Assimilation of SeaWiFS Data into a Global Ocean-Biogeochemical Model using a Local SEIK Filter

Lars Nerger [a,b,*] Watson W. Gregg [a]

[a] *Global Modeling and Assimilation Office, NASA/Goddard Space Flight Center, Greenbelt, Maryland*

[b] *Goddard Earth Sciences and Technology Center, University of Maryland, Baltimore County, Baltimore*

Received June 22, 2006; revised November 27, 2006; accepted November 28, 2006

Abstract

Chlorophyll data from the Sea-viewing Wide Field-of-view Sensor (SeaWiFS) is assimilated into the three-dimensional global NASA Ocean Biogeochemical Model (NOBM) for the period 1998-2004 in order to obtain an improved representation of chlorophyll in the model. The assimilation is performed by the SEIK filter which is based on the Kalman filter algorithm. The filter is implemented to univariately correct the concentration of surface total chlorophyll. A localized filter analysis is used and the filter is simplified by using a static state error covariance matrix. The assimilation provides daily global surface chlorophyll fields and improves the chlorophyll estimates relative to a model simulation without assimilation. The comparison with independent in situ data over the seven years also shows a significant improvement of the chlorophyll estimate. The assimilation reduces the RMS log error of total chlorophyll from 0.43 to 0.32, while the RMS log error is 0.28 for the in situ data considered. That is, the global RMS log error of chlorophyll estimated by the model is reduced by the assimilation from 53% to 13% above the error of SeaWiFS. Regionally, the assimilation estimate exhibits smaller errors than SeaWiFS data in several oceanic basins.

Key words: Data assimilation, ecosystem modeling, Kalman filter, SEIK, Ocean color, Ocean chlorophyll

* Corresponding author. NASA/GSFC, Code 610.1, Greenbelt, MD 20771, USA. Tel. +1 (301) 614 6802, Fax. +1 (301) 614 6246

Email address: lnerger@gmao.gsfc.nasa.gov (Lars Nerger).

Article published in Journal of Marine Systems 68 (2007) 237–254

1 Introduction

Satellite ocean chlorophyll data is the only direct global-scale source of information on marine ecosystems. Routine observations have been available for a decade, and have reached a level of maturity that assimilation of the data into biological and biogeochemical models is now practical. Assimilation systems for satellite data have been shown to produce impressive results in ocean physical applications [eg., Keppenne et al., 2005, Brusdal et al., 2003, Stammer et al., 2002] and can potentially provide similar improvements in functionality and results for ocean biology. While there are many biological assimilation efforts utilizing in situ data [eg., Spitz et al., 2001, Schartau and Oschlies, 2003, Schlitzer, 2002], there are relatively few utilizing satellite ocean chlorophyll data. Variational methods have been the most common assimilation methodology for satellite ocean chlorophyll data, spanning the model range from 0-dimensional [Hemmings et al., 2003, 2004, Losa et al., 2004] through 1-dimensional [Friedrichs, 2002], to 3-dimensional [Garcia-Gorriz et al., 2003]. The emphasis on these investigations was parameter estimation. Here model parameters are adjusted to improve the model performance with regard to observations. While improvements in model parameterizations have been obtained, the parameter estimates tend to be specific for the particular model formulation and configuration. Thus, they may not be suitable for other models. In addition, the model using the estimated optimal parameters will only provide a good representation of the data, if the model formulation is able to reproduce the observational information [see e.g. Fennel et al., 2001]. However, an unsuccessful parameter estimation can point to inadequacies in the model formulation [Spitz et al., 1998].

Other work focuses on state estimation. Here the model parameters remain fixed, while the model fields are constrained by the observations to obtain improved estimates of the model fields. There are two main motivations for performing state estimation with biogeochemical models. First, the representation of assimilated variables, both (partially) observed and unobserved, can be improved by combining the best features of a model and data set. Second, more accurate derived variables in the model can be obtained, such as primary production and biogeochemical constituents. Satellite data from the Coastal Zone Color Scanner (CZCS) has been assimilated with the aim of state estimation into a 3-dimensional model of the southeast US coast by direct insertion [Ishizaka, 1990]. CZCS data has also been used in the North Atlantic with a nudging method [Armstrong et al., 1995]. Using simulated satellite data, Carmillet et al. [2001] applied a singular "evolutive" extended Kalman (SEEK) filter to assimilate simulated observations into a 3-dimensional model in the North Atlantic to study the possibilities for multivariate data assimilation. Using very accurate data with a prescribed error of 10%, Carmillet et al. [2001] were able to constrain phytoplankton as well as other fields like nitrate

and ammonium over 70 days experiment length. Using almost the same ocean-biogeochemical model, Natvik and Evensen [2003], assimilated real SeaWiFS data with an Ensemble Kalman filter (EnKF) over the period April and May 1998. In this study, the EnKF was able to improve surface phytoplankton and to reduce the variance of surface nitrate fields. In addition, subsurface nitrate and zooplankton was affected, but the changes were difficult to interpret quantitatively.

Algorithms based on the Kalman filter (KF) [Kalman, 1960], like the SEEK filter, the EnKF, or the SEIK filter used here, have several interesting properties. They directly provide dynamic error estimates of the state estimate. The error estimate is propagated throughout the assimilation period by the model dynamics. The implementation of KF-based algorithms is rather simple, as e.g. no adjoint model is required. Further, the algorithms can easily account for imperfect models. Thus, the model is not required to reproduce the observational data, but the error estimate of the filter combines observation, errors, and model errors. With regard to operational data assimilation, KF-based algorithms share the advantage of being sequential. Thus observational data can be incorporated when it becomes available, without the need to rerun the model over an extended period of model time. However, the classical linear Kalman filter as well as its first-order extension to nonlinear models, the Extended Kalman filter [see Jazwinski, 1970], have a prohibitive cost for high-dimensional models. For this reason, several algorithms based on the Kalman filter have been developed during about the last decade, which are well suited for high-dimensional numerical models and which are, to some extent, able to handle the nonlinearity of models of the ocean or atmosphere. KF-based algorithms are typically applied for state estimation. However, parameter estimation is also possible with sequential assimilation algorithms, see e.g. Losa et al. [2001].

In a recent study [Gregg, 2007] the first global long-time assimilation of SeaWiFS ocean chlorophyll data was discussed. The data was assimilated into a global 3-dimensional ocean biogeochemical model over the period of 1997-2003 using a conditional relaxation method (CRAM). The rather simple assimilation method substantially improved the estimated surface chlorophyll of the model and was able to provide daily global surface chlorophyll fields. Compared to in situ data, the assimilation resulted in a smaller bias than SeaWiFS data while the root-mean square (RMS) error was slightly higher for the assimilation than for the satellite data. In addition, the estimate of primary production was improved.

While CRAM was successful in the univariate assimilation of surface chlorophyll, it cannot be extended to a multivariate scheme which would allow to other model fields, such as nutrients in conjunction with the surface chlorophyll. As an initial effort of the application of an advanced KF-based algorithm

for state and flux estimation with a global 3-dimensional ocean biogeochemical model, we apply here a simplified form of the singular "evolutive" interpolated Kalman (SEIK) filter [see, Nerger et al., 2005a] to assimilate SeaWiFS ocean chlorophyll data over a period of 7 years into an updated version of the model used by Gregg [2007]. The simplification consists in keeping the state error covariance matrix, which estimates the error in the model state, constant analogous to the application of the SEEK filter by [Carmillet et al., 2001]. This avoids the requirement of an expensive ensemble integration necessary in a full dynamic SEIK filter. To focus on the surface total chlorophyll, the filter is applied univariately to only update the surface total chlorophyll. The effectiveness of the assimilation is analyzed by comparing back to the assimilated SeaWiFS data and by a comparison with independent in situ chlorophyll data. In addition, the influence of the assimilation on the model estimate of primary production and surface nutrients is assessed.

2 The NASA Ocean-Biogeochemical Model

The NASA Ocean Biogeochemical Model (NOBM) is a fully coupled general circulation/biogeochemical/radiative model. The general structure of the NOBM is depicted in figure 1. The three major components simulate the ocean general circulation, radiative transfer processes, and biogeochemical processes.

The ocean general circulation is modeled by the Poseidon model [Schopf and Loughe, 1995]. It is a finite-difference, reduced gravity ocean model. Here, a global configuration extending from near the South Pole to 72°N is used which includes all regions with bottom depth > 200m. The configuration uses a uniform resolution of 2/3° in latitude and 5/4° in longitude. It contains 14

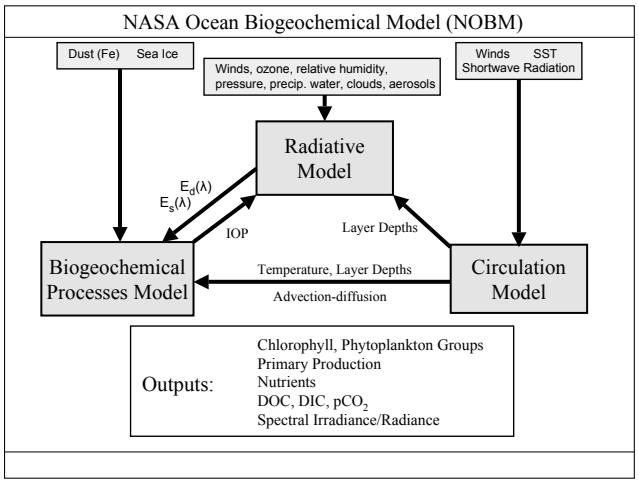

Fig. 1. General structure of the NOBM showing the interactions among the main components. In addition forcing fields and nominal outputs are shown.

vertical layers in quasi-isopycnal coordinates. The model is forced by wind stress, sea surface temperature, and shortwave radiation.

The radiative model, the Ocean Atmosphere Spectral Irradiance Model (OASIM, Gregg [2002]), provides underwater irradiance fields which drive the growth of the phytoplankton groups. The OASIM is based on the spectral model by Gregg and Carder [1990], expanded to the spectral regions 200 nm to 4 μm. It considers spectral and directional properties of radiative transfer in the oceans, and explicitly accounts for clouds. The radiative transfer calculations also interact with the heat budget. Three irradiance paths are enabled: downwelling direct and diffuse (scattered) paths, as well as an upwelling diffuse path. The oceanic radiative properties are driven by water absorption and scattering, chromophoric dissolved organic matter (CDOM), as well as the optical properties of the phytoplankton groups. The spectral nature of the irradiance is included in all oceanic radiative calculations. The forcing data sets for OASIM are shown in figure 1.

The biogeochemical processes model is described in detail in Gregg and Casey [2007]. The model consists of ecosystem and carbon components. The ecosystem component (figure 2a) contains 4 phytoplankton groups and 4 nutrient groups. In addition, a single herbivore group as well as 3 detrital pools are modeled. The phytoplankton groups have distinct growth and sinking rates, nutrient requirements. Also the optical properties, spectral absorption and scattering as well as light saturation constants, are distinct. The modeled nutrients are nitrate, regenerated ammonium, silica, and iron. Storage of organic material, sinking and eventual remineralization back to usable nutrients are simulated using the three detrital pools. The carbon component (figure 2b) simulates the interaction of dissolved organic and inorganic carbon with phytoplankton, herbivores and detritus and considers the exchange of carbon-dioxide with the atmosphere. The model uses a variable carbon:chlorophyll ratio while the carbon:nutrient ratios are constant. External forcing for the biogeochemical processes model is required in the form of atmospheric deposition of iron and sea ice fields as well as partial pressure of atmospheric CO_2.

The model is forced by transient monthly atmospheric fields. Ozone data is obtained from the Total Ozone Mapping Spectrometer. Soil dust is from Ginoux et al. [2001]. Data about cloud cover and liquid data path are obtained from the International Satellite Cloud Climatology project. The atmospheric CO_2 is from the Ocean Carbon-cycle Intercomparison Project [ICMIP, http://www.ipsl.jussieu.fr/OCMIP, derived from Enting et al., 1994] with the value for the year 2000 used as the climatological mean. All remaining forcing data is obtained from National Center for Environmental Prediction (NCEP) reanalysis products.

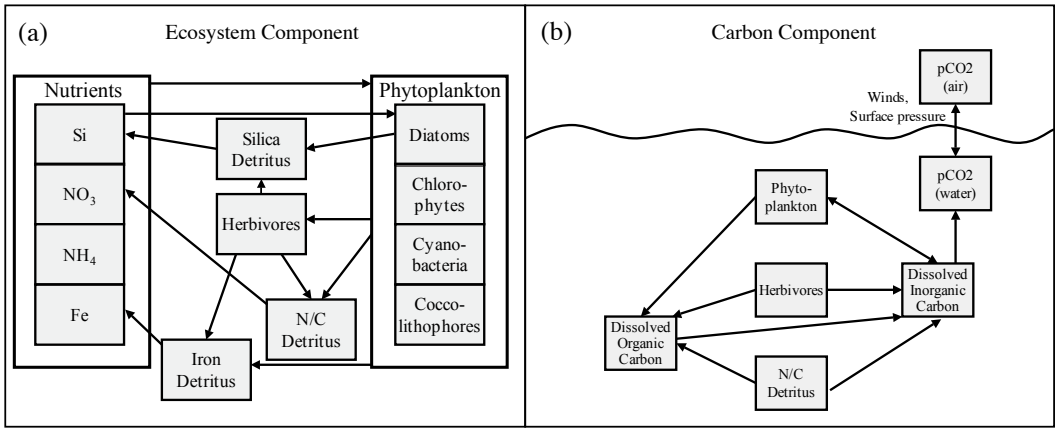

Fig. 2. Components o the Biogeochemical Processes Model: (a) Ecosystem component showing pathways and interaction between the 4 phytoplankton groups, 4 nutrients, one herbivore group and 3 detrital pools. (b) Carbon component depicting the interactions between dissolved organic and dissolved inorganic Carbon with phytoplankton, herbivores, and carbon detritus as well as exchange of CO_2 with the atmosphere.

3 Local SEIK Filter

The data assimilation is performed using the SEIK filter [Pham et al., 1998a]. The SEIK filter has been introduced as a variant of the SEEK filter [Pham et al., 1998a]. However, it has been found [Nerger et al., 2005a] that it can be considered as an ensemble-based Kalman filter which uses a preconditioned ensemble and a numerically very efficient scheme to incorporate the observational information during the analysis step. Like the SEEK filter, SEIK bases on an explicit low-rank approximation of the covariance matrix which estimates the error in the state estimate. The state correction (denoted "analysis") is computed very efficiently in the low-dimensional error sub-space which is represented by the low-rank approximated covariance matrix. In contrast, the EnKF [Evensen, 1994, Burgers et al., 1998] bases on a Monte-Carlo approach. For a detailed comparison of the EnKF, SEEK, and SEIK see Nerger et al. [2005a]. The SEIK filter algorithm demonstrated advantages over the widely used EnKF and the SEEK filter [Pham et al., 1998b] in recent studies [Nerger et al., 2005a, 2007] in which sea surface height observations were assimilated. For example, compared to the SEEK filter, the ensemble integration applied in both the EnKF and the SEIK filter showed to be better suited for nonlinear models. Compared to the EnKF, the SEIK filter requires much less computation time than the EnKF if the dimension of the observation vector is much larger than the ensemble size. In addition, the SEIK filter were able to obtain estimates with smaller errors than the EnKF.

Here, the experiments apply the localized variant of the SEIK filter [Nerger

et al., 2006] which restricts the analysis update of some horizontal location in the grid of the numerical model to consider only observations within some influence radius. The algorithm applied here is simplified by keeping the state error covariance matrix constant. Accordingly, the same error estimate for the model state is used for each analysis step. This simplification avoids the computational cost of integrating a full ensemble of model states during the assimilation process.

3.1 The (global) SEIK Filter

The SEIK filter, as other algorithms based on the Kalman filter, expresses the estimate of the state of a physical system, such as the ocean, at some time t_k in terms of the estimated analysis state vector $\mathbf{x}_k^a$ of dimension n and the corresponding covariance matrix $\mathbf{P}_k^a$ which represents the error estimate of the state vector. Being an ensemble-based Kalman filter scheme, the SEIK filter represents these quantities by an ensemble of state vectors

$$\mathbf{X}_k^a = \{\mathbf{x}_k^{a(1)}, \ldots, \mathbf{x}_k^{a(N)}\} \tag{1}$$

of N model state realizations. In this case, the state estimate is given by the ensemble mean

$$\overline{\mathbf{x}_k^a} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_k^{a(i)}, \tag{2}$$

while the ensemble covariance matrix

$$\tilde{\mathbf{P}}_k^a := \frac{1}{N-1}(\mathbf{X}_k^a - \overline{\mathbf{X}_k^a})(\mathbf{X}_k^a - \overline{\mathbf{X}_k^a})^T \approx \mathbf{P}_k^a \ , \tag{3}$$

with $\overline{\mathbf{X}_k^a} = \{\overline{\mathbf{x}_k^a}, \ldots, \overline{\mathbf{x}_k^a}\}$, is an estimate of the covariance matrix $\mathbf{P}_k^a$.

The SEIK algorithm can be subdivided into several phases and is prescribed by the following equations:
Initialization:
To initialize the filter algorithm, we assume an initial state estimate $\mathbf{x}_0^a$. Further we suppose that the initial covariance matrix $\mathbf{P}_0^a$ is estimated by a rank-r matrix which is given in decomposed form as

$$\mathbf{P}_0^a := \mathbf{V}_0\mathbf{U}_0\mathbf{V}_0^T \tag{4}$$

where $\mathbf{U}_0$ is an $r \times r$ matrix while $\mathbf{V}_0$ has size $n \times r$.

Based on these initial estimates, a random ensemble of minimum size $N = r + 1$ is generated whose statistics represent $\mathbf{x}_0^a$ and $\mathbf{P}_0^a$ exactly. For this, we transform the columns in matrix $\mathbf{V}_0$ by a random matrix with special properties. Let $\mathbf{C}_0$ be a square root of the matrix $\mathbf{U}_0$, i.e. $\mathbf{U}_0 = \mathbf{C}_0^T\mathbf{C}_0$. Then $\mathbf{P}_0^a$ can be written as

$$\mathbf{P}_0^a = \mathbf{V}_0\mathbf{C}_0^T\mathbf{\Omega}_0^T\mathbf{\Omega}_0\mathbf{C}_0\mathbf{V}_0^T, \tag{5}$$

where $\mathbf{\Omega}_0$ is a $N \times r$ random matrix whose columns are orthonormal and orthogonal to the vector $(1, \ldots, 1)^T$. The ensemble of state realizations is then given by

$$\mathbf{X}_0^a = \overline{\mathbf{X}_0^a} + \sqrt{N-1}\ \mathbf{V}_0\mathbf{C}_0^T\mathbf{\Omega}_0^T, \tag{6}$$

where each column of $\overline{\mathbf{X}_0^a}$ contains the vector $\overline{\mathbf{x}}_0^a$.

Forecast:
In the "forecast phase" the state ensemble is integrated by the numerical model to propagate the state and error estimates toward the next time when observations are available. Let $M_{i,i-1}$ be the nonlinear dynamic model operator that integrates a model state from time t_{i-1} to time t_i. Then each ensemble member $\{\mathbf{x}^{a(\alpha)}, \alpha = 1, \ldots, N\}$ is evolved up to time t_k by iterating the model equation

$$\mathbf{x}_i^{f(\alpha)} = M_{i,i-1}[\mathbf{x}_{i-1}^{a(\alpha)}] + \boldsymbol{\eta}_i^{(\alpha)}. \tag{7}$$

Here the superscript 'f' denotes the forecast while 'a' denotes the analysis. Each integration is subject to individual Gaussian noise $\boldsymbol{\eta}_i^{(\alpha)}$ which allows to simulate model errors.

For the experiments performed here, we simplify the forecast phase. For this, a matrix of ensemble perturbations ($\sqrt{N-1}\ \mathbf{V}_0\mathbf{C}_0^T\mathbf{\Omega}_0^T$ in Eq. 6) are stored and only the ensemble mean $\overline{\mathbf{x}_i^a}$ is integrated without applying the stochastic forcing $\boldsymbol{\eta}_i$. Subsequent to the integration, a forecast ensemble $\mathbf{X}_k^f$ is obtained by adding the ensemble perturbations to the forecast state $\overline{\mathbf{x}_k^f}$.

Analysis:
In the analysis phase the state and error estimates are updated on the basis of the observations, the ensemble covariance matrix, and the error covariance matrix of the observations. The SEIK filter uses a description of the covariance matrix $\mathbf{P}_k^f$ which allows for a very efficient algorithm. $\mathbf{P}_k^f$ can be computed from the state ensemble $\mathbf{X}_k^f$ in analogy to the covariance matrix in (4) according to

$$\mathbf{P}_k^f = \mathbf{L}_k\mathbf{G}\mathbf{L}_k^T \tag{8}$$

with

$$\mathbf{L}_k := \mathbf{X}_k^f \, \mathbf{T}, \quad \mathbf{G} := (N-1)^{-1} \left(\mathbf{T}^T \mathbf{T} \right)^{-1}. \tag{9}$$

Here, $\mathbf{T}$ is a $N \times r$ matrix with zero column sums, such as

$$\mathbf{T} = \begin{pmatrix} \mathbf{I}_{r \times r} \\ \mathbf{0}_{1 \times r} \end{pmatrix} - \frac{1}{N} \left(\mathbf{1}_{N \times r} \right) \tag{10}$$

where $\mathbf{0}$ represents the matrix whose elements are equal to zero. The elements of the matrix $\mathbf{1}$ are equal to one. Matrix $\mathbf{T}$ implicitly subtracts the ensemble mean when computing $\mathbf{P}_k^f$.

The analysis update of the state estimate, which is given by the mean of the forecast ensemble, can be expressed as the combination of the columns of $\mathbf{L}_k$ which span the error subspace represented by the ensemble:

$$\mathbf{x}_k^a = \overline{\mathbf{x}_k^f} + \mathbf{L}_k \mathbf{a}_k \tag{11}$$

The vector $\mathbf{a}_k$ of weights can be computed in the error subspace as

$$\mathbf{a}_k = \mathbf{U}_k (\mathbf{H}_k \mathbf{L}_k)^T {\mathbf{R}_k}^{-1} \left(\mathbf{y}_k^o - \mathbf{H}_k \overline{\mathbf{x}_k^f} \right), \tag{12}$$

$$\mathbf{U}_k^{-1} = \rho \mathbf{G}^{-1} + (\mathbf{H}_k \mathbf{L}_k)^T \mathbf{R}_k^{-1} \mathbf{H}_k \mathbf{L}_k. \tag{13}$$

Here, $\mathbf{H}_k$ is the measurement operator which computes what observations would be measured given the state $\mathbf{x}_k$. Further, $\mathbf{R}_k$ is the observation error covariance matrix and $\mathbf{y}_k^o$ denotes the vector of observations. The forgetting factor ρ, $(0 < \rho \leq 1)$ leads to an inflation of the estimated variances of the model state. It can stabilize the filter algorithm and, to some degree, account for model errors. The analysis covariance matrix is given by $\mathbf{P}_k^a := \mathbf{L}_k \mathbf{U}_k \mathbf{L}_k^T$, but does not need to be computed explicitly.

Re-Initialization:
In the re-initialization phase, the forecast ensemble is transformed such that it represents the analysis state $\mathbf{x}_k^a$ and the corresponding covariance matrix $\mathbf{P}_k^a$. Analogously to the generation of the initial ensemble it is

$$\mathbf{X}_k^a = \overline{\mathbf{X}_k^a} + \sqrt{N-1} \, \mathbf{L}_k \mathbf{C}_k^T \mathbf{\Omega}_k^T. \tag{14}$$

where a Cholesky decomposition is applied on the matrix $\mathbf{U}_k^{-1}$ to obtain $\mathbf{C}_k^{-1} (\mathbf{C}^{-1})_k^T = \mathbf{U}_k^{-1}$. The matrix $\mathbf{\Omega}_k$ has the same properties as in the initialization.

Here we shortly describe the local SEIK filter. For a detailed derivation of the local SEIK filter from the global SEIK filter see Nerger et al. [2006].

To localize the SEIK filter we restructure the analysis and re-initialization steps. We perform the operations in a loop through disjoint local analysis domains of the model grid, rather than updating the full state vector at once. For example, a local domain can be a single water column. This reformulation involves no approximation to the filter algorithm as long as all globally available observations are considered for the update of the state vector in each local domain.

For the localization of the analysis step, we neglect observations which are beyond a prescribed cut-off distance from a local domain. Since below all quantities refer to the time index k, we drop this index here for clarity of notation. Let the subscript σ denote a local analysis domain. The domain of the corresponding observations is denoted by the subscript δ. Then, the equations for the local SEIK analysis can be written analogously to the global analysis equations (11 – 13) as

$$\mathbf{x}_\sigma^a = \overline{\mathbf{x}_\sigma^f} + \mathbf{L}_\sigma \mathbf{a}_\delta, \tag{15}$$

$$\mathbf{a}_\delta = \mathbf{U}_\delta (\mathbf{H}_\delta \mathbf{L})^T {\mathbf{R}_\delta}^{-1} \left(\mathbf{y}_\delta^o - \mathbf{H}_\delta \overline{\mathbf{x}^f} \right), \tag{16}$$

$$\mathbf{U}_\delta^{-1} = \rho_\delta \mathbf{G}^{-1} + (\mathbf{H}_\delta \mathbf{L})^T \mathbf{R}_\delta^{-1} \mathbf{H}_\delta \mathbf{L} \; . \tag{17}$$

$\mathbf{H}_\delta$ is the observation operator which projects a (global) state vector onto the local observation domain. ρ_δ denotes the local forgetting factor, which can vary for different local analysis domains.

The localization of the re-initialization phase can be performed analogously to the analysis step. The local state ensemble is transformed according to

$$\mathbf{X}_\sigma^a = \overline{\mathbf{X}_\sigma^a} + \sqrt{N-1}\, \mathbf{L}_\sigma (\mathbf{C}_\delta)^T \mathbf{\Omega}^T \tag{18}$$

where $\mathbf{C}_\delta^{-1} (\mathbf{C}_\delta^{-1})^T = \mathbf{U}_\delta^{-1}$. Here, it is important that the same transformation matrix $\mathbf{\Omega}$ is used for each local analysis domain to ensure consistent transformations throughout all local domains. The rows of the ensemble matrix which correspond to a single analysis domain are transformed at once using the information from the matrix $\mathbf{U}_\delta^{-1}$ for the particular domain. This matrix corresponds to the local error subspace for the local domain and is determined by both the local state ensemble and the error covariance matrix of the local observations (Eq. 17).

Mathematically, the localization amounts to the neglect of long-range correlations in the state covariance matrix during the analysis step, see Nerger et al. [2006]. Viewed globally, the neglect of long-range correlations increases the rank of the covariance matrix and hence leads to a larger dimension of the error-subspace in which the update of the state estimate is computed. This larger dimension is only considered implicitly during the independent analysis updates in the local domains, because the state covariance matrix is never computed explicitly. The rank of the covariance matrix represented by the global state ensemble does not increase, since this rank depends on the ensemble size N and can be at most $N-1$. The re-initialization transforms each local state ensemble using the same matrix $\mathbf{\Omega}$. This consistent re-initialization results in an ensemble which represents a covariance matrix with the same rank as the covariance matrix of the forecast ensemble.

The localization by neglecting observations beyond the cut-off distance can be combined by a localizing weighting of the observations. For this the inverse variance estimates in the inverse local observation error covariance matrix $\mathbf{R}_\delta^{-1}$ for each local domain are reduced by a factor depending on the distance of the observation from the local analysis domain. A possibility is an exponential decrease, which reduces the influence of observations with growing distance from the local analysis domain. This weighting method is similar, and under some circumstances equivalent, to the method of "covariance localization" which involves the element-wise product of the state error covariance matrix by a correlation matrix holding correlations of compact support [Houtekamer and Mitchell, 2001].

4 SeaWiFS Ocean Chlorophyll Data

The experiments assimilate global chlorophyll data from SeaWiFS. Daily fields, of data set version 5.1, at 9km resolution have been obtained from the NASA Ocean Color Web site. The data fields have been remapped to the model grid for the assimilation.

The assimilation is performed daily at model midnight to reduce sampling errors. Clouds, sun glint, inter-orbit gaps, and high-aerosol concentrations obscure remote observations, producing prominent and sometimes persistent gaps in satellite data, especially at daily time scales. Thus, the underlying complete fields provided by the model, adjusted by the assimilation of observations where and when available on a daily time scale, alleviates the sampling problems incurred using remote-sensing data alone. A typical daily data set is shown in figure 3. The inter-orbit gaps as well as the gaps resulting from sun glint and the austral polar night are clearly visible. In addition, clouds obscure several regions like wide areas of the North Pacific. Typically, there are be-

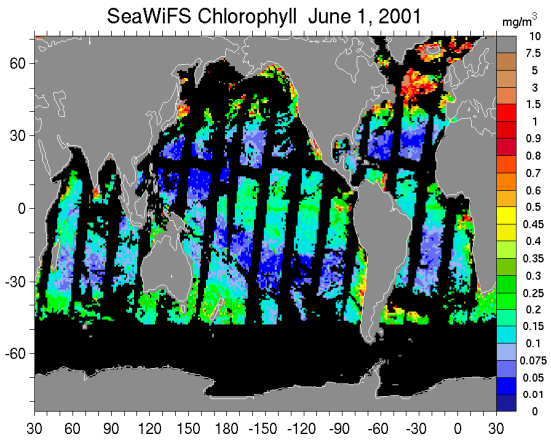

Fig. 3. SeaWiFS chlorophyll data for June 1, 2001. Grey indicates land and coast while black indicates missing data. Visible are the inter-orbit gaps as well as gaps resulting from sun glint, clouds and the austral polar night.

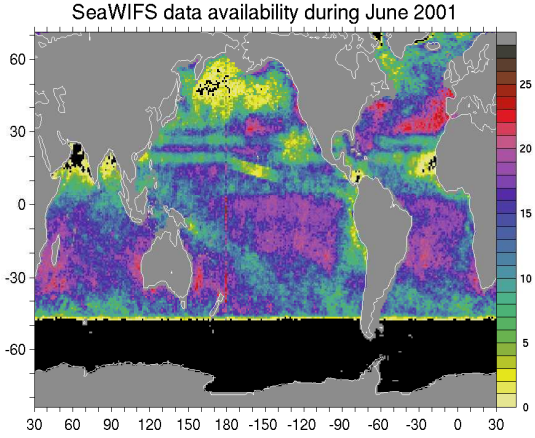

Fig. 4. Availability of SeaWiFS chlorophyll data in days during June 2001. No Data is available south of about 58°N due to the polar night. Noticeable is the region in the North Pacific where only very few data is available due to clouds as well as the small data availability in the Arabian Sea and offshore Mauritania.

tween 13000 to 18000 observed grid points daily. Due to clouds the sampling frequency of the data is irregular, as is visible from figure 4 which shows the amount of data per grid point available during June 2001. Caused by longer persistent clouds, the amount of data is strongly reduced in several regions. This is most noticeable in the Arabian Sea, where for most grid points no data was available at all during June 2001. In addition, a wide region of the North Pacific north of 40°N was obscured by clouds during almost all of this month. Due to this irregular temporal sampling, there will be regions which are not influenced by the data assimilation for significantly longer periods than the daily assimilation interval. This presents a challenge to the assimilation process, with larger errors of the model state being estimated.

For the assimilation all daily SeaWiFS chlorophyll data with concentrations larger than twice the monthly mean are considered as outliers and excluded. In addition, data are excluded which occur within a model grid point containing ice. These exclusions are motivated by the fact the remote sensing errors are typically expressed as overestimates as the most dominant error sources, absorbing aerosols, CDOM, sub-pixel scale clouds and ice most often lead to overestimates of chlorophyll. These overestimates can have a very deleterious effect on the quality and stability of the assimilation process.

The analysis equations of the Kalman filter assume a normal distribution of the state vector. The distribution of chlorophyll and errors in chlorophyll are assumed to be log-normal [Campbell, 1995]. Accordingly, the assimilation is performed on the logarithm of the observed and modeled chlorophyll concentrations.

The observation errors are assumed to be independent. Thus the observation error covariance matrix $\mathbf{R}_k$ is diagonal. Frequently a global error estimate of 35% is used for SeaWiFS chlorophyll data [see, e.g., Natvik and Evensen, 2003], because this accuracy was a major objective of the SeaWiFS project [Hooker et al., 1992]. However, the comparison of SeaWiFS chlorophyll data with independent in situ data shows significant variations around this estimate [Gregg and Casey, 2004] which ranged between 13% and 56% with a global mean error of 31%. Motivated by this study, the experiments here use regionally varying errors for the observations, similar to the weighting approach applied by Gregg [2007]. For June 1, 2001 the errors are shown in figure 5. These error estimates are not identical with those reported by Gregg and Casey [2004], but are chosen to minimize the estimation errors in the assimilation. The error in the North and Equatorial Indian Oceans is chosen to be larger motivated by the prevalence of light-absorbing dust [Wang et al., 2005]. This problem also occurs in the tropical Atlantic. Here, a special condition is assumed for the Mauritanian offshore region (region B in figure 5) where the error estimate is increased for larger satellite chlorophyll concentrations. Namely, the error is set to 0.8 for grid points with C(sat) between 1mg m^{-3} and 2mg m^{-3}, and to 5.0 for grid points with C(sat) $>$ 2mg m^{-3}. This approach has also been followed for the Amazon and Congo river outflow regions (regions A and C in figure 5, respectively). These regions are dominated by CDOM which produce erroneous chlorophyll values in the satellite data. Here an error of 0.8 is set for grid points with C(sat) between 1mg m^{-3} and 2mg m^{-3} and an error of 1.2 is specified when C(sat) $>$ 2mg m^{-3}. Using these special conditions also avoids the occurrence of negative concentrations, e.g. in nutrients, during the model integration which could occur as a reaction of the model dynamics on too large changes in the surface chlorophyll.

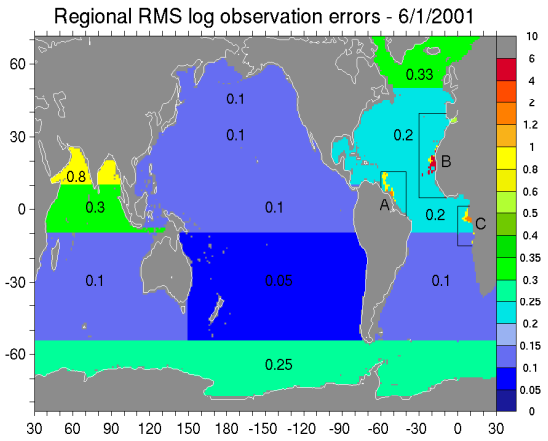

Fig. 5. Observation errors assumed for the assimilation of chlorophyll data on June 1, 2001. The regions A to C follow special conditions based on the value of the satellite data. Here a larger error is assumed for particularly large concentrations.

5 Data Assimilation Experiments

The SeaWiFS ocean chlorophyll data is assimilated daily into the NOBM over the period of seven years from 1998 to 2004. We will first compare the estimated total chlorophyll fields from the assimilation with the SeaWiFS data. Subsequently, we discuss the influence of the assimilation on primary production and on the nitrate fields which are not directly affected by the assimilation. Finally, we compare the assimilation results to independent in situ data.

5.1 Experimental setup

In the experiments global daily chlorophyll observations from SeaWiFS are assimilated into the NOBM at model midnight. Only the 4 phytoplankton groups in the surface layer are updated by the filter algorithm. Since only total chlorophyll is observed, the sum of the chlorophyll concentrations of the four phytoplankton groups is used as total chlorophyll of the model state. After adjusting the total chlorophyll concentration by the filter algorithm, the phytoplankton groups are updated under the constraint that their relative abundances remain unchanged.

The data assimilation process is initialized by a state estimate for January 1998 obtained from a spin-up run over 20 years with monthly climatological forcing. The initial state error covariance matrix $\mathbf{P}_0^a$ for the logarithm of the total chlorophyll concentration is estimated from a free model run over the 8 years 1997 to 2004 with monthly forcing data. One time slice per month is retained resulting in 96 state vectors. The decomposition according to equation

(4) is then obtained by the singular value decomposition of the perturbation matrix which holds in its columns the deviations of each single state vector from the 7-year mean state. This procedure directly yields the eigenvector matrix $\mathbf{V}_0$ and the square-roots of the eigenvalues which build the diagonal of the matrix $\mathbf{C}_0$ in equation (5). For the data assimilation experiment the leading 10 eigenvectors and eigenvalues are used to generate the state ensemble. This initialization of the covariance matrix is similar to that used by [Carmillet et al., 2001] and other assimilation studies which applied the SEEK or SEIK filters.

The simplified variant of the local SEIK filter is implemented within the parallel data assimilation framework PDAF [Nerger et al., 2005b] which provides fully-implemented filter algorithms which can be connected to existing models to generate a data assimilation system. In the experiments, only the ensemble mean state is propagated by the model without applying any stochastic forcing to the integration. The forgetting factor is set to one. For the localization, a small cut-off distance of 5 grid points in zonal and meridional directions is used to define rectangular local observation domains. The localizing weighting of the observation is performed by an exponential decrease with a length scale of 1 grid point to reduce the variance estimate by a factor of 1/e.

5.2 Estimated Total Chlorophyll Concentrations

Daily snapshots for June 15, 2001 of the total surface chlorophyll field from the free-run model and the assimilation are shown in figure 6. In addition, the SeaWiFS chlorophyll field for this day is shown. The free-run model shows a broad agreement with the satellite data. The agreement is significantly improved by the assimilation. In particular, the chlorophyll concentration is reduced in the Equatorial and South Pacific oceans. The concentrations of the blooms in the North Atlantic and North Pacific are increased, while the spatial extent of the bloom region in the North Pacific is reduced. These changes are in agreement with the SeaWiFS data. In addition, the assimilation provides a complete daily coverage of total chlorophyll which is obtained by a combination of extrapolating the satellite data within the local analysis domains into the data gaps and by propagating previous information by the model dynamics.

Figure 7 shows the monthly mean of daily differences between the logarithms of the chlorophyll fields from the assimilation and SeaWiFS data during June 2001. The difference between the assimilation estimate and the satellite data is generally small (below 0.05). There are regions with larger deviations which correspond to the regions in which an increased error of the SeaWiFS data was assumed. These regions are the North and Equatorial Indian Oceans, as well as near the Congo mouth, offshore Mauritania and north of mouth of the

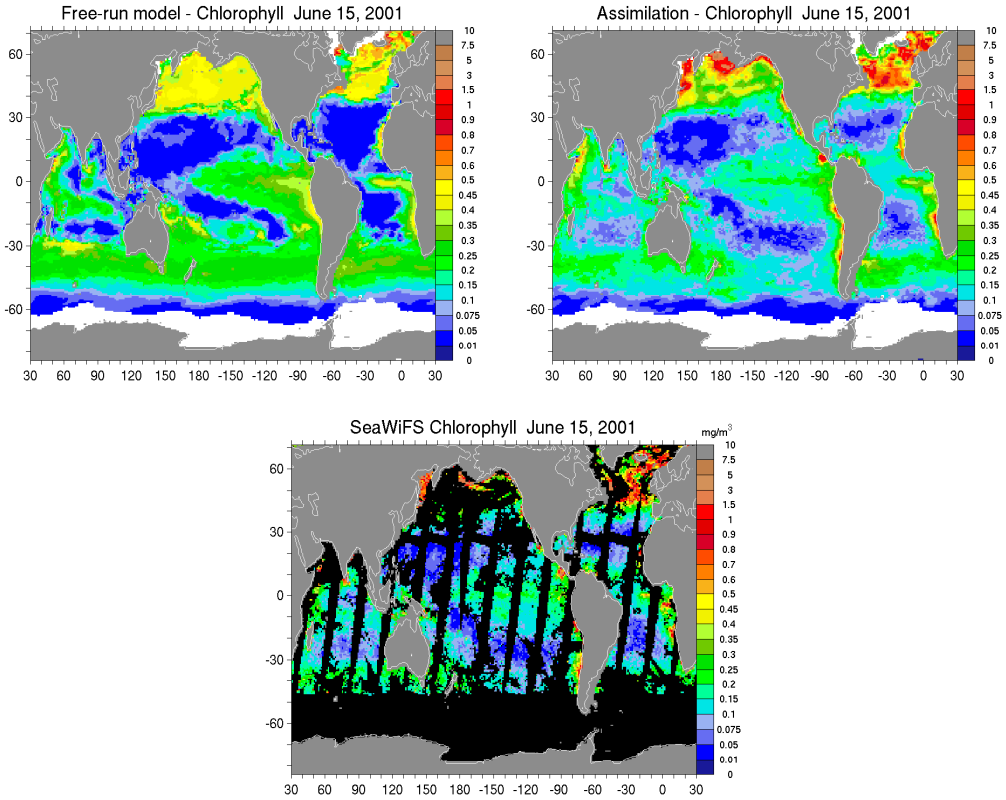

Fig. 6. Total surface chlorophyll for June 15, 2001 from the free-run model (upper left), the assimilation (upper right), and SeaWiFS (lower panel). White indicates sea ice. The assimilation significantly improves the chlorophyll estimate of the free-run model which shows broad agreement with SeaWiFS data.

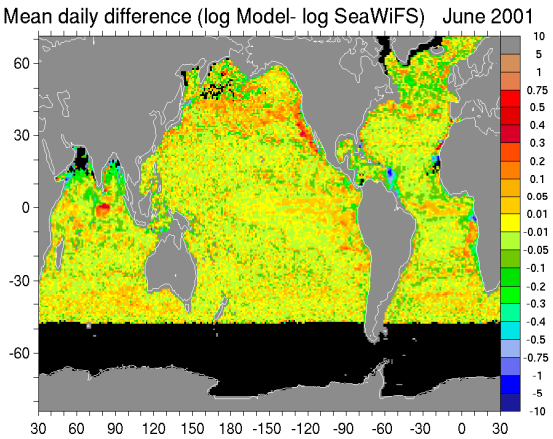

Fig. 7. Monthly mean of daily differences between the logarithms of chlorophyll concentration from the assimilation and SeaWiFS data.

Amazon.

Noticeable differences are also visible in the North Pacific and the North Atlantic Oceans. These differences with values up to about ± 0.2 lie in bloom regions with high chlorophyll concentrations. These misfits between the as-

similation and the satellite data have several reasons. For the North Pacific near the Bering strait, there is only a very limited amount of satellite data available during June 2001 as is visible in figure 4. Over a wide region data is available on less than 4 days during this month. Accordingly, the model is less constrained by the data which results in larger misfits between the assimilation estimate and the SeaWiFS data. Over the North Atlantic the availability of SeaWiFS data is generally higher and the model is more constrained by the data. However, in this region we assumed an error of the observations of 0.33 north of 50°N and 0.2 otherwise. Thus, the mean differences visible in the North Atlantic are smaller than the assumed errors in the observations.

Near the coast of California a band is visible in which the assimilation overestimates the total chlorophyll concentration from SeaWiFS. This deviation is due to the choice of the state error covariance matrix described in section 5.1. The logarithmic variance estimate of this matrix exhibits a small variance between 0.01 and 0.05 near the coast of California. Accordingly, the assimilation method considers the model to be very accurate in this region. This leads to a smaller influence of the satellite data here which results in a larger misfit between assimilation estimate and satellite data.

5.3 Primary Production and Nutrients

The univariate assimilation leads only to a direct update of the concentrations of phytoplankton groups at the surface. However, variables that are directly related to chlorophyll, such as primary production (PP), export, carbon:chlorophyll ratios, growth rates, and irradiance in the water column are affected as well in a positive manner by the univariate assimilation. Other state variables and processes are not directly affected, but will react on the changed chlorophyll fields. To examine the effects of the univariate assimilation we focus here on the primary production and the nutrients at the surface exemplified by nitrate.

PP is a flux quantity which is computed in the model as the depth-integrated growth rate multiplied by the carbon:chlorophyll ratio. Here, we focus on the annual total PP. We compare the PP estimated by the model with PP estimated directly from satellite data. A common algorithm to compute PP from satellite data is the Vertically Generalized Production Model [VGPM, Behrenfeld and Falkowski, 1997]. Next to chlorophyll, sea surface temperature (SST) and photosynthetically available radiation (PAR) are required as inputs by the VGPM. For the comparison, SeaWiFS chlorophyll is used. In addition, SST is used from the same source as used for model forcing. The atmospheric component of OASIM in the wavelength region 350-700 nm provides PAR. Figure 8 shows annual PP estimated by the free-run-model, the assimilation,

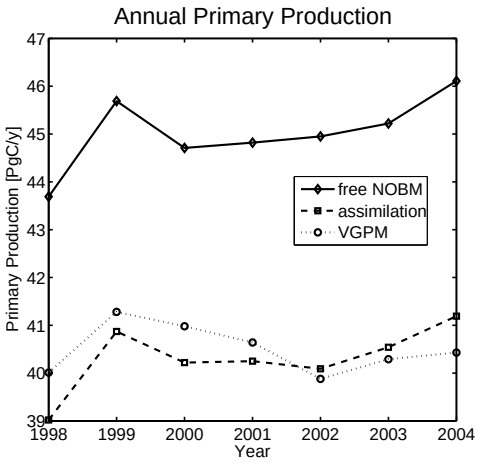

Fig. 8. Annual primary production over the period from 1998 to 2004 for the free-run model (blue), the assimilation (green), and the Vertically Generalized Production Model (VGPM, black).

and the VGPM for the years 1998 to 2004. The estimate from the free-run model is on average 11.2% higher that the estimate from VGPM. This deviation is larger than that reported by Gregg and Casey [2007] because of their use of climatological forcing while transient monthly forcing is used here. The assimilation succeeds in providing an estimate of PP which is consistent with that from the VGPM. However, the assimilation estimate still shows the same variability signature as the free-run model. On average, the PP estimate from the assimilation is 0.5% lower than the VGPM estimate.

The nutrient concentrations in the model are not directly modified by the univariate assimilation, but they react on the changed surface chlorophyll concentrations during the model integration. Thus, no systematic improvement of the nutrients can be expected and the reaction of the nutrients can lead to regional improvement or deterioration of the fields. For the stability

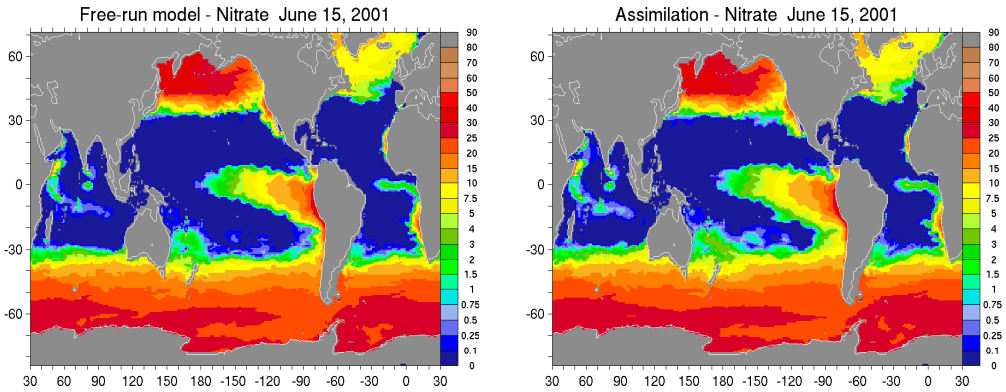

Fig. 9. Surface nitrate for June 15, 2001 from the free-run model (left) and the assimilation (right). White indicates sea ice. The nitrate field reacts to the assimilation with increased concentrations in several regions.

of the assimilation process, it is important that possible negative influences on the nutrients do not destroy the stability of the model integration, e.g. by negative nutrient concentrations as discussed in section 4. The nitrate field at the surface is shown in figure 9 for the free-run model and the assimilation for June 15, 2001. The assimilation resulted only in small changes in the nitrate concentrations compared to the free-run model. Most noticeable, the nitrate concentration in increased in the South Pacific between 35°S and 20°S. Overall, the assimilation leads to a small degradation of the nitrate field compared, e.g., to climatology. The effects of the assimilation on the other nutrients are similar and thus not shown here.

5.4 Comparison with Independent Data

The comparison of the assimilation estimate with the assimilated SeaWiFS data shows that the assimilation method works as expected. However, a real test of the efficiency of the assimilation relies in the comparison to independent data. This will be performed here by comparing the assimilation estimate with independent in situ data of total surface chlorophyll concentrations. The in situ chlorophyll data has been obtained from the SeaWiFS Bio-Optical archive and Storage Systems [SeaBASS, Werdell and Bailey, 2002] and the NOAA/National Oceanographic Data Center (NODC)/Ocean Climate Laboratory (OCL) archives [Conkright et al., 2002] which provides chlorophyll data from fluorimetric measurements. For the comparison, daily in situ data was mapped to the model grid by computing the average of measurements within each single grid cell.

In the following we discuss RMS log errors and bias of log quantities for the comparison with in situ data globally and separated over the 12 ocean basins. The basins are defined as follows: The Antarctic basin is considered to be south of 40°S. The southern basins lie between the Antarctic basin and the equatorial basins which extent from 10°S to 10°N. The North Indian as well as the North Central Pacific and Atlantic basins are located north of 10°N. The latter two basins have a northern boundary at 40°N. The North Pacific and North Atlantic basins are located north of this latitude. We show RMS and bias of the logarithmic value because of the log-normal distribution of the chlorophyll concentrations. Due to this, the errors show a normal distribution on log quantities and the distribution can be quantified consistently by a state estimate and an additive error which is symmetric with respect to the state estimate. However, to visualize the actual deviations from the in situ data, figure 10 shows the actual error of the assimilation estimate with respect to in situ data in the Pacific north of 30°S averaged over the years 1998-2004. Note, that the temporal averaging of errors of actual values which are log-normally distributed can lead to misleading results, if the values

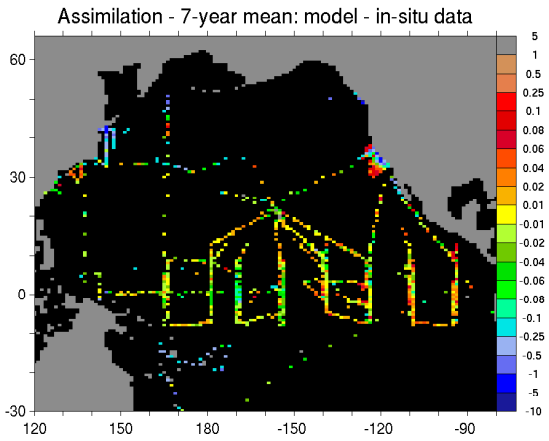

Fig. 10. Mean errors over seven years between the assimilation and in situ data.

vary strongly. However, the majority of the grid points is only observed once in which case the actual difference is shown. The Equatorial Pacific region exhibits a systematic large-scale sampling which follows the Tropical Ocean-Global Atmosphere program/Tropical Atmosphere Ocean (TOGA/TAO) array. Here up to 10 measurements at the same model grid point are available over the 7 year period of the comparison. The error of the assimilation in this region is very small with some overestimation of the concentrations in the eastern part and underestimation in the western part. In the other basins the large-scale sampling is quite irregular and only a single observation is available for most locations during the comparison period. A large amount of data is available in the North Central Pacific. However, the data is dominated by data from the CalCOFI (California Cooperative Oceanic Fisheries Investigations) project near the coast of California which accounts for about 69% of the data in the North Central Pacific. Significant mean errors are visible in the region of CalCOFI. Near the cost, the assimilation underestimates the in situ data up to about 3 mg/m^3. Distant from the coast the assimilation overestimates the in situ data by up to 0.3 mg/m^3. The reason for these larger errors in the assimilation estimate are the small estimated variances in this region as has been discussed in section 5.2. In the remaining North Central Pacific, the errors are rather small and both overestimates and underestimates occur. In the North Pacific the in situ data is mostly underestimated by the assimilation with largest errors near the coast. This effect results from the fact that only regions with bottom depth $>$ 200m are included in the model. In the South Pacific, in situ data is underestimated in the eastern part of Melanesia (between 10°S - 30°S and 160°E - 170°W). The reason for this are discussed in conjunction with the RMS log errors below.

Figure 11 shows the RMS log errors and the bias of the log quantities for the comparison with in situ data globally and separated over the 12 ocean basins. Shown are the errors for the chlorophyll estimate from the assimilation, the free run model, and SeaWiFS chlorophyll. The number of comparison points

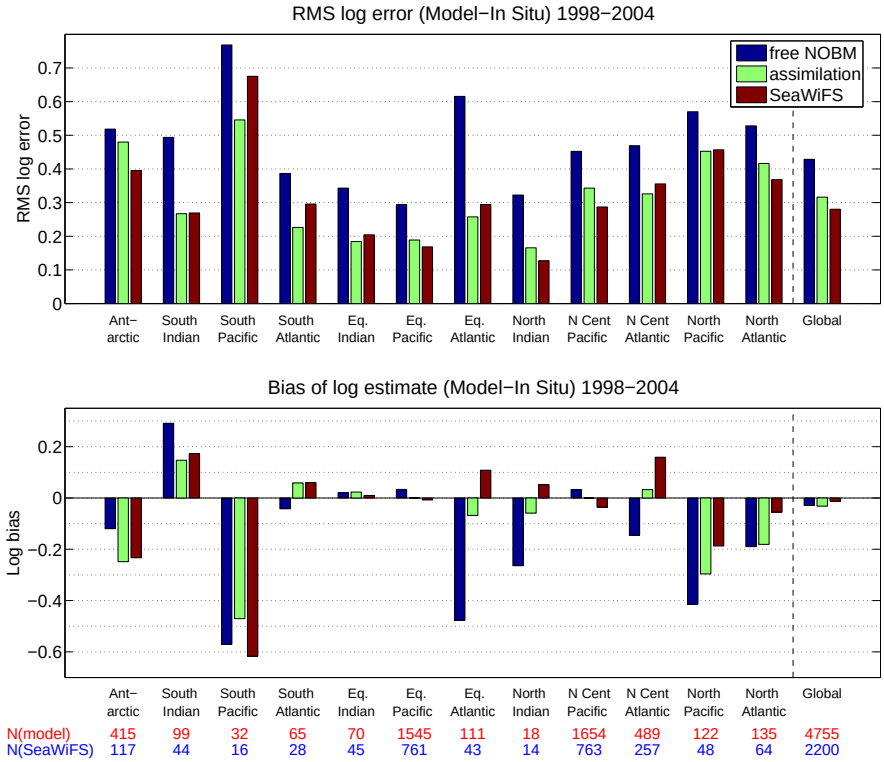

Fig. 11. Upper panel: RMS log error between model or SeaWiFS data and in situ data separated over the 12 major ocean basins and globally. Lower panel: bias of the log quantities for the comparison with in situ data. Shown are values for the free-run model (blue), the assimilation estimate (green), and SeaWiFS data (red). At the bottom the number of comparison points for the model and SeaWiFS data are listed.

is listed for each basin for the model-in situ data and satellite-in situ data comparisons. For the comparison of the model fields - from the free-run and the assimilation - with the in situ data more than twice the number of comparison points were available than for the comparison between satellite and in situ data. This is due to the gaps in the daily satellite data in contrast to the complete coverage of the model output. The availability of in situ data varies strongly between different basins. The basins with the largest amount of data are the Equatorial Pacific and the North Central Pacific basins. However, only the Equatorial Pacific shows a systematic large scale sampling.

Globally, the improvement of the surface chlorophyll field by assimilation of SeaWiFS data is well visible. The RMS log error is reduced from 0.43 for the unconstrained model to 0.32 with assimilation. However, the RMS log error of SeaWiFS data is smaller at 0.28. Thus, the assimilation reduces the global RMS log error from 53% above the error of SeaWiFS to 13%. When we consider only in situ data points collocated with the satellite data, the free run model error is about 51% and the assimilation error about 8% larger than

the error of SeaWiFS data. The larger assimilation error for the comparison involving all in situ data points shows that the information transfer into data gaps is not free of errors.

Regionally, the assimilation estimate shows smaller RMS log errors than SeaWiFS data in several basins. In particular the Atlantic basins, except for the North Atlantic (north of $40°N$), are better represented by the assimilation estimate than by the SeaWiFS data. In addition, the Equatorial Indian Ocean and the South Pacific show lower RMS log errors for the assimilation than for SeaWiFS data. The errors are generally smaller in the equatorial regions than for the northern basins for both the model and SeaWiFS. An exception for this is the North Indian Ocean. The very small error for the SeaWiFS data in this basin is due to sampling error caused by the very small amount of in situ data. As described in section 4, it is known that light-absorbing dust in the North Indian Ocean can result in overestimates of the chlorophyll concentration by the satellite [Wang et al., 2005]. Apparently, in situ data was only available at times or locations when and where this problem did not exist.

Over the whole North Central Pacific the assimilation estimate has an error which is about 20% larger than the error of SeaWiFS. As noted before, the data in this basin is dominated by the CalCOFI project. If we separate the basin into the region containing the data from CalCOFI and the remaining North Central Pacific, we obtain errors of which are about 23.4% and 7.9% larger than the SeaWiFS error, respectively. Thus, while the assimilation performs quite well in the most part of the North Central Pacific, it's performance is inferior in the small CalCOFI region. This is also reflected by the actual mean errors shown in figure 10.

The South Pacific exhibits the largest errors of all basins, both for the SeaWiFS data and the free-run and assimilation model. These errors are mainly caused by a large bias as is evident from the lower panel of figure 11. The chlorophyll concentrations in the eastern part of Melanesia are strongly underestimated by both the model and SeaWiFS. Just to the north of this region, Messié et al. [2006] found very high chlorophyll values near the Kiribati Islands (170°E, 0°N). This also occurred at the same time as the observations in our data set, May 1998, corresponding to the switch from El Niño to La Niña. Messié et al. [2006] suggested the blooms were caused by the topographic effects of the islands on the circulation patterns, and thus the nutrient fields, associated with the shift in the El Niño Southern Oscillation. These dynamics are similar to the eastern Melanesia observations, at the same time. However, our model resolution was unable to capture the dynamics. The re-gridded SeaWiFS data to the model grid shows some elevated chlorophyll concentrations up to about 0.4 mg/m^3 at single grid points next to islands. However, these high-value points were to sparse to have a significant effect in the assimilation and were not collocating with data in the used in situ data set. If the in situ data in this

region is removed from the comparison, 9 collocation points for the model and 2 for the SeaWiFS data remain. In this case the log bias is reduced to -0.21 for SeaWiFS, -0.13 for the assimilation, and -0.05 for the free-run model. The RMS log error is reduced to 0.51 for the free-run model, 0.21 for SeaWiFS, and 0.16 for the assimilation. While this comparison only included very limited collocation points, it indicates that the assimilation strongly improves the model estimate also in the South Pacific.

The global log bias is very small for the free model (-0.030) and the assimilation (-0.032), and even smaller for SeaWiFS data (-0.012). In most basins, the assimilation effectively reduces the bias of the free model. A noticeable exception from this is the Antarctic Ocean where the bias is amplified from about -0.12 to -0.25. For the comparison with in situ data, this region is rather problematic. This is due to the fact that the sea ice coverages considered in the model and by SeaWiFS can be distinct from the real coverage which limits the in situ measurements. Accordingly, in situ measurements are available also at grid points at which the model assumes a non-vanishing ice concentration or at excluded SeaWiFS data points. The RMS log error and log bias shown in figure 11 in the Antarctic Ocean is based on neglecting the presence of sea ice. If we consider only grid points at which the sea ice concentration in the model is zero, we obtain for the free-run model an RMS log error of 0.44 with a log bias of 0.02. For the assimilation the RMS log error is 0.38 with a log bias of -0.16. The SeaWiFS data shows an RMS log error of 0.4 with a log bias of -0.23 taking into consideration all collocation points of satellite and in situ data. Thus, the assimilation reduces the error of the free model to a level slightly below the error of SeaWiFS at points without sea ice. In addition, the bias of the assimilation estimate lies in between the biases of the free model and the satellite data.

To obtain a better insight in the distribution of the errors, figure 12 shows histograms of the frequency distribution of log errors. The histograms show a nearly normal distribution of the log errors. This supports our assumption of a log-normal distribution of chlorophyll errors. However, for the free-run model the distribution is skewed with a larger extent of underestimated than overestimated chlorophyll concentrations. In addition, next to the maximum at zero, a second relative maximum is visible for values around 0.3. The assimilation strongly reduces the spread such that it is only slightly higher than the spread for the SeaWiFS data. Some skewness in the distribution remains. The second maximum at positive values for the free run and maximum the assimilation at positive values is mainly caused by the errors in the North Central Pacific. Here, the error distribution exhibits the maximum at positive values while the skewness of the distribution toward negative values leads to an overall negligible bias as is visible in figure 11.

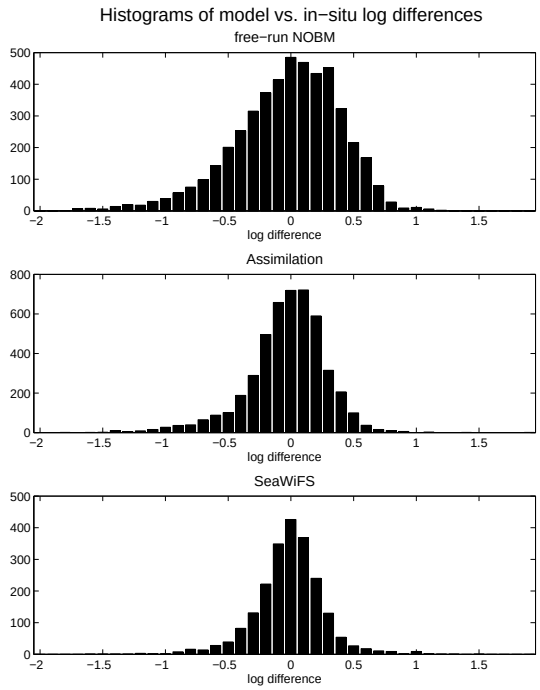

Fig. 12. Histograms of the logarithmic differences between model and in situ data.

6 Discussion

The assimilation of daily SeaWiFS chlorophyll data using the local SEIK filter in the simplified univariate form applied here, resulted in a significant improvement of the surface chlorophyll fields estimated by the NOBM. The assimilation provides daily full global chlorophyll fields. The comparison with in situ data has shown that these fields are have similar errors as the SeaWiFS data. However, globally the fields estimated by the assimilation have a log error which is 13% higher than the error of the SeaWiFS data.

The regional comparison showed that there are regions in which the assimilation provides an estimate of smaller error than the satellite data and those in which the assimilation estimate is inferior. In particular these are the North Central Pacific, but also the Antarctic basin and the North Atlantic. In the North Central Pacific, the dominating data from the CalCOFI project results in a larger error. The small estimated variances in this region point to limitations of the use of a static covariance matrix and the dependence of the assimilation result on the particular choice of the covariance matrix. In the experiments, the covariance matrix has been computed from the monthly variability of model surface chlorophyll with respect to the 7-year mean of the model. This choice has certain limitations. The temporal coverage of the covariance matrix results in small variances for areas with small annual variability while large variances are obtained, e.g. for the areas in high latitudes which show strong spring blooms. Further, the initialization of the covariance matrix

assumes that the model is perfect. This results in a general underestimation of error estimates. These particularities of the covariance matrix are addressed in the assimilation system by adjusting the error estimates of the observations to values which minimize the estimation error over the 7-year period as has been discussed in section 4. There are obvious variations of the covariance matrix. The covariance matrix could be computed from state vectors with a higher temporal resolution. In addition, a running mean over weeks to months could be used instead of a long-time mean. Finally, the state vectors could be generated from ensemble runs which consider the possible model errors. These runs could include variations the model parameters or a stochastic component, for example in the atmospheric forcing. While these variation likely lead to more realistic variance estimates of the model, it is unknown, whether they would lead to smaller estimation errors in the assimilation process.

The decision to use a static covariance matrix can also be expected to have a significant influence on a data assimilation application. A static covariance matrix neglects dynamical changes in the variances and of correlations between the model variables caused by the evolution of the pelagic system. An example for this are correlations between the chlorophyll at the surface and in lower model layers. The vertical distributions of ocean chlorophyll are typically either decreasing with depth from the surface, or increasing to a maximum near the bottom of the mixed layer. These two distributions can change regionally and seasonally. Accordingly, the vertical correlations are expected to change seasonally. To estimate the changing correlations, the dynamic propagation of the covariance matrix is required. In the case of the univariate assimilation performed here, these issues have only a limited influence, as the assimilation is governed by the estimated variances and spatial covariances within each analysis domain.

The experiments only updated the surface chlorophyll field. We note, that it is desirable to also update the deeper model layers, at least in the euphotic layer, to preserve the consistency of the chlorophyll profiles. However, as was outlined above, this is hardly possible when a static covariance matrix is used, because the multi-layer updates involve the problem of estimating dynamically changing correlations between chlorophyll at the surface and in deeper layers.

While the assimilation was only performed univariately, variables that are directly related to chlorophyll are affected in a positive manner by the univariate assimilation. Other state variables and processes are only indirectly affected. They will react on the changed chlorophyll concentrations during the model integration and will tend to push the model results in the same direction as the free-run model. In the assimilation, the frequent assimilation updates lead to a balance between improvements by the assimilation and the dynamical tendency toward the free-run model result. To improve the estimation, other state variables could be updated using a multivariate assimilation which uses

estimated covariances between the surface chlorophyll and other variables. The ability of a multivariate assimilation will depend on the possibility to obtain meaningful covariances between the different model fields. The results by Carmillet et al. [2001] showed that this is possible, at least for synthetic data which is fully consistent with the model formulation.

For a comparison of our assimilation results with previous studies only that by Gregg [2007] allows for a meaningful comparison. Gregg [2007] applied the CRAM method to a previous version of the NOBM. This method provides slightly better results in the comparison to in situ data. This is mainly due to an inferior performance of the CRAM method in the CalCOFI region. Natvik and Evensen [2003] assimilated SeaWiFS chlorophyll data into a 3-dimensional model in the North Atlantic over a period of two months using a multivariate implementation of the EnKF. Their method was able to reduce the difference between the free-run model and the satellite data at the times of the analysis update of the filter. However, the short period of their experiment does not allow for a comparison with our results.

7 Conclusion

A local SEIK filter has been applied to assimilate real SeaWiFS ocean chlorophyll data univariately into the surface layer of the NASA Ocean Biogeochemical Model. The filter has been simplified by using a constant error estimate for the state, thus avoiding the need of a costly ensemble integration. The assimilation is performed on the logarithm of the total chlorophyll field because of the log-normal distribution of chlorophyll. While the satellite provides only a measurement of total chlorophyll, the model simulates four phytoplankton groups. Because direct information about the relative abundances of the phytoplankton groups is not available from the satellite data, the assimilation was performed under the constraint that the relative abundances of the phytoplankton groups remain unchanged during each assimilation update of the model state.

The assimilation of SeaWiFS ocean chlorophyll data into the NASA Ocean Biogeochemical Model over the 7-year period from 1998 to 2004 resulted in a significant improvement of the surface chlorophyll estimate compared to the free-run model. Realistic complete daily chlorophyll fields were provided by the assimilation.

Compared to in situ data over the assimilation period, the global logarithmic error was 0.32 for the assimilation with a bias of -0.032. The free-run model error was larger with 0.43 while the bias was almost the same with -0.030. The SeaWiFS data showed slightly smaller error than the assimilation with 0.28

and a bias of -0.012. However, regionally the assimilation provided in several basins estimates of total chlorophyll with a smaller deviation from in situ data than SeaWiFS data did.

This study is the initial step of work which is intended to lead to a full-featured implementation of a SEIK filter with dynamic error evolution. Here only the total surface chlorophyll concentration was directly modified by the assimilation. The ultimate goal of a comprehensive data assimilation system would involve multi-variate assimilation, in which also variables like nutrients are updated during the analysis step of the filter algorithm. Also, the inclusion of lower model layers in the analysis update is required. In addition, the dynamic error estimation in terms of an ensemble integration is expected to improve the assimilation. However, this technique will increase the computing requirements significantly. In the experiments with the simplified SEIK filter the error estimates of the observations are chosen for good performance of the filter. However, with more realistic estimates of the estimation error in the model, the error estimates of the observations need to be revised for better realism.

Acknowledgements

We would like to acknowledge Orbimage Corp. for collecting SeaWiFS data and the NASA Ocean Biology Processing Group for processing and distribution. We would thank NODC for acquisition and distribution of in situ chlorophyll data. Nancy Casey, SSAI, acquired and provided model forcing and validation data sets from a wide variety of sources and formats. We are also thankful for the helpful comments of two anonymous reviewers. This work was supported by NASA RTOP (grant) 621-30-39.

References

R. A. Armstrong, J. L. Sarmiento, and R. D. Slater. Monitoring ocean productivity by assimilating satellite chlorophyll into ecosystem models. In Powell and Steele, editors, *Ecological Time Series*, pages 371–390. Chapman and Hall, London, 1995.

M. J. Behrenfeld and P. G. Falkowski. Photosynthetic rates derived from satellite-based chlorophyll concentration. *Limnol. Oceanogr.*, 42:1–20, 1997.

K. Brusdal, J. M. Brankart, G. Halberstadt, G. Evensen, P. Brasseur, P. J. van Leeuwen, E. Dombrowsky, and J. Verron. A demonstration of ensemble based assimilation methods with a layered OGCM from the perspective of operational ocean forecasting systems. *J. Mar. Syst.*, 40-41:253–289, 2003.

G. Burgers, P. J. van Leeuwen, and G. Evensen. On the analysis scheme in the Ensemble Kalman Filter. *Mon. Wea. Rev.*, 126:1719–1724, 1998.

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

V. Carmillet, J.-M. Brankart, P. Brasseur, H. Drange, G. Evensen, and J. Verron. A singular evolutive extended Kalman filter to assimilate ocean color data in a coupled physical-biochemical model of the North Atlantic ocean. *Ocean Modeling*, 3:167–192, 2001.

M. E. Conkright, J. I. Antonov, O. Baranova, T. P. Boyer, H. E. Garcia, R. Gelfeld, D. Johnson, T. D. OBrien, I. Smolyar, and C. Stephens. *World ocean database 2001, Vol. 1: Introduction*. NOAA Atlas NESDIS 42, US Govt. Printing Office, Washington, DC, 2002.

I. G. Enting, T. M. L. Wigley, and M. Heimann. Future emissions and concentrations of carbon dioxide: key ocean/atmosphere/land analyses. Technical Paper 31, CSIRO Aust. Div. Atmos. Res., 1994.

G. Evensen. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.*, 99(C5):10143–10162, 1994.

K. Fennel, M. Losch, J. Schröter, and M. Wenzel. Testing a marine ecosystem model: Sensitivity analysis and parameter optimization. *J. Mar. Syst.*, 28: 45–63, 2001.

M. A. M. Friedrichs. Assimilation of JGOFS EqPac and SeaWiFS data into a marine ecosystem model of the central equatorial Pacific Ocean. *Deep-Sea Res. II*, 49:289–320, 2002.

E. Garcia-Gorriz, N. Hoepffner, and M. Ouberdous. Assimilation of SeaWiFS data in a coupled physical-biological model of the Adriatic Sea. *J. Mar. Syst.*, 40-41:233–252, 2003.

P. Ginoux, M. Chin, I. Tegen, J. M. Prospero, B. Holben, O. Dubovik, and S.-J. Lin. Sources and distributions of dust aerosols simulated with the GOCART model. *J. Geophys. Res.*, 106:20255–20273, 2001.

W. W. Gregg. A coupled ocean-atmosphere radiative model for global ocean biogeochemical models. Technical Report 2002-104606, Vol. 22, NASA, 2002.

W. W. Gregg. Assimilation of SeaWiFS ocean chlorophyll data into a three-dimensional global ocean model. *J. Mar. Syst.*, 2007. in press; doi:10.1016/j.marsys.2006.02.15.

W. W. Gregg and K. L. Carder. A simple spectral solar irradiance model for cloudless maritime atmospheres. *Limnology and Oceanography*, 35:1657–1675, 1990.

W. W. Gregg and N. W. Casey. Global and regional evaluation of the SeaWiFS chlorophyll data set. *Rem Sens. Env.*, 93:463–479, 2004.

W. W. Gregg and N. W. Casey. Modeling coccolithophores in the global oceans. *Deep-sea Res. II*, 54:447–477, 2007.

J. C. P. Hemmings, M. A. Srokosz, P. Challenor, and M. J. R. Fasham. Assimilating satellite ocean-colour observations into oceanic ecosystem models.

Philosophical Transactions of the Royal Society of London A, 361:33–39, 2003.

J. C. P. Hemmings, M. A. Srokosz, P. Challenor, and M. J. R. Fasham. Split-domain calibration of an ecosystem model using satellite ocean colour data. *J. Mar. Syst.*, 50:141–179, 2004.

S. B. Hooker, W. E. Esaias, G. C. Feldmann, W. W. Gregg, and C. R. McClain. An overview of seawifs and ocean color. In S. B. Hooker and E. R. Firestone, editors, *NASA Technical Memorandum 104566*, volume 1 of *SeaWiFS Technical Report Series*. NASA Goddard Space Flight Center, Greenbelt, Maryland, 1992.

P. L. Houtekamer and H. L. Mitchell. A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.*, 129:123–137, 2001.

J. Ishizaka. Coupling of Coastal Zone Color Scanner data to a physical-biological model of the southeastern United-States continental-shelf ecosystem. 3. Nutrient and phytoplankton fluxes and CZCS data assimilation. *J. Geoph. Res.*, 95:20201–20212, 1990.

A. H. Jazwinski. *Stochastic Processes and Filtering Theory*. Academic Press, New York, 1970.

R. E. Kalman. A new approach to linear filtering and prediction problems. *Trans. ASME, J. Basic Eng.*, 82:35–45, 1960.

C. L. Keppenne, M. M. Rienecker, N. P. Kurkowski, and D. A. Adamec. Ensemble Kalman filter assimilation of temperature and altimeter data with bias correction and application to seasonal prediction. *Nonl. Proc. Geoph.*, 12:491–503, 2005.

S. N. Losa, G. A. Kivman, J. Schröter, and M. Wenzel. Sequential weak constraint parameter estimation in an ecosystem model. *J. Mar. Syst.*, 43: 31–49, 2001.

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

M. Messié, M.-H. Radenac, J. Lefèvre, and P. Marchesiello. Chlorophyll bloom in the western Pacific at the end of the 1997-1998 El Niño: The role of the Kiribati Islands. *Geophys. Res. Lett.*, 33:L14601, doi:10.1029/2006GL026033, 2006.

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

L. Nerger, W. Hiller, and J. Schröter. A comparison of error subspace Kalman filters. *Tellus*, 57A:715–735, 2005a. doi:10.1111/j.1600-0870.2005.00141.x.

L. Nerger, W. Hiller, and J. Schröter. PDAF - the Parallel Data Assimilation Framework: Experiences with Kalman filtering. In W. Zwieflhofer and G. Mozdzynski, editors, *Use of High Performance Computing in Meteorology - Proceedings of the 11. ECMWF Workshop*, pages 63–83. World Scientific, 2005b.

L. Nerger, S. Danilov, W. Hiller, and J. Schröter. Using sea level data to

constrain a finite-element primitive-equation ocean model with a local SEIK filter. *Ocean Dynamics*, 56:634–649, 2006. doi:10.1007/s10236-006-0083-0.

L. Nerger, S. Danilov, G. Kivman, W. Hiller, and J. Schröter. Data assimilation with the ensemble Kalman filter and the SEIK filter applied to a finite element model of the North Atlantic. *J. Mar. Syst.*, 65:288–298, 2007. doi:10.1016/j.jmarsys.2005.06.009.

D. T. Pham, J. Verron, and L. Gourdeau. Singular evolutive Kalman filters for data assimilation in oceanography. *C. R. Acad. Sci., Ser. II*, 326(4): 255–260, 1998a.

D. T. Pham, J. Verron, and M. C. Roubaud. A singular evolutive extended Kalman filter for data assimilation in oceanography. *J. Mar. Syst.*, 16: 323–340, 1998b.

M. Schartau and A. Oschlies. Simultaneous data-based optimization of a 2D-ecosystem model at three locations in the north atlantic: Part I - method and parameter estimates. *J. Mar. Res.*, 61:765–793, 2003.

R. Schlitzer. Carbon export fluxes in the Southern Ocean: Results from inverse modeling and comparison with satellite-based estimates. *Deep-sea Res. II*, 49:1623–1644, 2002.

P. S. Schopf and A. Loughe. A reduced gravity isopycnal ocean model: Hindcasts of El Niño. *Mon. Wea. Rev.*, 123:2839–2863, 1995.

Y. H. Spitz, J. R. Moisan, M. R. Abbott, and J. G. Richman. Data assimilation and a pelagic ecosystem model: Parameterization using time series observations. *J. Mar. Syst.*, 16:51–68, 1998.

Y. H. Spitz, J. R. Moisan, and M. R. Abbott. Configuring an ecosystem model using data from the Bermuda Atlantic Time Series (BATS). *Deep-Sea Res. II*, 48:1733–1768, 2001.

D. Stammer, C. Wunsch, R. Giering, C. Eckerts, P. Heimbach, J. Marortzke, A. Adcroft, C. Hill, and J. Marshall. The global ocean circulation during 1992-1997, estimated from ocean observations and a general circulation model. *J. Geophys. Res.*, 107(C9):3001, 2002. doi:10.1029/2001JC000888.

M. Wang, K. D. Knobelspiesse, and C. R. McClain. Study of the Sea-Viewing Wide Field-of-View Sensor (SeaWiFS) aerosol optical property data over ocean in combination with the ocean color producets. *J. Geophys Res.*, 110: D10S06, 2005. doi:10.1029/2004JD004950.

P. J. Werdell and S. W. Bailey. The SeaWiFS bio-optical archive and storage system (SeaBASS): Current architecture and implementation. NASA Technical Memorandum 2002-211617, NASA Goddard Space Flight Center, Greenbelt, MD, 2002.