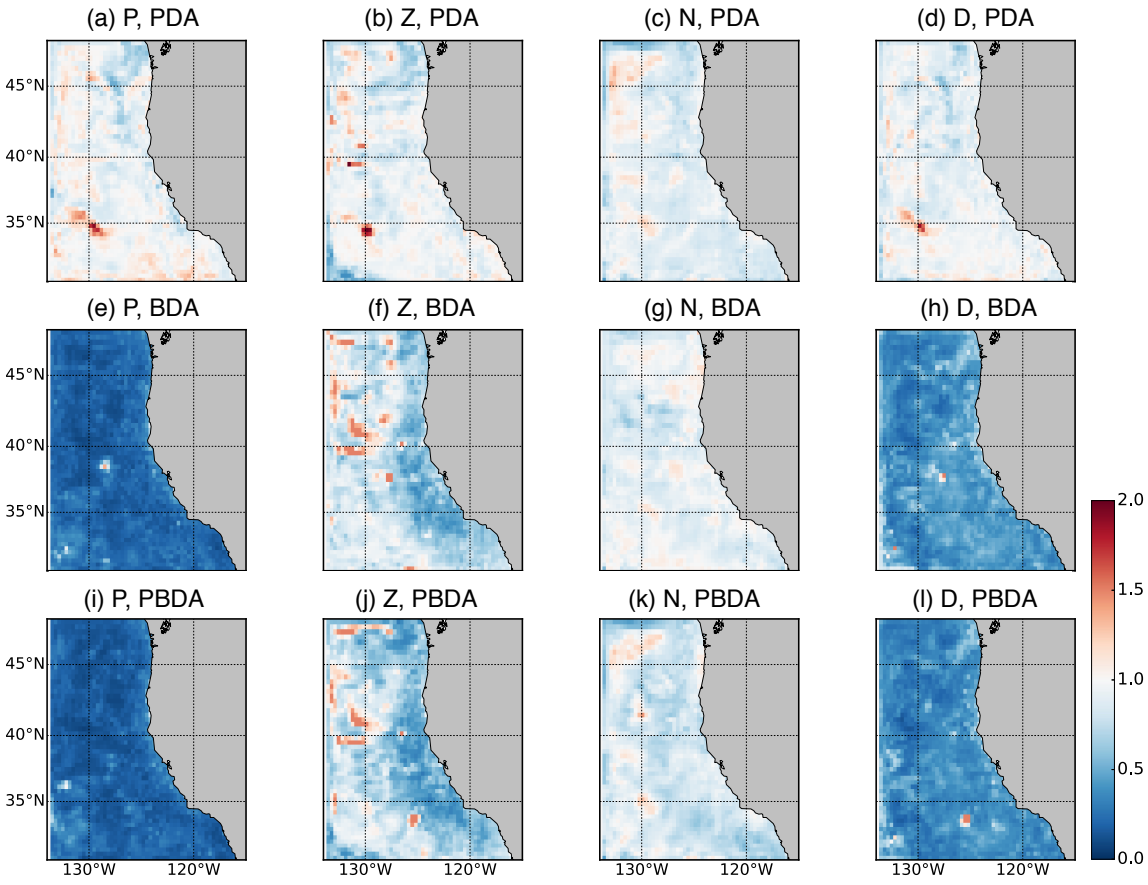

Figure 5: Same as Figure 3, but RMSE of P, Z, N and D at the surface.

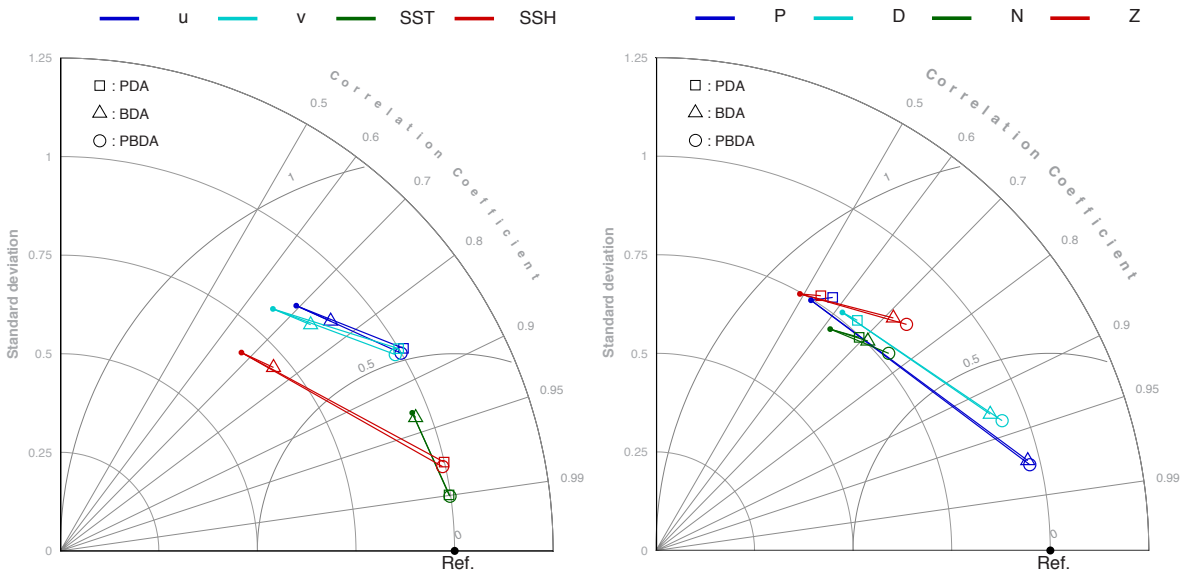

Figure 6: Taylor diagram showing the improvements of three statistical values in (left) physical and (right) biological variables. Each line shows the statistical improvement by data assimilation. The dots represent the statistical states of the prior solution, and square, triangular and circular terminators represent the statistical states of the posterior solution from PDA, BDA and PBDA, respectively. Physical variables on the right panel include u (blue), v (cyan), SST (dark green) and SSH (red). On the right panel, lines represent P (blue), Z (cyan), N (dark green) and D (red). Black dots represent the reference or true states.

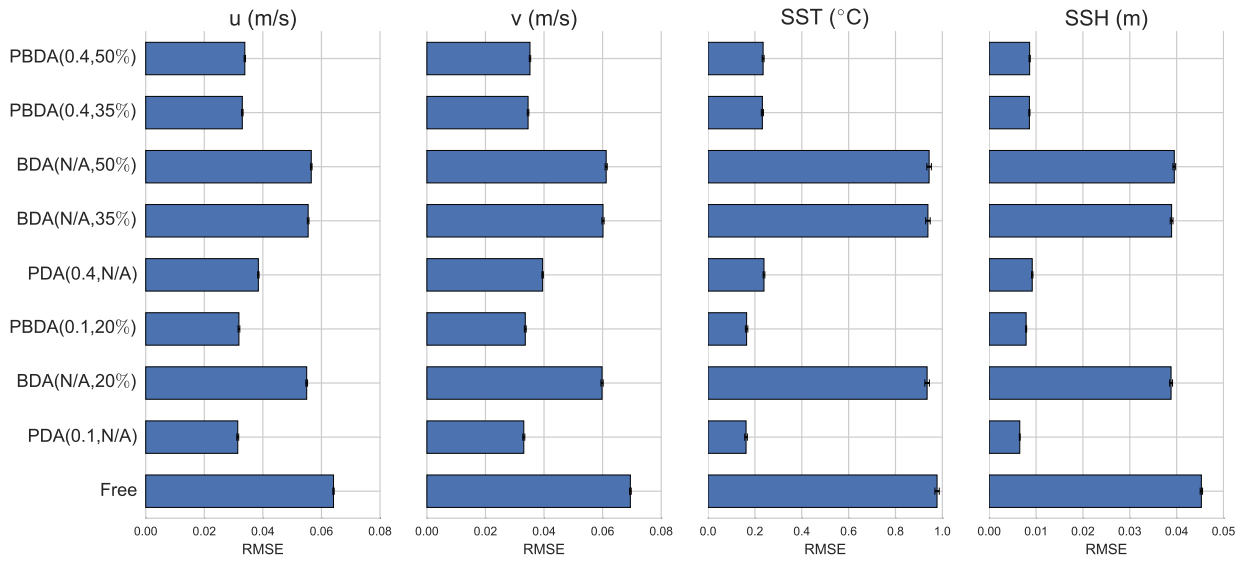

Figure 7: The RMSEs for u, v, SST and SSH (from the left to the right) in the year 2001. The RMSEs at the bottom are from the free run, overlaid by RMSEs from data assimilation runs whose observational errors for SST and phytoplankton are indicated in the parenthesis. For example, PBDA(0.4, 50%) is the data assimilation run where the SST and phytoplankton observational errors are 0.4°C and 50%, respectively. 'N/A' indicates that the corresponding observation is not assimilated.

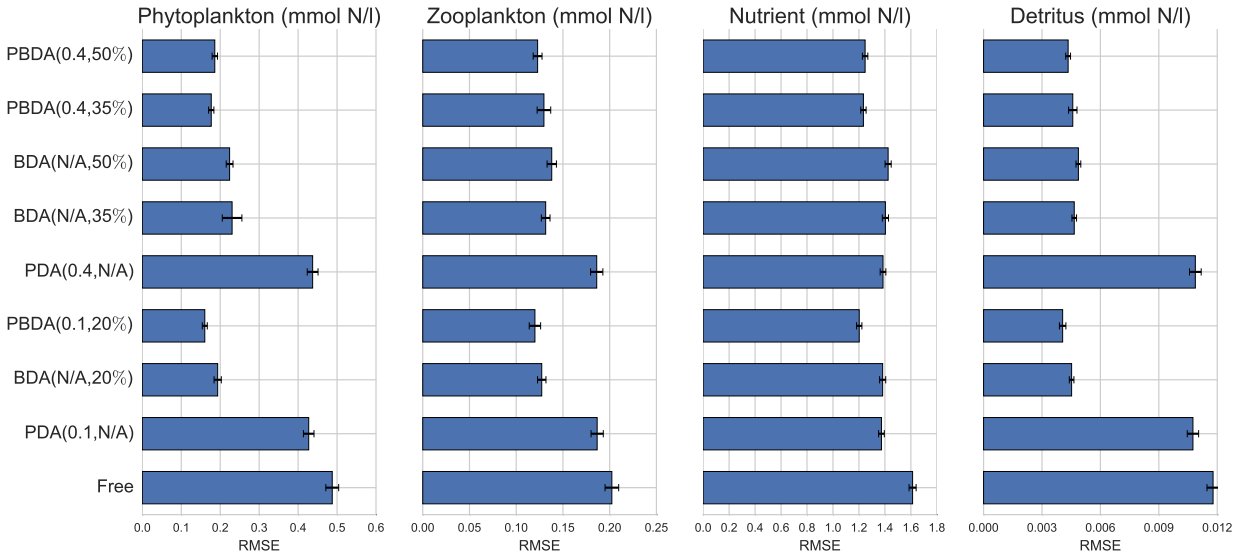

Figure 8: Same as Figure 7, but for P, Z, N and D.

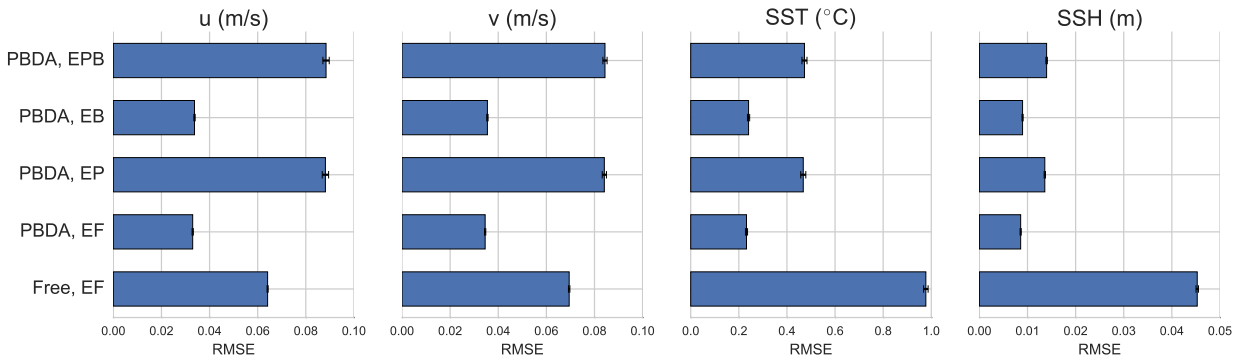

Figure 9: The RMSEs for u, v, SST and SSH (from the left to the right) in the year 2001. The RMSEs at the bottom are from the free run with no model error, overlaid by RMSEs from data assimilation runs whose observational errors for SST, phytoplankton and the label associated with the model error are indicated in the parenthesis. The labels 'EF', 'EP' and 'EPB' represent 'Error Free', 'Error in Physics' and 'Error in Physics and Biology', respectively. The physical model error is introduced by using surface forcing of the year 2002, and the biological model error comes from different biological parameter values in the assimilation runs (Table 1).

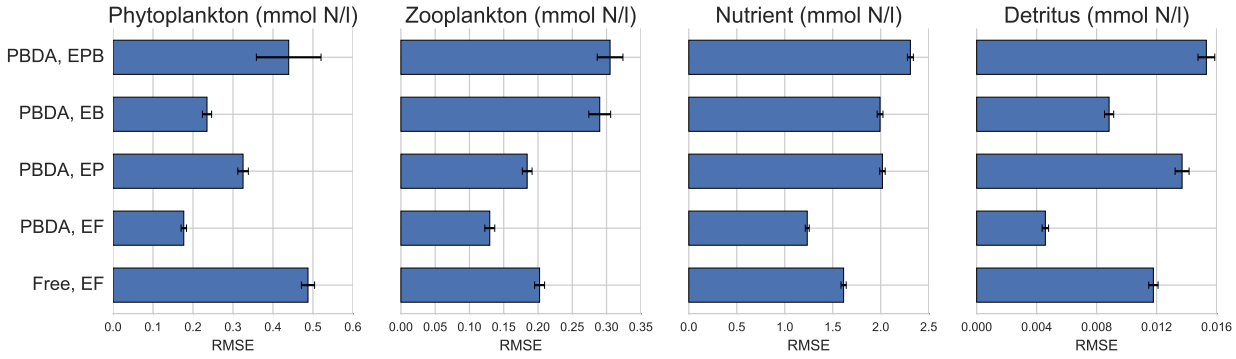

Figure 10: Same as Figure 9, but for P, Z, N and D.

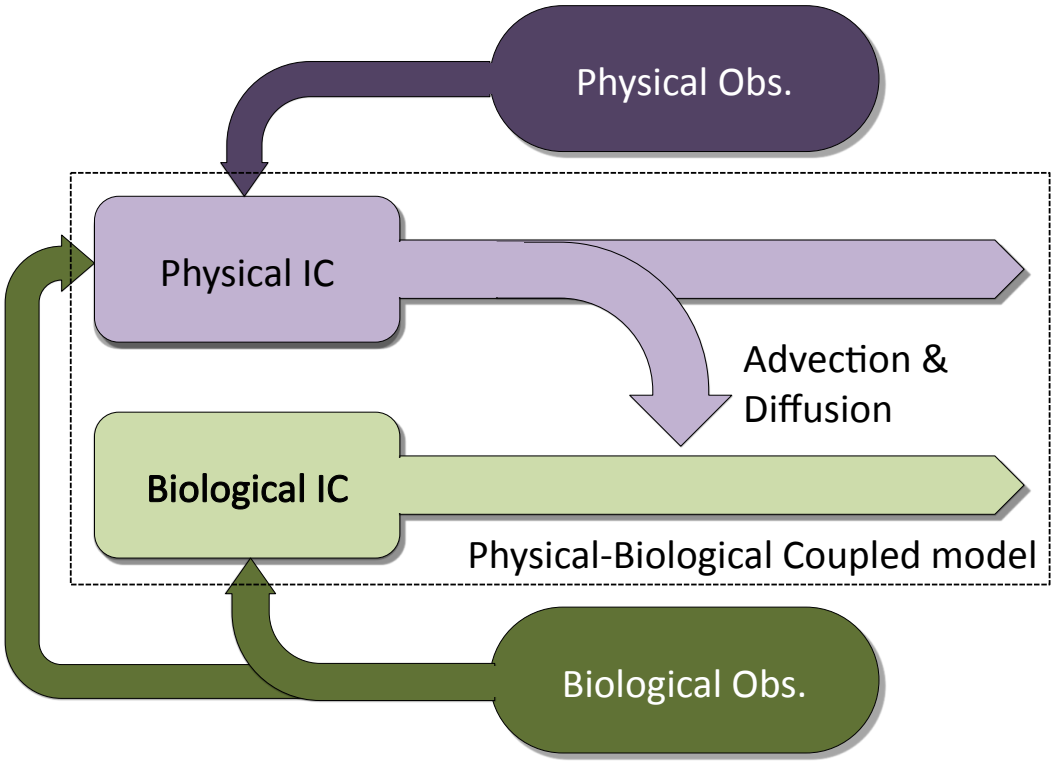

Figure 11: Diagram that shows the flow of the information from observations. Physical observations (Obs.) provide the information to adjust the physical initial condition (IC). This information is spread to biological variables through advection and diffusion. Biological observations provide the information to adjust both physical and biological initial condition. This is because the dynamics in the adjoint model pass the information only from biological component to physical component in the coupled system used in this study.