Improving Assimilation of SeaWiFS Data by the Application of Bias Correction with a Local SEIK Filter

Lars Nerger [a,b,c,*] 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*

[c] *Current affiliation: Alfred Wegener Institute for Polar and Marine research, Bremerhaven, Germany*

Received May 15, 2007; revised August 30, 2007; accepted September 12, 2007

Abstract

Ocean-biogeochemical models show typically significant errors in the representation of chlorophyll concentrations. The model state can be improved by the assimilation of satellite chlorophyll data with algorithms based on the Kalman filter. However, these algorithms do usually not account for the possibility that the model prediction contains systematic errors in the form of model bias. Accounting explicitly for model biases can improve the assimilation performance. To study the effect of bias estimation on the estimation of surface chlorophyll concentrations, chlorophyll data from the Sea-viewing Wide Field-of-view Sensor (SeaWiFS) are assimilated on a daily basis into the NASA Ocean Biogeochemical Model (NOBM). The assimilation is performed by the ensemble-based SEIK filter combined with an online bias correction scheme. The SEIK filter is simplified here by the use of a static error covariance matrix. The performance of the filter algorithm is assessed by comparison with independent in situ data over the 7-year period 1998–2004. The bias correction results in significant improvements of the surface chlorophyll concentrations compared to the assimilation without bias estimation. With bias estimation, the daily surface chlorophyll estimates from the assimilation show about 3.3% lower error than SeaWiFS data. In contrast, the error in the global surface chlorophyll estimate without bias estimation is 10.9% larger than the error of SeaWiFS data.

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

1 Introduction

Since September 1997, SeaWiFS has provided an uninterrupted time series of high quality global ocean chlorophyll data. This unprecedented data set has led to many advances in our understanding of ocean biological processes, and has been an invaluable resource for assessing the skill of ocean biological models. Recently there has been expanding interest in using the data set in data assimilation studies, see Gregg [2007] for an overview of previous studies using SeaWiFS data for assimilation. Such applications hold potential for improving estimates of model state fields and model parameters. As an emerging field, researchers have only recently begun to use satellite data for improving the model state directly, a process known as state estimation.

Aiming at a direct state estimation without changing model parameters in a 3-dimensional model, Carmillet et al. [2001] applied a singular "evolutive" extended Kalman (SEEK) filter to assimilate simulated observations into a model in the North Atlantic. These twin experiments focused on the general possibilities for multivariate assimilation with ocean chlorophyll data. Using very accurate data with a prescribed error of 10%, Carmillet et al. [2001] were able to constrain phytoplankton as well as 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.

The first global long-time assimilation of SeaWiFS ocean chlorophyll data was discussed by Gregg [2007]. The data were assimilated daily into the global 3-dimensional NASA Ocean Biogeochemical Model (NOBM) over the period of 1997-2003 using a conditional relaxation method. 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 independent 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.

Recently, Nerger and Gregg [2007] used a simplified form of the singular "evolutive" interpolated Kalman (SEIK) filter with localized analysis [Nerger et al.,

* Corresponding author. Alfred Wegener Institute for Polar and Marine Research, 27570 Bremerhaven, Germany. Phone +49 (471) 4831 1558, fax +49 (471) 4731 1590

Email address: lars.nerger@awi.de (Lars Nerger).

2005a] to assimilate SeaWiFS ocean chlorophyll data over a period of 7 years into an updated version of the NOBM used by Gregg [2007]. The univariate application of the SEIK filter to surface chlorophyll provided daily global surface chlorophyll fields that exhibited a lower RMS log error than the SeaWiFS data in the majority of the oceanic basins when compared to in situ data. Globally, the RMS log error for the assimilation was 13% larger than the error of the SeaWiFS data. In connection to the estimation of surface chlorophyll, the primary production in the model was improved. The univariate assimilation scheme was not configured to constrain other fields. However, it was shown that the univariate assimilation did not harm other model fields, like nutrients, which remained constant during the assimilation updates but reacted on the changed chlorophyll concentrations during the model integration.

The previous assimilation studies with ocean-biogeochemical models applied algorithms that do not explicitly account for model bias. While even in this case biases can be reduced by the assimilation, this reduction is not systematic and suboptimal because the noise in the assimilation estimate is increased [see Dee and Da Silva, 1998]. The application of bias correction algorithms together with sequential assimilation algorithms has been demonstrated to improve the performance of the assimilation systems in atmospheric applications [Dee and Todling, 2000, Baek et al., 2006] as well as with physical ocean circulation models [Chepurin et al., 2005, Keppenne et al., 2005].

To study the influence of bias estimation on the assimilation of satellite chlorophyll data, the SEIK algorithm applied by Nerger and Gregg [2007] is extended here with an online bias correction scheme. This extended assimilation algorithm is used to estimate daily surface chlorophyll over the period September 1997 to December 2004. The assimilation performance is assessed by comparison with independent in situ data. Our objective is to apply data assimilation with an established biogeochemical model to produce improved estimates of global chlorophyll. Although the model is more complex than necessary to demonstrate the effectiveness of the data assimilation application in principle, its use provides an opportunity to quantitatively evaluate improvements in a realistic context with an extensively validated model.

2 Methods

The configuration of the assimilation experiment performed here is similar to the assimilation discussed by Nerger and Gregg [2007]. The differences of both studies are summarized in table 1. Below, the parts of the assimilation system and the experimental configuration are described.

2.1 The NASA Ocean-Biogeochemical Model

The NOBM used here for the data assimilation is a fully coupled general circulation/biogeochemical/radiative model. It consists of three major components, which are depicted in figure 1. Simulated are the ocean general circulation, radiative transfer processes, and biogeochemical processes as outlined below.

The Poseidon model [Schopf and Loughe, 1995] simulates the ocean general circulation. This finite-difference, reduced gravity ocean model is used here in a global configuration extending from near the South Pole to 72°N including all regions with bottom depth $>$ 200m. The discretization uses a uniform resolution of 2/3° in latitude and 5/4° in longitude. 14 layers in quasi-isopycnal coordinates are used in the vertical. The model is forced by wind stress, sea surface temperature, and short-wave radiation.

The biogeochemical processes model is described in detail in Gregg and Casey [2007]. It consists of ecosystems and carbon components. The ecosystem component simulates four phytoplankton groups: Diatoms, Chlorophytes, Cyanobacteria, and Coccolithophores. These are characterized by distinct growth and sinking rates, as well as differences in nutrient requirements. Further, spectral absorption and scattering, as well as light saturation constants are distinct. The phytoplankton groups interact with 4 nutrients: nitrate, regenerated ammonium, silica, and iron. Storage of organic material, sinking and eventual remineralization back to usable nutrients are simulated using three detrital pools. In addition, a single herbivore group is modeled. The carbon component models the interaction of dissolved organic and inorganic carbon with phytoplankton, herbivores and detritus. In addition, the exchange of carbon-dioxide with the atmosphere is considered. The model is forced by transient monthly atmospheric fields.

2.2 SeaWiFS Ocean Chlorophyll Data

The assimilation uses daily global chlorophyll data from SeaWiFS, version 5.1 at 9km resolution obtained from the NASA Ocean Color Web site. The data fields were remapped to the model grid for the assimilation.

For the assimilation daily SeaWiFS chlorophyll data with concentrations larger than twice the monthly mean are considered as outliers and excluded. These exclusions are motivated by the fact the remote sensing errors are typically expressed as overestimates as the most dominant error sources, absorbing aerosols, chromophoric dissolved organic matter (CDOM), sub-pixel scale clouds and ice most often lead to overestimates of chlorophyll [Gregg and Casey, 2004]. These overestimates can have a very deleterious effect on the quality

and stability of the assimilation process.

The distribution of chlorophyll and errors in chlorophyll are assumed to be log-normal [see, Campbell, 1995, Nerger and Gregg, 2007]. As the analysis of the Kalman filter is only variance minimizing for normal distributions of the state and observation errors, 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 $\mathbf{R}_k$ (see appendix, equations 13 and 14) is diagonal. Regionally varying errors are specified for the observations as shown in figure 2. These error estimates are chosen to minimize the estimation errors in the assimilation. In the North and Equatorial Indian Ocean observations with concentrations above 1 mg m^{-3} are excluded. This is motivated by the prevalence of light-absorbing dust [Wang et al., 2005], which can result in overestimates of the chlorophyll concentration. This problem also occurs in the tropical Atlantic where all observations with concentrations larger than 1 mg m^{-3} are excluded in the Mauritanian offshore region (region B in figure 2). The same approach is used in the Amazon and Congo river outflow regions (regions A and C in figure 2, respectively). These regions are dominated by CDOM, which produce erroneous chlorophyll values in the satellite data.

2.3 Local SEIK Filter

The data assimilation is performed using the SEIK filter [Pham et al., 1998], which was also utilized by Nerger and Gregg [2007] to assimilate ocean chlorophyll data. Here, only a brief description of the concepts of the filter algorithm is provided. Details of the mathematical formulation of the SEIK filter and its localization are given in the appendix.

The SEIK filter is an ensemble-based Kalman filter that uses a preconditioned ensemble and a numerically very efficient scheme to incorporate the observational information during the analysis step. SEIK is based on an explicit low-rank approximation of the covariance matrix that estimates the error in the state estimate. The merging of model state and observational data, denoted "analysis", is computed efficiently in the low-dimensional error subspace, which is represented by the low-rank approximated covariance matrix. Compared to the EnKF, which has also been applied for assimilation with ocean-biogeochemical models [Allen et al., 2003, Natvik and Evensen, 2003], the SEIK filter requires much less computation time if the dimension of the observation vector is much larger than the ensemble size. In addition, SEIK can be applied with smaller ensembles than the EnKF. A detailed comparison of the SEIK filter with other, more common, filter algorithms like the EnKF

can be found in Nerger et al. [2005a].

Here, the assimilation applies the localized variant of the SEIK filter [Nerger et al., 2006]. The analysis update of some horizontal location in the model grid considers only observations within some influence radius. The localization is combined with a spatial weighting of covariances, which reduces the influence of remote observations within the influence radius. This method is typically denoted as "covariance localization", [see, e.g. Houtekamer and Mitchell, 2001]. The SEIK filter algorithm applied here is simplified by keeping the state error covariance matrix constant. Hence, the same error estimate for the model state is used for each analysis step. With this simplification only the ensemble mean state is propagated by the model. This avoids the computational cost of integrating a full ensemble of model states during the assimilation process.

2.4 Online Bias Correction

Biases in the model fields are generated by systematic errors in the formulation of the discretized model operator. The time propagation of the true state $\mathbf{x}^t$ of a modeled system is given by the stochastic-dynamic time discretized model equation

$$\mathbf{x}_i^t = M_{i,i-1}[\mathbf{x}_{i-1}^t] + \boldsymbol{\eta}_i \ . \tag{1}$$

Here $M_{i,i-1}$ is a, possibly nonlinear, operator describing the state propagation between the two consecutive time steps $i-1$ and i. The vector $\boldsymbol{\eta}_i$ is the model error. Data assimilation algorithms that are not bias-aware base on the assumption that $\boldsymbol{\eta}_i$ is a stochastic perturbation with zero mean and covariance matrix $\mathbf{Q}_i$. If the mean of $\boldsymbol{\eta}_i$ is non-zero, the model is biased. This leads to systematic differences between the model state and observations, because the modeled state drifts away from the true state during the integration of the numerical model, which does not include the model error term.

In the case of the state modeled by the NOBM, systematic differences between the model state and the SeaWiFS chlorophyll data exist. These are typically slowly-changing differences. For example, the underestimation of the spring bloom in the North Atlantic can be considered to be a bias. The SEIK filter without bias correction assumes that the errors of the model and the observations are unbiased, e.g. they are distributed with alternating signs around the mean state estimate of the filter. If the errors are biased, e.g. if for a longer period of time the model underestimates the data within some region, the efficiency of the filter analyses is reduced.

To account for model bias in conjunction with the SEIK filter, an online bias correction (OBC) scheme is applied, which follows the formulation by Dee and Da Silva [1998]. Assuming that the observations are unbiased, this two-stage bias correction scheme estimates a bias vector analogous to the estimation of the state vector performed by the SEIK filter. The analysis of the state estimate is then performed on the basis of deviations of the de-biased state estimate from the observations. The bias vector is kept constant during the forecast phase while the bias state estimate is propagated by the model. The error estimate for the bias correction uses a fraction of the ensemble covariance matrix that is also used for the state estimation. This scheme separates the state correction into a systematic bias component and a random unbiased component. The mathematical formulation of the OBC scheme in combination with the SEIK filter is provided in the appendix.

2.5 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 constant. With this, the assimilation of chlorophyll does not affect the relative abundances of the phytoplankton groups directly. However, the relative abundances can be indirectly modified. For example, changing to total chlorophyll can change the irradiance availability and its spectral nature. This in turn can benefit or hinder one or more phytoplankton group relative to the others. Similarly, assimilation of total chlorophyll can change the vertical and horizontal gradients of the phytoplankton or nutrients, thus potentially leading to a different environment favoring different groups from the free-run conditions.

The data assimilation process is initialized by a model state estimate for January 1997 obtained from a spin-up run over 20 years with monthly climatological forcing. This state is integrated with transient monthly forcing to obtain a model state estimate for September 1997, the start month of the assimilation experiment. 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 6 years 1998 to 2003 with monthly forcing data. A perturbation matrix is assembled by computing the differences of the state at each 15th day from a running mean over three months. The decomposition $\mathbf{P}_0^a = \mathbf{V}_0\mathbf{U}_0\mathbf{V}_0^T$ (see equation 5 in the appendix) is then obtained by the singular value decom-

position of this perturbation matrix. The covariance matrix obtained by this procedure showed overall variances that were too small to represent a realistic estimate of the initial error of the model state. For this reason, the variance was inflated to obtain a magnitude of variance that is comparable to differences between the initial model state and observations. A factor of 0.06^{-1} resulted in the best assimilation performance when varying the inflation factor. This inflation results in maximum variance estimates of the same order as those in the covariance matrix used by Nerger and Gregg [2007]. The singular value decomposition directly yields the eigenvector matrix $\mathbf{V}_0$ and the square-roots of the eigenvalues, which build the diagonal of the matrix $\mathbf{C}_0$ with $\mathbf{U}_0 = \mathbf{C}_0^T\mathbf{C}_0$. For the data assimilation experiment the leading 30 eigenvectors and eigenvalues are used to generate the state ensemble. The ensemble of 31 members proved to be sufficiently large as larger ensembles resulted only in marginal improvements of the state estimate, while the assimilation performance degraded for smaller ensembles.

The simplified variant of the local SEIK filter is implemented within a data assimilation framework [PDAF, Nerger et al., 2005b], which provides fully-implemented filter algorithms to be connected to existing models in order 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. A localizing weighting of the observations is performed by an exponential decrease from the grid point to be updated with a length scale of 1 grid point to reduce the variance estimate by a factor of 1/e. A fraction of 10% of the ensemble-represented covariance matrix is used for the OBC, while 90% are utilized for the state estimation. The influence of this choice to partition the covariance matrix is further discussed in section 4.

2.6 Performance Evaluation of the Assimilation

To assess the performance of the assimilation, we compare the estimated total surface chlorophyll concentrations with independent in situ chlorophyll data. The in situ data were 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]. For the comparison with the model, daily in situ data were mapped to the model grid by computing the average of measurements within each single grid cell. For the comparison, we excluded in situ data for May 1998 in the South Pacific. Nerger and Gregg [2007] found that these points lead to a large bias in the comparison of the in situ data with both the

modeled chlorophyll and SeaWiFS data. This was caused by high transient and localized concentrations in the eastern part of Melanesia that couldn't be represented by the model at its spatial resolution and monthly forcing.

The analysis is performed globally and separated over 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 Atlantic basins are located north of 40°N.

To quantify the assimilation performance in accordance with the log-normal distribution of total chlorophyll, statistics for the logarithms of the chlorophyll concentrations are computed. Specifically, the RMS log error, the bias as mean deviation between the logarithms of the state estimate and the in situ data, as well as the correlation coefficient between the logarithms of the fields are computed for daily collocated data.

3 Results

We will first discuss the filter performance obtained from the assimilation using the SEIK filter combined with the OBC scheme. Subsequently, the improvements resulting from the OBC are discussed.

3.1 Assimilation Performance with Bias Correction

To assess the assimilation performance of the filter with OBC, the estimated daily chlorophyll concentrations for the years 1998 to 2004 were compared with independent in situ data. Figure 3 shows RMS log errors, bias, and correlations coefficients for the comparison as described in section 2.6. In addition, the number of collocation points over the 7 years is shown below the uppermost panel for both the model estimates and SeaWiFS data. More than twice as much collocation points exist for the complete model fields than for the daily SeaWiFS data. However, the data availability varies strongly by region. The Equatorial Pacific and North Central Pacific oceans dominate the data, followed by the North Central Atlantic and Antarctic oceans. These four basins comprise about 86% of the available in situ data.

The panels of Fig. 3 show a significant improvement of the model-predicted surface chlorophyll field by the assimilation. The RMS log error is reduced

globally from 0.441 to 0.266 by the assimilation. The assimilation reduces the error to be about 3.3% below that of SeaWiFS data, which exhibits an RMS log error of 0.275. In addition, the global bias is reduced from -0.076 to -0.034. However, the SeaWiFS data show an even lower bias of -0.008. The correlation coefficient is increased by the assimilation, too. While the free-run model showed only a positive correlation of 0.52, it was increased by the assimilation to 0.83. The correlation of SeaWiFS data is slightly lower at 0.81. Overall, the assimilation strongly improves the estimate of surface chlorophyll with regard to all three considered statistics.

Regionally, RMS log errors are lower for the assimilation than for SeaWiFS data in all basins except the Equatorial Pacific, the Antarctic and the North Indian Oceans. In the latter basin, the errors are equal. In the Antarctic, the RMS error for the assimilation estimate is about 5% larger than the error of SeaWiFS data. In the Equatorial Pacific Ocean, the error of the assimilation estimate is about 9% larger than that of the SeaWiFS data. However, the error level is quite small here, with 0.185 for the assimilation and 0.169 for SeaWiFS. The improvements from the free run model are significant ($P< 0.05$) in all basins and globally. With regard to the satellite data, only the differences between assimilation and SeaWiFS data in the South Atlantic and North Central Pacific are significant.

While the global bias is small for the free-run model, the assimilation, as well as the SeaWiFS data, larger biases occur regionally. In particular, the free-run model strongly underestimates the concentrations in the in situ data in several basins. The amount of bias is significantly reduced by the assimilation in most basins and globally. The bias in the assimilation estimate remains within the limits of 0.13 for the South Indian Ocean and -0.23 for the Antarctic.

The assimilation provides estimates that are significantly correlated ($P< 0.05$) with the in situ data in all basins, except the South Pacific. However, in this basin only 9 collocation points are available for the comparison. Due to this, also the negative correlation of the free-run model with the in situ data is not significant. In addition, the 100% correlation of SeaWiFS with the in situ data is based on only 2 collocation points. The significant correlation for the assimilation in all other basins is an improvement over the free-run model in which also the Equatorial Atlantic and the Antarctic were not significantly correlated with the in situ data. (Note, that the model shows a significant positive seasonal correlation with SeaWiFS data in all basins [see Gregg, 2007]). Next to the increase of the correlation in all basins by the assimilation in comparison to the free-run model, the assimilation results also in improved correlations compared to the SeaWiFS data. In particular, the SeaWiFS data are not significantly correlated with the in situ data in the North Pacific. Here, the assimilation estimate shows a larger significant correlation than both the free-run model and SeaWiFS data.

The green and yellow bars in the panels of figure 3 compare the assimilation performance for the algorithms without and with OBC. The OBC significantly reduces the global RMS errors from about 10.9% above the error of SeaWiFS data to about 3.3% below. Regionally, the error reduction is significant in all northern and equatorial basins except the Indian Ocean.

Next to the RMS log errors, also the log bias between the model and in situ data is reduced in several basins. The slight global increase in bias is not significant. However, the changes in the bias between the algorithms without and with OBC are significant in all basins, except the South Pacific and Atlantic, and the North Indian Ocean. These are the basins with the smallest number of collocation points. Further, the correlation is increased in all basins, except the Antarctic, South Pacific and Equatorial Indian Oceans.

Figure 4 shows a snapshot of the estimated logarithmic model bias for April 1, 2000. The bias reaches values between about ± 1. The general pattern of the bias estimate is persistent over the full multi-year assimilation period. In particular, the band of positive bias, i.e. the model overestimates the concentration, in the North Central Pacific around 30°N, and the band of positive bias at about 30°S in all basins exist during all seasons with slightly varying amplitude. This is also true for the large negative bias in the tropical Atlantic and the North Indian Oceans. In addition, the pattern of positive bias in the eastern Equatorial Pacific and negative bias in the western part of the tropical Pacific is persistent with the exception of the first months of the assimilation where the El Niño conditions lead to a larger extend of the positive bias to the western part of the Equatorial Pacific. The model bias varies seasonally in the North Pacific and Atlantic Oceans. In general the bias is rather small here, but negative bias is apparent during bloom periods, e.g. in the Atlantic north of 60°N in figure 4, while negative bias occurs at the end of the blooms. The largest variability in the bias is visible in the Antarctic south of 50°S. Again, the bias is negative during boreal spring while during summer and autumn the bias is widely positive.

To examine the actual influence of the bias estimate on the concentration of surface chlorophyll, figure 5 shows the chlorophyll fields for the assimilation without OBC, the biased and de-biased fields for the assimilation with OBC as well as the chlorophyll field from the free-run model for April 1, 2000. Comparing the concentrations from the assimilation without OBC and the de-biased concentration from the assimilation with OBC (upper panels), it is obvious that both estimates are very similar. The concentration in the North and Equatorial Pacific is slightly smaller in the de-biased field. In addition, the concentration of the bloom in the North Atlantic is slightly lower in the de-

biased field. With these differences, the chlorophyll field from the assimilation with OBC corresponds slightly better to the satellite data on the same day than the estimate without OBC.

The effect of the OBC on the chlorophyll estimate that is actually evolved with the model in form of the four phytoplankton groups is visible in the lower panels of figure 5. The biased total chlorophyll estimated by the assimilation with OBC is similar to the chlorophyll field from the free-run model. Generally, the assimilation with OBC adds finer structures to the model field while preserving the large-scale structure of the chlorophyll field from the free-run model. Major changes are visible in the North Atlantic where the phytoplankton bloom is strongly reduced north of about 55°N and in the North and North Central Pacific. In the South Pacific and the Antarctic Oceans, the concentrations are mostly increased but decreased in several regions south of 60°S.

4 Discussion

Several aspects of the daily assimilation of SeaWiFS data, such as the improvement of the chlorophyll estimate from the free-run model, the possibility to obtain complete daily chlorophyll fields, the influence of the univariate assimilation of chlorophyll on nitrate, and the positive influence of the assimilation on the estimate of primary production have been discussed by Nerger and Gregg [2007]. For this reason, we focus here on aspects that relate to the influence of the OBC.

The independent in situ data comparison shows that the application of OBC in the assimilation of daily SeaWiFS chlorophyll data significantly improves the surface chlorophyll estimates. Without OBC, all data assimilation algorithms based on the Kalman filter assume that the model forecast is unbiased. The positive influence of OBC is possibly a general conclusion, and other data assimilation algorithms using ensemble-based Kalman filters that assume the model forecast is unbiased, e.g. Natvik and Evensen [2003], Allen et al. [2003] or Triantafyllou et al. [2003], may also profit from the application of OBC if the models exhibit bias.

The OBC scheme reduces the influence of the SEIK filter on the chlorophyll concentrations evolved by the model. A large part of the correction to the state estimate is contained in the model bias estimate, which is kept static during the forecast phases. The smaller changes to the evolved chlorophyll fields can generally stabilize the evolution of all model fields. An example for the influence of the OBC is the Costa Rica upwelling dome (CRD). Here, the upwelling can result in elevated chlorophyll concentrations as visible in figure 5a. However, the free-run model (Fig. 5c) does not show elevated concentra-

tions, because the upwelling is not well represented by the reduced gravity model. Accordingly, if the assimilation without OBC increases the chlorophyll concentration in the region of the CRD without changing the hydrography, the forecast will reduce the concentrations again. This effect will lead to persistent daily increases in the chlorophyll concentration at the CRD at each analysis. The model dynamics react on this change, e.g. with a decrease in nutrient concentrations. When the increments in chlorophyll are sufficiently large, this can even result in the occurrence of negative nutrient concentrations because the model does intentionally not contain a constraint for positiveness [see Nerger and Gregg, 2007]. With OBC, however, the repeated positive increment to the chlorophyll result in an elevated estimate of negative bias (see figure 4) while the change to the chlorophyll concentrations evolved by the model is very small. Accordingly, the deteriorating effect of the artificial increase in chlorophyll by the analysis followed by its decrease during the forecast will be minimized. In a multivariate analysis, like that applied by Natvik and Evensen [2003], nitrate concentrations are updated directly in conjunction with the correction of chlorophyll. However, the bias correction is still beneficial when switching to a multivariate analysis because also a multivariate ensemble will not be able to represent model bias, if the assimilation algorithm does not account for it.

Some basins need special care when interpreting the statistical comparison of the assimilation estimates with in situ data. One has to keep in mind that the availability of in situ data is very irregular. In addition, the amount of data (4732 collocation points for the comparison of the model fields with in situ data; 2186 points for the comparison of SeaWiFS data) over the 7-year period of the comparison is very limited.

The largest amount of data is available in the Equatorial Pacific and the North Central Pacific. In the North Central Pacific, the RMS log error for the assimilation is 0.26 and about 9% below the error of SeaWiFS data. The in situ data in this basin are dominated by data from the California Cooperative Oceanic Fisheries Investigations (CalCOFI) project, which accounts for about 69% of the data. For this reason, the major amount of this data allows only to access the assimilation performance in a small area near the coast of California. In this region, the RMS log errors are larger than for the total North Central Pacific with 0.37 for the assimilation without OBC, 0.27 for the assimilation with OBC, and 0.30 for SeaWiFS data. According to these numbers, the OBC has a large positive influence in this region by reducing the RMS log error by 27%.

The Equatorial Pacific exhibits a systematic large-scale sampling that follows the Tropical Ocean-Global Atmosphere program/Tropical Atmosphere Ocean (TOGA/TAO) array. Here up to 10 measurements at the same grid point are available over the 7-year period. The assimilation results in a very small RMS

log error and in the Equatorial Pacific. However, the RMS error and bias of SeaWiFS data is even lower. If we only consider in situ data points collocated with satellite data, thus using the same in situ data for SeaWiFS and assimilation, the error of the assimilation estimate in the Equatorial Pacific is reduced to about 1% below the error of the SeaWiFS data. The larger error for the comparison with all available collocation points shows that the information transfer into the gaps of the satellite data incurs some error. The effect of the OBC is very small in the Equatorial Pacific, because biases are very small in this region.

A basin of particularly strong difficulties for both the assimilation and the performance evaluation is the Antarctic Ocean. This is due to the presence of sea ice, the limited and irregular availability of SeaWiFS data, and the irregular availability of in situ data. The availability of SeaWiFS data is limited by the presence of clouds and sea ice. Poleward of 60°S typically less than 10 days of data per month are available per grid point. Accordingly, the model is less constrained and more frequent extrapolation of the data into gaps occurs. This will likely result in an inferior quality of the data assimilation estimate. In addition, the model is less constrained due to the presence of sea ice. At grid points with non-zero ice concentration, the assimilation update of the state is scaled relative to the percentage of open water. This ensures a gradual transition from the constrained ice-free grid points to the unconstrained points with 100% ice concentration. However, the model is less constrained at points with, at least partial, ice coverage. This becomes evident when we neglect all points covered with ice from the comparison of the assimilation estimate with the in situ data. In this case with 296 collocation points, the RMS error drops from being 5% above to 5% below the error of the satellite data. In addition, the amount of bias is reduced from -0.23 to -0.17. The performance evaluation is also influenced by the distribution of the available in situ data. The majority of the data are measured in the Drake passage and near the coast of the Antarctic Peninsula. Thus, the comparison provides mostly information about the errors in the model state and the SeaWiFS data in this very limited region and only during the seasons in which blooms of the phytoplankton occur. This is likely the reason for the rather large positive bias in the Antarctic shown in figure 3. Regarding the influence of the OBC, the Antarctic is the only basin in which the bias correction significantly increases the RMS error and the amount of bias compared to the assimilation without OBC. This shows the difficulty to obtain reasonable estimates of the model bias from the very limited amount of satellite data available in this region.

While the OBC algorithm assumes that the observations are unbiased, the second panel of figure 3 shows that this assumption is not fulfilled. There are significant regional biases in the SeaWiFS data that will reduce the assimilation performance. The OBC scheme is able to reduce the model bias with respect to the SeaWiFS data in most basins. However, not in all of these

basins the bias from the in situ data is reduced. An example for this effect is the North Central Pacific. As discussed above, the error in the CalCOFI region is strongly reduced, but the bias with regard to in situ data remains almost constant.

The efficiency of the OBC also depends on the choice of the weight given for the OBC compared to the state correction. This is defined by the fraction γ of the ensemble-represented state covariance matrix that is used for the OBC (see equations 23 and 24 in the appendix). The results presented here were obtained from experiments using $\gamma = 0.1$. This value resulted in the smallest global RMS log error and smallest log bias when varying γ. However, the global RMS log error and bias only changed by up to 0.5% when varying γ between 0.05 and 0.3. Regionally, the variations of the RMS log error and log bias are larger in some regions. Further, some regions benefit from a higher bias weight while others show better results for smaller γ. For example, the RMS log error in the Antarctic ocean exhibited the smallest RMS log error for $\gamma = 0.05$, while for $\gamma = 0.3$ the RMS log error increased by about 5%. In contrast, the North Central Pacific showed a 2% larger RMS log error for $\gamma = 0.05$ compared to $\gamma = 0.3$.

Since the assimilation was only performed univariately here, only the chlorophyll concentrations are directly modified while other fields, like nutrients, herbivores, and detritus react on the changes chlorophyll during the model integration. Accordingly, we cannot expect a systematic improvement of these fields in our assimilation experiments. Nerger and Gregg [2007] discussed that, without OBC, nitrate concentrations exhibit only small changes due to data assimilation. With OBC these changes are even smaller, because the change in the model-integrated biased chlorophyll fields is smaller.

The assimilation experiments discussed here are influenced by the assumption that the errors in chlorophyll concentrations exhibit a log-normal distribution. This assumption is in contrast to other studies [Carmillet et al., 2001, Natvik and Evensen, 2003, Allen et al., 2003] who assimilated actual concentrations and hence implicitly assumed a normal distribution of the errors. The studies, however, assumed a fixed relative error of the observations. Analogously, the regional RMS log observations errors assumed here correspond to relative errors. The log-normal assumption influences the generation of the initial ensemble, as well as the analysis and re-initialization phases of the SEIK filter. All these operations have to be performed using the logarithm of the concentrations. An immediate advantage of the log-normal assumption for the filter analysis, apart from the mathematical consistency, is that unrealistic negative concentrations, e.g. reported by Natvik and Evensen [2003], cannot be caused by the analysis update. A detailed assessment of the influence of the log-normal assumption compared to an assumption of a normal error distribution is beyond the scope of this work. However, the general behavior in

the scalar case of a single state vector element and a single observation is the following. If the analysis update increases concentrations, the log-normal assumption will lead to smaller increments compared to a filter with normal error assumption. If the increment is negative, its size will be larger for a log-normal error distribution compared to a normal distribution.

5 Conclusion

An online bias correction scheme has been applied with the SEIK filter with localized analysis and constant error covariance matrix to estimate total surface chlorophyll from September 1997 to the end of 2004. For this, daily SeaWiFS chlorophyll data have been assimilated into the surface layer of the global 3-dimensional NASA Ocean Biogeochemical Model.

The assimilation performance was assessed by comparison to independent in situ data. The application of bias correction resulted in improvements of the estimated surface chlorophyll. Compared to the assimilation without OBC, the improvements are significant globally and regionally in the equatorial and northern regions of the Pacific and Atlantic Oceans. The improvements obtained by the application of the OBC also show that there are significant systematic errors in the model estimate and that these can be estimated as model bias.

The estimated surface chlorophyll shows smaller RMS errors than SeaWiFS data in almost all of the major oceanic basins. Globally, the assimilation estimate shows a 3.3% smaller RMS log error and a larger correlation with the in situ data, than the SeaWiFS data. The assimilation estimate shows smaller errors than SeaWiFS data in all basins, except the Equatorial Pacific, the North Indian, and Antarctic oceans. In the Equatorial Pacific, the satellite data are already very accurate and practically unbiased. The assimilation estimate is inferior to the satellite data, because of the extrapolation of this information into data gaps. In the Antarctic ocean, the assimilation performance is reduced by limited availability of satellite data and by the occurrence of sea ice.

Despite the bias correction scheme applied here, the assimilation estimate shows significant biases with regard to in situ data. Partly, this is due to biases in the SeaWiFS data. These are neglected by the bias correct scheme, which assumes unbiased observations. To improve the assimilation performance it would be necessary to account also for the bias in the SeaWiFS data.

This study is an intermediate step toward a full-featured multivariate assimilation system based on a SEIK filter with dynamic evolution. It shows the importance of bias correction in the assimilation system. Further, in conjunc-

tion with [Nerger and Gregg, 2007], it was shown, that daily assimilation of satellite chlorophyll data into a global ocean-biogeochemical model over a period of several years is feasible with an advanced data assimilation method. It results in significant improvements of the model-estimated surface chlorophyll concentrations with errors that are below those of SeaWiFS chlorophyll data.

Acknowledgements

We thank the NASA Ocean Biology Processing Group for SeaWiFS and in situ data, and NODC for in situ data. We thank Nancy Casey, SSAI for obtaining, quality control, combining, and re-formatting in situ data sets. We are also thankful for the careful review by Markus Schartau. This work was supported by the NASA EOS and MAP programs.

Appendix

Here, the details of the mathematical formulation of the SEIK filter, its localization, and the online bias correction are described in their general form for Gaussian errors. Finally, implementation aspects for the experiments discussed above are discussed. In particular the handling of log-normal distributions is considered.

A The (global) SEIK Filter

In the SEIK filter, the estimate of the state of a physical system, such as the ocean, is expressed 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 $\mathbf{P}_k^a$ that represents the error estimate of the state vector. The ensemble-based filter scheme represents these quantities by an ensemble of N vectors $\{\mathbf{x}^{a(\alpha)}, \alpha = 1, \ldots, N\}$ of model state realizations:

$$\mathbf{X}_k^a = \{\mathbf{x}_k^{a(1)}, \ldots, \mathbf{x}_k^{a(N)}\} \tag{2}$$

The state estimate is given by the ensemble mean

$$\overline{\mathbf{x}_k^a} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_k^{a(i)}, \tag{3}$$

and the covariance matrix $\mathbf{P}_k^a$ is approximated by the ensemble covariance matrix

$$\tilde{\mathbf{P}}_k^a := \frac{1}{N-1}(\mathbf{X}_k^a - \overline{\mathbf{X}_k^a})(\mathbf{X}_k^a - \overline{\mathbf{X}_k^a})^T \approx \mathbf{P}_k^a \ , \tag{4}$$

with $\overline{\mathbf{X}_k^a} = \{\overline{\mathbf{x}_k^a}, \ldots, \overline{\mathbf{x}_k^a}\}$.

The SEIK algorithm can be subdivided into several phases and is prescribed by the following equations:

Initialization:

To initialize the filter algorithm, a state estimate $\mathbf{x}_0^a$ is required. In addition, an initial covariance matrix $\mathbf{P}_0^a$ is estimated by a rank-r matrix, which is given in decomposed form as

$$\mathbf{P}_0^a := \mathbf{V}_0\mathbf{U}_0\mathbf{V}_0^T. \tag{5}$$

Here $\mathbf{U}_0$ and $\mathbf{V}_0$ are matrices of size $r \times r$ and $n \times r$, respectively.

From $\mathbf{x}_0^a$ and $\mathbf{P}_0^a$ a random ensemble of minimum size $N = r + 1$ is generated whose statistics represent $\mathbf{x}_0^a$ and $\mathbf{P}_0^a$ exactly. This can be achieved by transforming the columns of matrix $\mathbf{V}_0$ by a random matrix with special properties. $\mathbf{P}_0^a$ can be written as

$$\mathbf{P}_0^a = \mathbf{V}_0\mathbf{C}_0^T\mathbf{\Omega}_0^T\mathbf{\Omega}_0\mathbf{C}_0\mathbf{V}_0^T, \tag{6}$$

were $\mathbf{C}_0^T\mathbf{C}_0 = \mathbf{U}_0$ is a square root of the matrix $\mathbf{U}_0$. If the decomposition in equation (5) is obtained by a singular value decomposition, $\mathbf{C}_0$ will be a diagonal matrix holding singular values. $\mathbf{\Omega}_0$ is a $N \times r$ random matrix generated form uniformly distributed random numbers. The columns of $\mathbf{\Omega}_0$ are constrained to be orthonormal and orthogonal to the vector $(1, \ldots, 1)^T$. The ensemble of state realizations is then given by

$$\mathbf{X}_0^a = \overline{\mathbf{X}_0^a} + \sqrt{N-1}\ \mathbf{V}_0\mathbf{C}_0^T\mathbf{\Omega}_0^T. \tag{7}$$

Forecast:

The state ensemble is integrated using the numerical model in the "forecast phase". Each ensemble member $\{\mathbf{x}^{a(\alpha)}, \alpha = 1, \ldots, N\}$ is evolved up to time t_k by iterating the model equation

$$\mathbf{x}_i^{f(\alpha)} = M_{i,i-1}[\mathbf{x}_{i-1}^{a(\alpha)}] + \boldsymbol{\eta}_i^{(\alpha)}. \tag{8}$$

where $M_{i,i-1}$ denotes the nonlinear dynamic model operator that integrates a model state from time t_{i-1} to time t_i. The superscript 'f' denotes the fore-

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

Analysis:
In the "analysis phase", the information from the model state and the data are merged to provide updated state and error estimates. $\mathbf{P}_k^f$ can be computed from the state ensemble $\mathbf{X}_k^f$ in analogy to the covariance matrix in (5) according to

$$\mathbf{P}_k^f = \mathbf{L}_k \mathbf{G} \mathbf{L}_k^T \tag{9}$$

with

$$\mathbf{L}_k := \mathbf{X}_k^f \, \mathbf{T}, \quad \mathbf{G} := (N-1)^{-1} \left(\mathbf{T}^T \mathbf{T} \right)^{-1}. \tag{10}$$

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

$$\mathbf{T} = \begin{pmatrix} \mathbf{I}_{r \times r} \\ \mathbf{0}_{1 \times r} \end{pmatrix} - \frac{1}{N} \left(\mathbf{1}_{N \times r} \right) \tag{11}$$

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 is given by

$$\mathbf{x}_k^a = \overline{\mathbf{x}_k^f} + \mathbf{L}_k \mathbf{a}_k \tag{12}$$

where the vector $\mathbf{a}_k$ of size $N-1$ is computed by

$$\mathbf{a}_k = \mathbf{U}_k (\mathbf{H}_k \mathbf{L}_k)^T {\mathbf{R}_k}^{-1} \left(\mathbf{y}_k^o - \mathbf{H}_k \overline{\mathbf{x}_k^f} \right), \tag{13}$$

$$\mathbf{U}_k^{-1} = \rho \mathbf{G}^{-1} + (\mathbf{H}_k \mathbf{L}_k)^T \mathbf{R}_k^{-1} \mathbf{H}_k \mathbf{L}_k. \tag{14}$$

Here, $\mathbf{H}_k$ is the observation 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. ρ is denoted forgetting factor ($0 < \rho \leq 1$). It leads to an inflation of the estimated variances of the model state and 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:

Before the next forecast phase, the "re-initialization phase" is performed in which 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

$$\mathbf{X}_k^a = \overline{\mathbf{X}_k^a} + \sqrt{N-1}\,\mathbf{L}_k\mathbf{C}_k^T\mathbf{\Omega}_k^T. \tag{15}$$

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.

B Localized Analyses and Re-initializations in SEIK

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

For the localization of the analysis and re-initialization phases in the SEIK filter the operations in these phases are performed in a loop through disjoint local analysis domains of the model grid, rather than performing the update of the full state vector at once. In a simple case, a local analysis domain can be a single water column. The results of the analysis and re-initialization phases remain unchanged by this reformulation, as long as all globally available observations are considered in the analysis phase.

In the analysis step, the localization is performed by neglecting observations that are beyond a prescribed influence distance from a local domain. Below, we omit the time index k for clarity, as all quantities refer to the same time. 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 (12 – 14) as

$$\mathbf{x}_\sigma^a = \overline{\mathbf{x}_\sigma^f} + \mathbf{L}_\sigma\mathbf{a}_\delta, \tag{16}$$

$$\mathbf{a}_\delta = \mathbf{U}_\delta(\mathbf{H}_\delta\mathbf{L})^T(\mathbf{R}_\delta)^{-1}\left(\mathbf{y}_\delta^o - \mathbf{H}_\delta\overline{\mathbf{x}^f}\right), \tag{17}$$

$$\mathbf{U}_\delta^{-1} = \rho_\delta\mathbf{G}^{-1} + (\mathbf{H}_\delta\mathbf{L})^T(\mathbf{R}_\delta)^{-1}\mathbf{H}_\delta\mathbf{L}\ . \tag{18}$$

$\mathbf{H}_\delta$ is the observation operator, which projects a (global) state vector onto the local observation domain. Thus, it combines the operation of a global observation operator with the restriction of the observation vector to 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 is performed analogously to the analysis step. The local state ensemble is transformed according to

$$\mathbf{X}^a_\sigma = \overline{\mathbf{X}^a_\sigma} + \sqrt{N-1}\,\mathbf{L}_\sigma(\mathbf{C}_\delta)^T\mathbf{\Omega}^T \tag{19}$$

where $\mathbf{C}_\delta^{-1}(\mathbf{C}_\delta^{-1})^T = \mathbf{U}_\delta^{-1}$. Here, the same transformation matrix $\mathbf{\Omega}$ is used for each local analysis domain to ensure consistent transformations throughout all local domains.

The localization above can be combined with a "covariance localization" [Houtekamer and Mitchell, 2001]. For this, a Schur (element-wise) product is applied to multiply the ensemble covariance matrix element by element with a matrix holding correlations of typically local support. In contrast to the EnKF, the ensemble covariance matrix or its projection onto the observation space is never explicitly computed in the SEIK filter. However, the covariance localization can be performed on the matrix $\mathbf{H}_\delta\mathbf{L}$. In the case that the local analysis domain consists of a single water column and the localization is applied horizontally, we can rewrite equations (17) and (18) as

$$\mathbf{a}_\delta = \mathbf{U}_\delta(\mathbf{D}\circ\mathbf{H}_\delta\mathbf{L})^T(\mathbf{R}_\delta)^{-1}\left(\mathbf{y}^o_\delta - \mathbf{H}_\delta\overline{\mathbf{x}^f}\right), \tag{20}$$

$$\mathbf{U}_\delta^{-1} = \rho_\delta\mathbf{G}^{-1} + (\mathbf{D}\circ\mathbf{H}_\delta\mathbf{L})^T(\mathbf{R}_\delta)^{-1}\,\mathbf{D}\circ\mathbf{H}_\delta\mathbf{L}\ . \tag{21}$$

Here $\circ$ denotes the Schur product and $\mathbf{D}$ is the matrix holding the correlations. Possible functions for the correlations are, for example, an exponential decrease or the use of a polynomial representing a correlation function of compact support [Gaspari and Cohn, 1999], which reduce the influence of observations with growing distance from the local analysis domain. Depending on the form of $\mathbf{R}^{-1}$, this formulation can be equivalent to the down-weighting of distant observations discussed by Nerger and Gregg [2007] and to the covariance localization applied in the EnKF, e.g. by Houtekamer and Mitchell [2001].

C Online bias correction

To correct for model bias, a two-stage online bias correction scheme [Dee and Da Silva, 1998] is applied. Here, a bias vector $\mathbf{b}_k$ of dimension $\tilde{n}$ is estimated analogously to the state estimation described in section 5. For simplicity, we discuss only the online bias estimate for the simplified SEIK filter using a static

covariance matrix. We only show the equations for the global filter. Notes on the localization of the two-stage bias correction scheme are provided after the description of the scheme.

Denoting $\hat{\mathbf{x}}_k$ the de-biased analysis estimate of the state is given by

$$\hat{\mathbf{x}}_k^a = \mathbf{x}_k^a - \mathbf{b}_k^a. \tag{22}$$

We assume that the bias error can be estimated by the state ensemble $\mathbf{X}_k$ analogously to the error of the state estimate. Specifically, we assume that some fraction γ, $0 < \gamma < 1$, accounts for the bias error. Accordingly, the covariance matrix for the online bias estimation is given by

$$\tilde{\mathbf{P}}_k^f = \gamma \mathbf{P}_k^f, \tag{23}$$

while the remaining fraction of the covariance matrix provides the error estimate for the state update:

$$\hat{\mathbf{P}}_k^f = (1-\gamma)\, \mathbf{P}_k^f \tag{24}$$

The evolution of the bias is typically much slower than that of the state. This motivates to assume a persistence model for the forecast of the bias. Thus, the forecast for the bias vector is given by

$$\mathbf{b}_k^f = \mathbf{b}_{k-1}^a. \tag{25}$$

The analysis of the SEIK filter with online bias correction is conducted as a two-stage algorithm. First, the bias vector is updated according to

$$\mathbf{b}_k^a = \overline{\mathbf{x}_k^f} - \gamma \mathbf{L}_k \tilde{\mathbf{a}}_k, \tag{26}$$

$$\tilde{\mathbf{a}}_k = \tilde{\mathbf{U}}_k (\tilde{\mathbf{H}}_k \mathbf{L}_k)^T \tilde{\mathbf{R}}_k^{-1} \left[\tilde{\mathbf{y}}_k^o - \tilde{\mathbf{H}}_k \left(\overline{\mathbf{x}_k^f} - \mathbf{b}_k^f \right) \right], \tag{27}$$

$$\tilde{\mathbf{U}}_k^{-1} = \tilde{\rho} \mathbf{G}^{-1} + (\tilde{\mathbf{H}}_k \mathbf{L}_k)^T \tilde{\mathbf{R}}_k^{-1} \tilde{\mathbf{H}}_k \mathbf{L}_k. \tag{28}$$

Here, $\tilde{\mathbf{H}}$ is the observation operator that provides the part of the observations, which are considered for the bias estimation, that would be measured given the de-biased model state $\overline{\mathbf{x}_k^f} - \mathbf{b}_k^f$. The corresponding observation error covariance matrix is $\tilde{\mathbf{R}}_k$ while the corresponding observation vector is $\tilde{\mathbf{y}}_k^o$. Further, a forgetting factor $\tilde{\rho}$ distinct from that used for the state estimation can be applied.

Subsequently, the state estimate is updated by

$$\mathbf{x}_k^a = \overline{\mathbf{x}_k^f} + (1-\gamma)\,\mathbf{L}_k\hat{\mathbf{a}}_k, \tag{29}$$

$$\hat{\mathbf{a}}_k = \hat{\mathbf{U}}_k(\mathbf{H}_k\mathbf{L}_k)^T{\mathbf{R}_k}^{-1}\left[\mathbf{y}_k^o - \mathbf{H}_k\left(\overline{\mathbf{x}_k^f} - \mathbf{b}_k^a\right)\right], \tag{30}$$

$$\hat{\mathbf{U}}_k^{-1} = \frac{\hat{\rho}}{1-\gamma}\mathbf{G}^{-1} + (\mathbf{H}_k\mathbf{L}_k)^T\mathbf{R}_k^{-1}\mathbf{H}_k\mathbf{L}_k. \tag{31}$$

In equation 29 the biased analysis state estimate is obtained. Together with the bias obtained from equation 26 the de-biased state can be computed using equation 22. Following Keppenne et al. [2005], the forecast is performed by integrating the biased state $\mathbf{x}_k^a$.

The localization of the two-stage analysis can be performed analogously to the localization of the state estimation discussed in section 5. However, it is possible to use different influence distances for the observations for the bias and state updates. In addition, the use of different localizing weightings of the observations is possible.

D Implementation Aspects

For the experiments performed here, the forecast phase is simplified by keeping the state error covariance matrix static. For this, a matrix of ensemble perturbations ($\sqrt{N-1}\,\mathbf{V}_0\mathbf{C}_0^T\mathbf{\Omega}_0^T$ in equation 7) is stored and only the ensemble mean $\overline{\mathbf{x}_i^a}$ is integrated without applying a 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 $\overline{\mathbf{x}_k^f}$.

With regard to the log-normal distribution of chlorophyll, the assimilation algorithm is implemented such that the ensemble $\mathbf{X}$ of model states holds the logarithm of the chlorophyll concentrations. For the model forecast the logarithmic concentrations are back-transformed to actual concentrations which are used in the routines of the numerical model. In addition, the observation vectors hold logarithmic concentrations. Due to this the observation operator $\mathbf{H}_k$ contains only the selection of grid points holding observations at a particular analysis time.

References

J. I. Allen, M. Eknes, and G. Evensen. An Ensemble Kalman Filter with a complex marine ecosystem model: hindcasting phytoplankton in the Cretan Sea. *Ann. Geoph.*, 21:399–411, 2003.

S.-J. Baek, B. R. Hunt, E. Kalnay, E. Ott, and I. Szunyogh. Local ensemble Kalman filtering in the presence of model bias. *Tellus*, 58A:293–306, 2006.

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.

G. A. Chepurin, J. A. Carton, and D. Dee. Forecast model bias correction in ocean data assimilation. *Mon. Wea. Rev.*, 133:1328–1342, 2005.

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.

D. P. Dee and A. M. Da Silva. Data assimilation in the presence of forecast bias. *Q. J. R. Meteorol. Soc.*, 124:269–295, 1998.

D. P. Dee and R. Todling. Data assimilation in the presence of forecast bias: The GEOS moisture analysis. *Mon. Wea. Rev.*, 128:3268–3282, 2000.

G. Gaspari and S. E. Cohn. Construction of correlation functions in two and three dimensions. *Q. J. Roy. Meteor. Soc.*, 125:723–757, 1999.

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 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.

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

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.

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, 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 and W. W. Gregg. Assimilation of SeaWiFS data into a global ocean-biogeochemical model using a local SEIK filter. *J. Mar. Syst.*, 2007. in press, doi:10.1016/j.jmarsys.2006.11.009.

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 Me-*

teorology - Proceedings of the 11. ECMWF Workshop, pages 63–83. World Scientific, 2005b.

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, 1998.

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

G. Triantafyllou, I. Hoteit, and G. Petihakis. A singular evolutive interpolated Kalman filter for efficient data assimilation in a 3-D complex physical-biogeochemical model of the Cretan sea. *J. Mar. Syst.*, 40-41:213–231, 2003.

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.

	Nerger and Gregg [2007]	**this work**
Assimilation algorithm	LSEIK	LSEIK with online bias correction
Update at grid points with ice	neglect data; no update at grid points with non-zero ice concentration	consider all data; perform update relative to percentage of open water
Model	older forcing fields (e.g. dust)	new forcing fields (e.g. dust)
Observation errors	special regions with prescribed larger errors for concentrations $>$ 1mg m^{-3}	neglecting data with concentrations $>$ 1mg m^{-3} in special regions
Model error covariance matrix	based on monthly deviations from 8-year mean state	based on deviation of state at each 15th day from 3-month running mean during 1998–2003
Assimilation period	1/1998–12/2004	9/1997–12/2004

Table 1
Comparison of the differences in the experimental configurations between Nerger and Gregg [2007] and the current work.

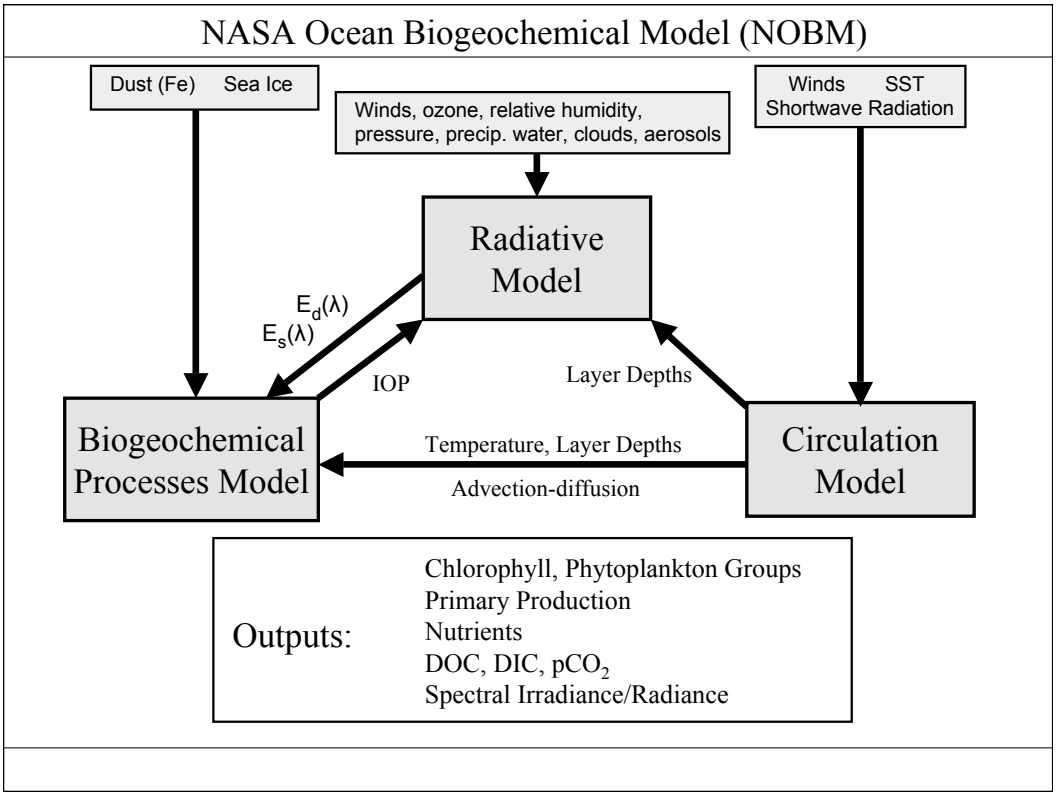

Fig. 1. General structure of the NOBM showing the interactions among the main components. Also shown are forcing fields and nominal outputs.

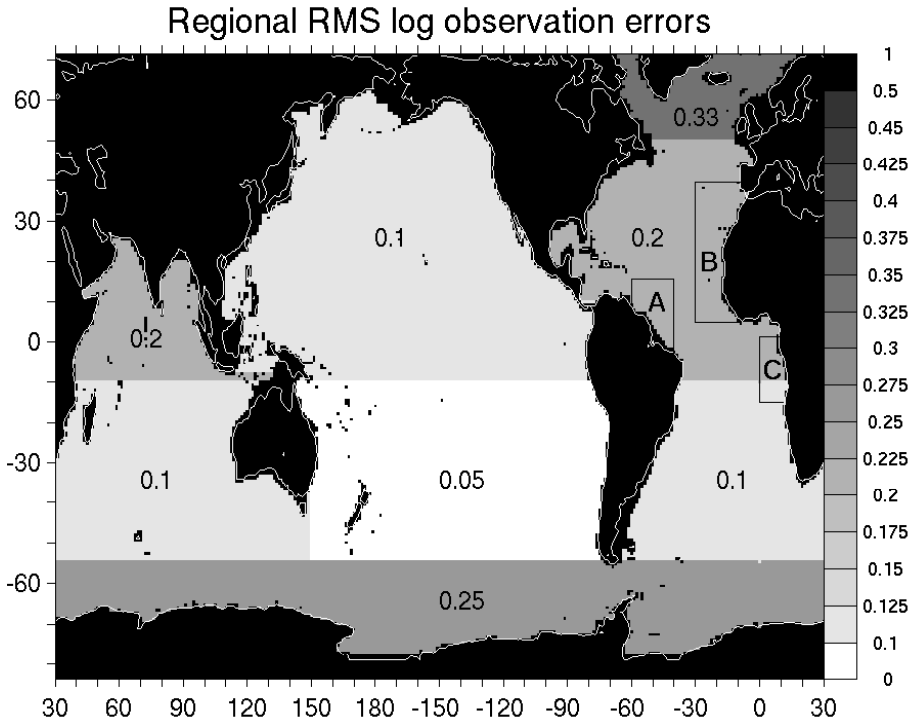

Fig. 2. Observation errors assumed for the assimilation of chlorophyll data. In the regions A to C as well as in the Equatorial and North Indian Ocean, observations with concentrations > 1mg m^{-3} are considered as outliers.

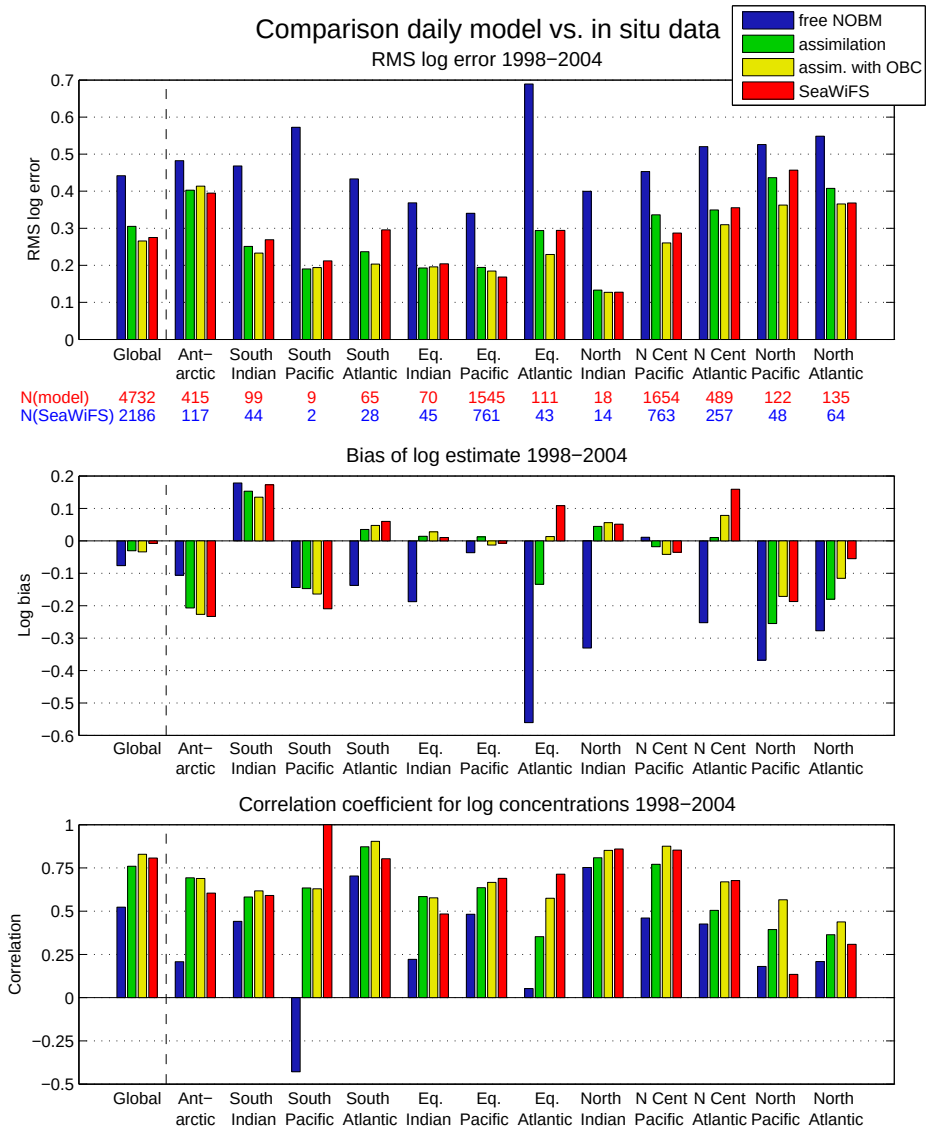

Fig. 3. Comparison of the surface chlorophyll from model and SeaWiFS with in situ data globally and separated over 12 major oceanographic basins for 1998–2004. Top: RMS log error. Middle: Mean 7-year bias of log concentrations. Bottom: Correlation coefficient. Shown are values for (blue) the free-run model, (green) the assimilation estimate without OBC, (yellow) the de-biased estimate with OBC, and (red) SeaWiFS data. Below the uppermost panel the number of collocation points is shown for the model and the SeaWiFS data.

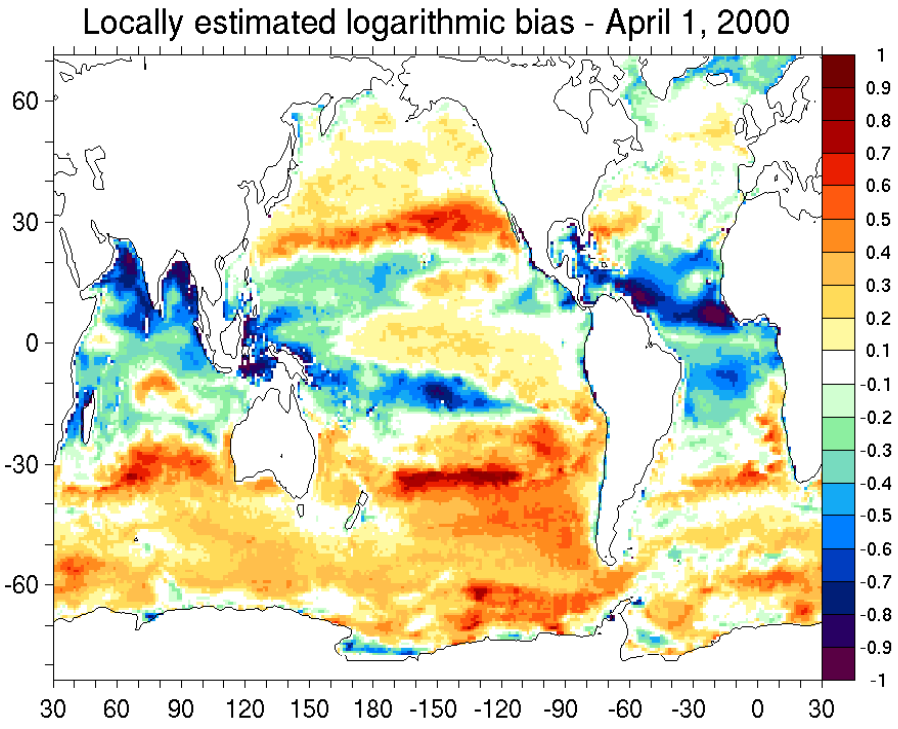

Fig. 4. Bias of logarithmic chlorophyll concentrations estimated by the OBC algorithm for April 1, 2000.

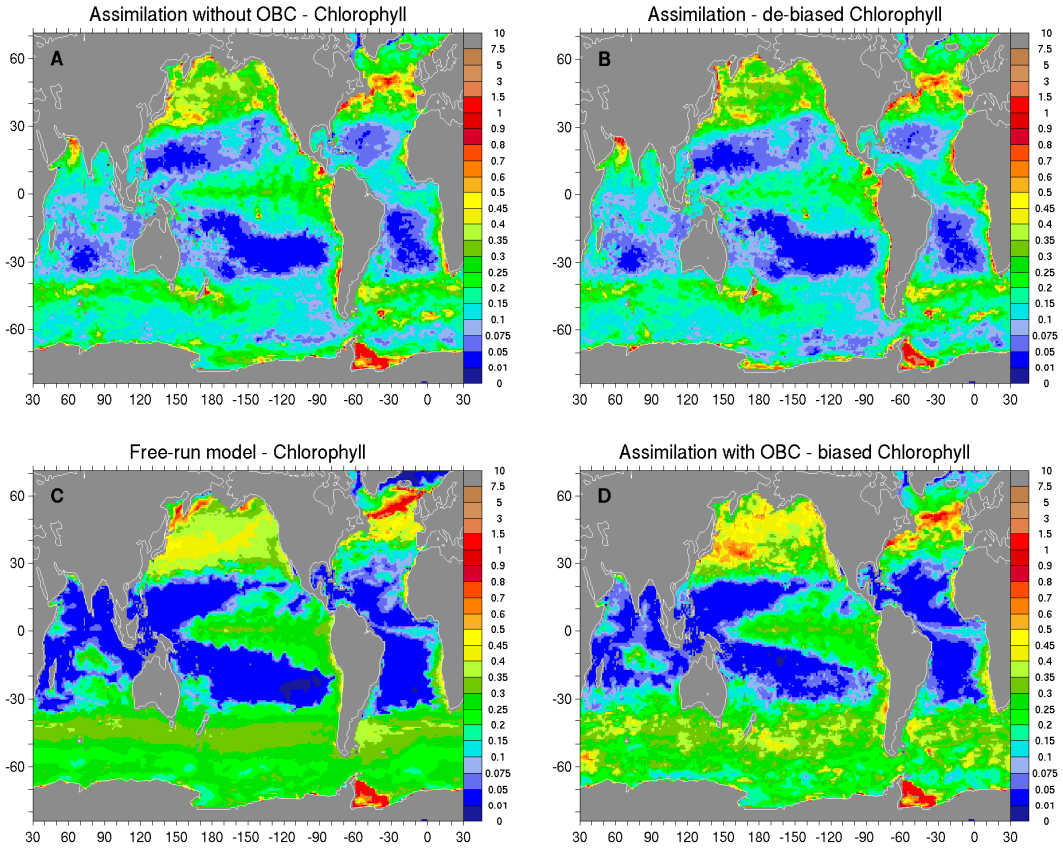

Fig. 5. Comparison of surface total chlorophyll on April 1, 2000: (A) from assimilation without OBC, (B) de-biased estimate from assimilation with OBC, (C) from free-run model, (D) de-biased estimate from assimilation with OBC.