Estimation of global CO_2 fluxes at regional scale using the maximum likelihood ensemble filter

R. S. Lokupitiya,[1] D. Zupanski,[2] A. S. Denning,[1] S. R. Kawa,[3] K. R. Gurney,[4] and M. Zupanski[2]

Received 5 December 2007; revised 21 July 2008; accepted 15 August 2008; published 22 October 2008.

[1] We use an ensemble-based data assimilation method, known as the maximum likelihood ensemble filter (MLEF), which has been coupled with a global atmospheric transport model to estimate slowly varying biases of carbon surface fluxes. Carbon fluxes for this test consist of hourly gross primary production and ecosystem, respiration over land, and air-sea gas exchange. Persistent multiplicative biases intended to represent incorrectly simulated biogeochemical or land-management processes such as stand age, soil fertility, or coarse woody debris were estimated for 1 year at 10° longitude by 6° latitude spatial resolution and with an 8-week time window. We tested the model using a pseudodata experiment with an existing observation network that includes flasks, aircraft profiles, and continuous measurements. Because of the underconstrained nature of the problem, strong covariance smoothing was applied in the first data assimilation cycle, and localization schemes have been introduced. Error covariance was propagated in subsequent cycles. The coupled model satisfactorily recovered the land biases in densely observed areas. Ocean biases, however, were poorly constrained by the atmospheric observations. Unlike in batch mode inversions, the MLEF has a capability of assimilating large observation vectors and hence is suitable for assimilating hourly continuous observations and satellite observations in the future. Uncertainty was reduced further in our pseudodata experiment than by previous batch methods because of the ability to assimilate a large observation vector. Propagation of spatial covariance and dynamic localization avoid the need for prescribed spatial patterns of error covariance centered at observation sites as in previous grid-scale methods.

Citation: Lokupitiya, R. S., D. Zupanski, A. S. Denning, S. R. Kawa, K. R. Gurney, and M. Zupanski (2008), Estimation of global CO_2 fluxes at regional scale using the maximum likelihood ensemble filter, *J. Geophys. Res.*, *113*, D20110, doi:10.1029/2007JD009679.

1. Introduction

[2] The CO_2 concentration in the atmosphere is increasing every year because of anthropogenic activities. About half of the CO_2 released to the atmosphere is absorbed by various land and ocean processes, but spatial and temporal variability of these carbon sinks is not well understood [*Denman et al.*, 2007]. When making policy decisions about CO_2 emissions, it is important to know the spatial distribution of these carbon sinks, how they function, and for how long they will keep operating.

[3] Inverse modeling has been widely used to locate the spatial distribution of the carbon sink by using observed CO_2 concentrations in the atmosphere [e.g., *Gurney et al.*, 2002; *Rödenbeck et al.*, 2003; *Michalak et al.*, 2004; *Bruhwiler et al.*, 2005; *Peters et al.*, 2005; *Baker et al.*, 2006]. The outcome of the inversions varies because of differences in the transport and the spatial representation of the prior fluxes. It depends strongly on the prescribed prior and observation error covariance matrices, which define weighting between the priors and the data for these under-determined problems.

[4] *Gurney et al.* [2002] introduced a model intercomparison experiment, which is widely known as TransCom3, with 16 global atmospheric transport models and model variants. They found a terrestrial carbon sink that is distributed almost evenly among the northern hemispheric continents. The magnitude of the sink was sensitive to transport differences among models. They also found that the CO_2 uptake in the southern ocean was less than calculated from ocean measurements, and this result was not sensitive to the transport models. Early inversions were carried out by dividing the globe into several large regions and by solving for fluxes in monthly or annual time scales. Optimization was done by estimating a single vector of unknowns. This technique is known as a "batch mode" inversion. For

[1]Department of Atmospheric Science, Colorado State University, Fort Collins, Colorado, USA.
[2]Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, Colorado, USA.
[3]NASA Goddard Space Flight Center, Greenbelt, Maryland, USA.
[4]Department of Earth and Atmospheric Science, Purdue University, West Lafayette, Indiana, USA.

Copyright 2008 by the American Geophysical Union.
0148-0227/08/2007JD009679

example, in the TransCom3 experiment, the globe was divided into 22 regions, which consisted of 11 land regions and 11 ocean regions [*Gurney et al.*, 2002]. Large regions were used because of the sparseness of the CO_2 observation sites. One advantage of this technique is that the problem is mathematically over-determined, because the number of unknowns (number of regions times number of time levels) is much less than the available observed information. However, a recent study by *Bruhwiler et al.* [2007] indicates that some regions like South America and Africa are poorly constrained by the current observation network because of the weaker signal from these regions. The batch problem is computationally efficient, even for monthly estimation over many years. However, lumping small basis regions into larger combined regions may lead to aggregation errors [*Kaminski et al.*, 2001; *Engelen et al.*, 2002], because observed CO_2 fields are sensitive to the distribution of sources and sinks within large basis regions. Batch inversion for large regions cannot adjust these finer scale patterns of fluxes, so that errors in subregional spatial or temporal patterns are unavoidably aliased into errors in the mean fluxes. Usually the sampling sites are biased toward the fluxes from nearby grid cells and hence cannot properly represent heterogeneous larger regions. Also in the Transcom3 experiment, the spatial distributions of fluxes within the large source regions were demarcated by hard boundaries, which do not exist in the real situation.

[5] In a Bayesian framework for data assimilation, we optimize a cost function, which consists of two components. Mathematically we define the cost function as

$$C(\beta) = \frac{1}{2}[\mathbf{y} - H(\beta)]^T \boldsymbol{R}^{-1}[\mathbf{y} - H(\beta)] + \frac{1}{2}[\beta - \beta_b]^T \boldsymbol{P}_f^{-1}[\beta - \beta_b], \tag{1}$$

where $\boldsymbol{y}$ is a vector of observations, H is an observation operator, $\boldsymbol{\beta}$ is a vector of unknowns (the state vector we are solving for), $\boldsymbol{\beta}_b$ is the prescribed prior (background) estimate, $\boldsymbol{R}$ is the observation error covariance matrix, and $\boldsymbol{P_f}$ is the forecast (prior) error covariance matrix. The first term of the cost function (equation (1)) controls the difference between the observations and the predicted values. The second term constrains the solution by an a priori (or "background") flux distribution, which is necessary to stabilize the solution in an under-constrained problem. From a statistical point of view, the cost function is the kernel (core of the distribution that depends on the variable $\boldsymbol{\beta}$) of the posterior distribution. The posterior distribution is defined as a product of the likelihood function and the prior distribution. The two terms of the cost function are the kernel of the likelihood function and the kernel of the prior distribution, respectively. Here we find an optimal solution for the variable $\boldsymbol{\beta}$ by maximizing the posterior distribution. This corresponds to the minimizing kernel of the posterior distribution or the cost function. A solution which minimizes the above cost function can be found assuming a linear observation operator ($\boldsymbol{H}$) as

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta_b} + \boldsymbol{PH^T}\left(\boldsymbol{HPH^T} + \boldsymbol{R}\right)^{-1}(\boldsymbol{y} - \boldsymbol{H\beta_b}), \tag{2}$$

$$\boldsymbol{P}_{\hat{\beta}} = \boldsymbol{P} - \boldsymbol{PH^T}\left(\boldsymbol{HPH^T} + \boldsymbol{R}\right)^{-1}\boldsymbol{HP}, \tag{3}$$

where $\hat{\boldsymbol{\beta}}$ is the posterior estimate of the state vector $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}$ is its corresponding posterior covariance [*Tarantola*, 1987].

[6] In large region inversions like the TransCom3 experiment, it is assumed that grid points within a given region are perfectly correlated in space with a constant flux value over some period of time (e.g., monthly). In finer-scale (grid-scale) inversions, if we were to assume that grid boxes are uncorrelated, the number of unknowns becomes extremely large compared to the number of observations. Hence the problem becomes under-determined, but can be solved by the priors. The best practical approach to solving the problem lies between these two extremes: perfectly correlated larger regions and uncorrelated grid boxes. In a grid-scale inversion, we find a solution, which lies in between these two extremes, by correlating the grid cells. The first grid-scale inversion of CO_2 was introduced by *Kaminski et al.* [1999]. They estimated a coarse grid of fluxes at 8° latitude by 10° longitude in monthly time scales. The problem was highly under-constrained and a unique solution was found by gathering a priori information on surface fluxes. *Rödenbeck et al.* [2003] performed a grid-scale inversion accounting for the spatial covariance of flux uncertainties. They assumed different de-correlation length scales over the land and the ocean and monthly fluxes were estimated on 8° latitude by 10° longitude spatial resolution from monthly mean observations. *Michalak et al.* [2004] developed a geostatistical approach, which avoids prescribing a priori fluxes. In their method, $\boldsymbol{\beta}_b$ in equation (1) was replaced by a trend term $\boldsymbol{X\beta}$. This modification to the cost function allows one to include additional information such as vegetation cover, leaf area index, and greenness fraction etc. that varies with the mean behavior of the fluxes. For example, if we assume that the mean behavior of ocean fluxes differs from that of the land, it can be incorporated into the trend term by simply including a variable, which separates the two fluxes. This separation is usually done by including an indicator variable that represents the land by 1 and the ocean by 0 or vise versa. They estimated the parameters of the state covariance matrix, such as de-correlation length scales and variances (land/ocean), as a byproduct of the optimization scheme, rather than prescribing them.

[7] All of these methods utilize the batch mode or synthesis inversion technique and they perform satisfactorily with the existing observation network. Every year new observation sites become available, many of which record CO_2 hourly rather than weekly. The observation vector will be tremendously large when Orbiting Carbon Observatory (OCO) data are available [*Crisp and Johnson*, 2005]. As more observations become available, we will be able to optimize the fluxes in much finer scales. However, batch mode inversions are unwieldy in this case because of the need to invert excessively large matrices. Finer-scale estimation of surface sources and sinks is now becoming feasible, but the computational burden and under-constrained nature of the problem requires innovative assimilation methods. *Bruhwiler et al.* [2005] introduced a fixed-lag Kalman smoother to estimate fluxes. Their method steps through the observations sequentially, which avoids the difficulties of using large observation vectors as in the batch mode inversion technique. However, this method requires

the pre-calculation of observation operators, which is still expensive in the case of assimilating hourly observations. Further developing the fixed-lag Kalman smoother, *Peters et al.* [2005] introduced an ensemble-based approach to carbon inversions, in which the Kalman gain matrix was approximated by using ensemble members. They used the ensemble square root filter, which assimilates observations serially (one at a time) [*Whitaker and Hamill*, 2002]. Serial assimilation could be troublesome in carbon problems because of the need for repeated integration of the transport model. This could be computationally expensive with very large observation vectors as in the case of a satellite experiment. *Baker et al.* [2006] and *Chevallier et al.* [2005] introduced variational data assimilation schemes to atmospheric CO_2 assimilation. Their methods also showed promising results with large observation vectors such as OCO data. However, the variational method requires the calculation of backward-in-time transport, also known as the model adjoint. Frequent improvements to the models are introduced so that the computation and maintenance of model adjoints, which is required in variational methods, also becomes complicated and troublesome. For example, in atmospheric transport models, reversing advection schemes is fairly simple but reversing the convective schemes can be rather difficult because of complicated parameterization schemes with many logical branches. Ensemble methods have the advantage that there is no need to compute model adjoints. The computational cost of both ensemble and variational methods are similar, but ensemble methods are more efficient in a parallel computing environment.

[8] In this paper, we apply a new ensemble-based method called the maximum likelihood ensemble filter (MLEF) [*Zupanski*, 2005; *Zupanski and Zupanski*, 2006] to global CO_2 inversion. The MLEF has also been applied to regional CO_2 inversion [*Zupanski et al.*, 2007a]. This regional-scale study was focused on estimating the biases for GPP and respiration in North America by assimilating continuous CO_2 observations from the WLEF tall tower and the "ring of towers" in northern Wisconsin. They investigated the model performance with a wide range of ensemble sizes and found that a reasonable solution can be reached even with small ensembles by applying covariance localization. For very large ensemble sizes, localization was not essential. In this study, we introduce a pseudodata experiment to test the performance of the MLEF by assimilating currently available (flasks, continuous, and aircraft profiles) observation sites on the global domain. We allow net surface fluxes of CO_2 to vary on an hourly basis, and solve for persistent multiplicative biases of each component flux in each model grid cell. We have assumed an idealized case in this experiment such that the biases stay constant throughout the year, corresponding to errors in slowly-varying biogeochemical or land-management parameters such as forest stand age, nitrogen deposition, or coarse woody debris which are difficult to simulate accurately everywhere. In reality, model biases may vary seasonally or in some other time scale, but these variations are not considered in the present study. The MLEF is feasible for applications with very large observation vectors (e.g., satellite observations) since no serial assimilation of observations is required and hence can be a useful tool in future CO_2 studies. Serial processing of observations was introduced in ensemble square root filter schemes for the purposes of covariance localization [e.g., *Whitaker and Hamill*, 2002; *Peters et al.*, 2005]. In the MLEF, a different approach for covariance localization is taken, so serial processing of observations is not necessary.

[9] The remainder of this paper is organized as follows. In section 2, we describe the inversion scheme we used in this study. Section 3 presents the results along with a discussion based on a pseudodata experiment. Finally, section 4 includes the concluding remarks and future directions of our work.

2. Method

[10] *Evensen* [1994] introduced the first ensemble-based approach to the data assimilation literature. Since then several versions have been introduced by improving the original version [*Houtekamer and Mitchell*, 1998; *Burgers et al.*, 1998; *Whitaker and Hamill*, 2002; *Zupanski*, 2005]. The MLEF has been developed by incorporating ideas from variational methods, iterated Kalman filters, and Ensemble Transform Kalman Filter (ETKF). The cost function is minimized numerically, which allows one to incorporate nonlinear models if necessary. Unlike other ensemble-based methods, the MLEF incorporates iterative minimization of a non-linear cost function with advanced Hessian preconditioning, which makes it more robust for non-linear processes. The method is based on maximum likelihood (rather than minimum variance) estimation and thus the optimal solution is given by the mode (rather than the mean) of the posterior distribution. As explained in the work of *Fletcher and Zupanski* [2006] the maximum likelihood solution is robust for non-Gaussian error distributions and extreme observations (outliers). Mathematical derivations in the study of Fletcher and Zupanski indicate that, when using non-Gaussian Probability Density Functions (PDFs), a different cost function is obtained. Typically, non-Gaussian PDFs (e.g., lognormal) result in more complicated cost functions including extra non-linear terms, which require additional minimization iterations in order to obtain a satisfactory solution. We do not explore this capability of the MLEF in this paper since our observation operator is linear and the PDFs are close to Gaussian. The following section includes a discussion of the assimilation scheme followed by a description of the MLEF method.

2.1. Assimilation Scheme

[11] Previous studies of carbon flux estimation were mostly focused on estimation of Net Ecosystem Exchange (NEE). NEE estimation has been done on weekly, monthly, or yearly time scales. However, much of the variation of fluxes on land lies in sub-daily time scales, which are often neglected. NEE is defined as the difference between two component fluxes, ecosystem respiration and Gross Primary Productivity (GPP). An annual sink at a given location may occur either because of high GPP or low respiration. Similarly, a source can occur with low GPP or high respiration.

[12] In this study, we estimate GPP and respiration by introducing unknown persistent multiplicative adjustments (biases) to them. We solve for the biases by assuming that

they are constant over longer time periods (in this example 8 weeks) compared to the component fluxes. By doing so, we can allow high-frequency time variations in respiration and photosynthesis (i.e., GPP) assuming that they are driven by relatively well-understood and easily-modeled processes [*Zupanski et al.*, 2007a]. This assumption is valid for a pseudodata experiment. However, for real data, any incorrect specification of temporal patterns of fluxes will be aliased into a bias in the recovered "biases."

[13] Currently, it is difficult to differentiate these two flux components from the observations. Nighttime concentrations of CO_2 over land are sensitive to respiration, but they are rarely used in flux inversions because the transport models cannot adequately simulate nocturnal boundary layers. Other tracers such as Carbonyl Sulfide (COS) [*Montzka et al.*, 2007] that are sensitive only to GPP and could be helpful in separating the component fluxes. Further investigation is required in this area. A mesoscale problem of bias estimation was discussed by *Zupanski et al.* [2007a]. In the global-scale problem, however, we need to consider ocean fluxes as well. The optimization problem can be represented as solving for unknown multiplicative biases:

$$F(x,y,t) = \beta_{RESP}(x,y)RESP(x,y,t) - \beta_{GPP}(x,y)GPP(x,y,t) + \beta_{Ocean}(x,y)Ocean(x,y,t), \quad (4)$$

where x and y denote the spatial coordinates and t represents the time, which is at hourly resolution. β's represent persistent multiplicative biases in the grid-scale component fluxes. The rationale for equation (4) is as follows. A persistent bias in photosynthesis might result (for example) from underestimation of available nitrogen, forest management, or agricultural land-use, whereas a persistent bias in respiration might result from overestimation of soil carbon or coarse woody debris. Sub-daily variations in the simulated component fluxes respiration and GPP are primarily controlled by the weather (especially changes in radiation due to clouds and the diurnal cycle of solar forcing), whereas seasonal changes are derived from phenological calculations parameterized from satellite imagery. Fine-scale spatial variations are driven by changes in vegetation cover, soil texture, and soil moisture. In any case, it is reasonable to assume that the biases $\boldsymbol{\beta}_{RESP}$ and $\boldsymbol{\beta}_{GPP}$ vary much more slowly than the fluxes themselves in longer time scales. Our method allows for respiration and GPP to vary on hourly, synoptic, and seasonal time scales, but assumes that biases in these fluxes persist for a period of approximately 2 months. Since the current study is a pseudodata experiment, we neglected the contribution from fossil fuel burning.

[14] Hourly component fluxes (respiration and GPP) were derived from the Simple Bioshere-version 3 (SiB3) model (*Denning et al.* [1996]; *Schaefer et al.* [2002]; *Baker et al.* [2003]; *Baker et al.* [2007], ORNL data set). Ocean fluxes were considered [*Takahashi et al.*, 2002] on a monthly time scale. We interpolated mid monthly values to hourly time resolution to be consistent with the land fluxes. Each flux is prescribed on 10° longitude by 6° latitude spatial resolution and hence the state vector has 1097 (265 land points × 2 + 567 ocean points) unknowns (degrees of freedom). We can write the state vector (in equation (1)) as follows:

$$\beta = \begin{bmatrix} \beta_{GPP} \\ \beta_{RESP} \\ \beta_{Ocean} \end{bmatrix}_{1097\times 1} \quad (5)$$

[15] Unlike in regional inversions, global inversions require a longer period of transport because of the scale of the problem and sparseness of the observation network. Usually, a CO_2 pulse from a given grid point has to travel over several weeks to months in order to reach a far distant observation site. Running large ensembles of months-long forward transport calculations is the most computationally intensive component of the data assimilation scheme. *Bruhwiler et al.* [2005] showed that 4–6 months of transport would effectively capture most of the signal from each source region. *Peters et al.* [2005] investigated the sensitivity of the flux estimates to the length of the assimilation window, and found that an 8–10 week assimilation window would reasonably recover the fluxes, according to their assimilation scheme. Although some influence on very distant observing stations is neglected by using such a short assimilation window, the signals are diluted over time by atmospheric dispersion. The added information from a longer window is not worth the added computational cost. In case of a densely observed system as in satellite data assimilation, a much shorter window can be considered. In NEE flux estimation, the window length would also control the number of unknowns in the state vector, but this is not the case in our bias inversion because the biases are assumed to vary slowly and we assume that they are constant over the window length. So, in this experiment, the window length controls the duration of the biases. We chose an eight-week window in our assimilation scheme. Though we know that biases are slow varying compared to the corresponding component fluxes, their actual persistence is unknown. Thus in a real life experiment, the window length would serve as an important parameter.

[16] In the first assimilation cycle, also known as a "cold start", the forward transport calculation for each ensemble member was started from a single 3-D CO_2 field, which was saved at the end of a three-year spin-up process (see section 2.3). A background (or first guess) $\boldsymbol{\beta}$ (=1), as described in section 2.7, along with the perturbed background vectors (ensemble members) was used to compute the hourly CO_2 fluxes using equation (4). Then each hourly tracer was run through the transport model for 8 weeks (the window length) to simulate CO_2 at the observation sites. The MLEF optimizes $\boldsymbol{\beta}$ by minimizing the distance between the simulated and observed CO_2 concentrations. During each assimilation cycle, the optimized $\boldsymbol{\beta}$ was run through the transport model and at the end of the cycle, the 3-D CO_2 field was saved, which was then used to start the next cycle ("warm start"). Similarly, 3-D CO_2 fields for the ensemble members were also saved, which were used as the starting point for ensemble members in the warm start. Note that unlike in other ensemble methods, in MLEF, analysis perturbations (ensemble members) are propagated to the next cycle through the forecast model so that a new drawing of random perturbations was not required in the next cycle

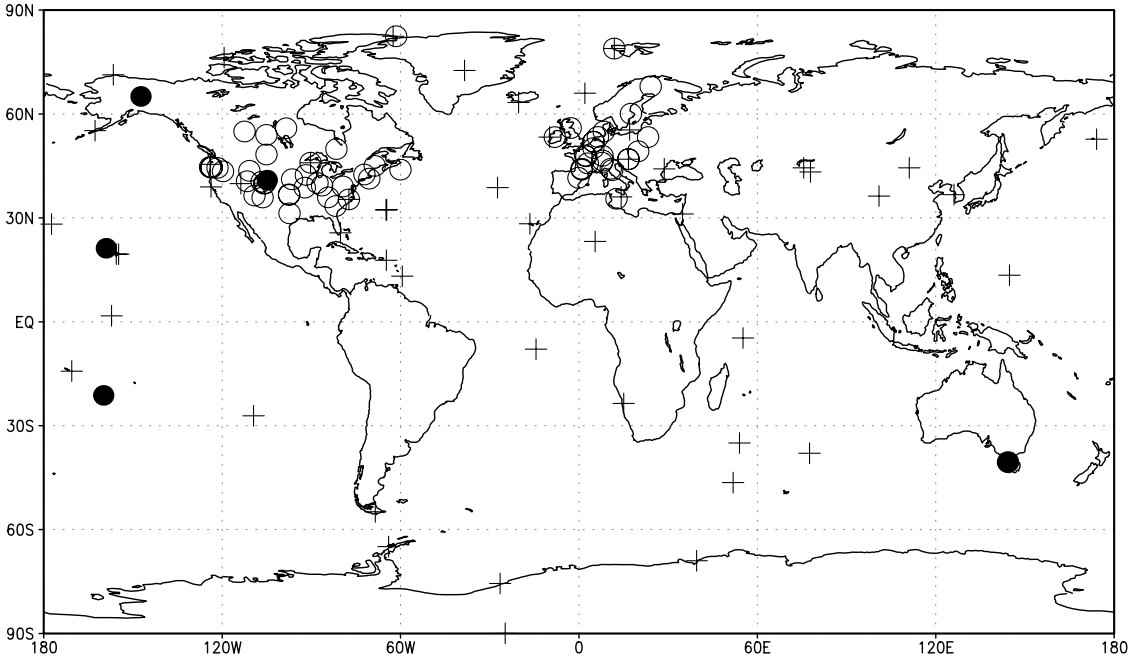

Figure 1. A map of stations used in this study. Solid circles—aircraft profiles, open circles—continuous sites, and plus sign—flask stations.

[*Zupanski*, 2005]. This process was followed for the remaining cycles.

[17] *Peters et al.* [2005, 2007] estimated weekly net carbon fluxes using a moving overlapping assimilation window. Weekly fluxes were estimated repeatedly using subsequent weeks of observations after allowing five weeks of atmospheric transport to propagate information from surface flux locations to relatively distant observation sites in the model. The three-dimensional model state (CO_2 distribution) was propagated between subsequent five-week forward simulations initiated one week apart. This technique provided atmospheric "memory" of fluxes to smooth the solution, avoiding unrealistic discontinuities in simulated CO_2 concentrations, which might result from sudden changes in estimated fluxes from one assimilation interval to the next. We have introduced an explicit temporal decomposition of the flux optimization problem (equation (4)), by which we separate large subdiurnal variations in net fluxes driven by physical forcing from more subtle but persistent multiplicative biases in each component flux because of incorrect biogeochemical or land-management parameters. Persistent grid-scale biases are estimated every eight weeks using observations from the same period, which is sufficiently long for the influence of surface fluxes to be "felt" across much of the observing network. The temporal decomposition used here prevents CO_2 discontinuities in much the same way as the moving window used by *Peters et al.* [2005, 2007], but avoids the need for computationally inefficient overlapping transport integrations.

2.2. Observations

[18] We assimilated three types of pseudo-observations; 53 CMDL surface flask observations that are collected on weekly basis, 5 aircraft profiles that are also collected on weekly basis at different vertical levels, and 67 continuous sites that are measured in-situ on an hourly basis (Note that some continuous sites sample CO_2 at several vertical levels). For locations and names of the observation sites see Table 1 and Figure 1. In this pseudodata experiment, we assumed that each flask station sampled observations with 1 ppm (parts per million) uncertainty. The actual model-data mismatch errors would be estimated for our transport model in a real data experiment. At the continuous sites, uncertainty is added according to the local time and station height as given by Table 2. Higher uncertainties are given to nighttime observations to account for the strong variations in the nighttime CO_2 due to variably stable nighttime boundary layers that are difficult to simulate. Hence they have minimal representation in the assimilation process. The observation error, which is also known as model-data mismatch, corresponds to the diagonal of the observation covariance matrix, $\boldsymbol{R}$. We usually assume that the observation stations are far from each other so that the correlations among their errors are negligible (off-diagonal elements of $\boldsymbol{R}$ are zero). However, model-data mismatches at continuous sites can be correlated over time and may lead to non-zero off-diagonal values in the matrix $\boldsymbol{R}$. We do not address this issue in this paper. Observation error consists of measuring instrument error at the site, transport error, and the error due to scale mismatch between the observations and the transport model (the so-called representativeness error, e.g., *Cohn* [1997] and *Engelen et al.* [2002]). The measurement errors are very small compared to the model transport error.

2.3. Pseudodata

[19] This experiment was conducted with artificially generated biases for GPP, respiration, and ocean fluxes. They served as the "truth" we tried to estimate and were generated as follows. First, each map of bias was created by generating random numbers from Gaussian distributions;

Table 1. CO_2 Measurement Sites Used in This Study

Code	Name	Latitude (deg)	Longitude (deg)	Altitude/Height (m)
	Flasks			
ALT	Alert, Nunavut, Canada	82.45	−62.52	210
AMS	Amsterdam Island, Indian ocean (France)	−37.95	77.53	150
ASC	Ascension Island, UK	−7.92	−14.42	54
ASK	Assekrem. Algeria	23.18	5.42	2728
AVI	St. Croix, Virgin Islands, USA	17.75	−64.75	3
AZR	Terceira Island, Azores, Portugal	38.77	−27.38	40
BAL	Baltic Sea, Poland	55.50	16.67	28
BME	St. Davis Head, Bermuda, UK	32.37	−64.65	30
BMW	Tudor Hill, Bermuda, UK	32.27	−64.88	30
BRW	Barrow, Alaska, USA	71.32	−156.60	11
BSC	Black Sea, Constanta, Romania	44.17	28.68	3
CBA	Cold Bay, Alaska, USA	55.20	−162.72	25
CGO	Cape Grim, Tasmania, Australia	−40.68	144.68	94
CHR	Christmas Island, Republic of Kiribati	1.70	−157.17	3
CMO	Cape Meares, Oregon, USA	45.48	−123.97	30
CRZ	Crozet Island, France	−46.45	51.85	120
EIC	Easter Island, Chile	−27.15	−109.45	50
GOZ	Dwejra Point, Gozo, Malta	36.05	14.18	30
GMI	Mariana Island, Guam	13.43	144.78	6
HBA	Halley Station, Antarctica, UK	−75.58	−26.50	33
HUN	Hegyhatsal, Hungary	46.95	16.65	344
ICE	Storhofdi, Vestmannaeyjar, Iceland	63.25	−20.15	127
ITN	Grifton, North Carolina, USA	35.35	−77.38	60
IZO	Tenerife, Canary Islands, Spain	28.30	−16.48	2360
KEY	Key Biscayne, Florida, USA	25.67	−80.20	3
KUM	Cape Kumukahi, Hawaii, USA	19.52	−154.82	3
KZD	Sary Taukum, Kazakhstan	44.45	77.57	412
KZM	Plateau Assy, Kazakhstan	43.25	77.88	2519
LEF	Park Falls, Wisconsin, USA	45.93	−90.27	483
MBC	Mould, Northwest Territories, Canada	76.25	−119.35	58
MHD	Mace Head, County Galway, Ireland	53.33	−9.90	25
MID	Sand Island, Midway, USA	28.22	−177.37	4
MLO	Mauna Loa, Hawaii, USA	19.53	−155.58	3397
NMB	Nambia	−23.58	15.03	408
NWR	Niwot Ridge, Colorado, USA	40.05	−105.58	3475
PSA	Palmer Station, Antarctica, USA	−64.92	−64.00	10
PTA	Point Arena, California, USA	38.95	−123.73	17
RPB	Ragged Point, Barbados	13.17	−59.43	45
SEY	Mahe Island, Seychelles	−4.67	55.17	3
SHM	Shemya Island, Alaska, USA	52.72	174.10	40
SMO	Tutuila, American Samoa	−14.25	−170.57	42
SPO	South Pole, Antarctica, USA	−89.98	−24.80	2810
STC	Ocean Station C, North Atlantic Ocean, USA	−35.00	54.00	6
STM	Ocean Station M, Norway	66.00	2.00	7
SUM	Summit, Greenland	72.58	−38.48	3238
SYO	Syowa Station, Antarctica, Japan	−69.00	39.58	11
TAP	Tae-ahn Peninsula, South Korea	36.73	126.13	20
TDF	Tierra Del Fuego, La Redonda Island, Argentina	−54.87	−68.48	20
UTA	Wendover, Utah, USA	39.90	−113.72	1320
UUM	Ulaan Uul, Mongolia	44.45	111.10	914
WIS	Sede Boker, Negev Desert, Israel	31.13	34.88	400
WLG	Mt. Waliguan, Peoples Republic of China	36.29	100.90	3810
ZEP	Ny-Alesund, Svalbard, Norway and Sweden	78.90	11.88	475
	Continuous Sites			
ALT	Alert, Nunavut, Canada	82.45	−62.52	210
AMT	Argyle, Maine, USA	45.03	−68.68	159
ARM	Atmospheric Radiation Measurement Site, Oklahoma, USA	36.78	−97.50	314
CBW		52.00	4.90	Multiple
CDL		53.87	−104.65	489
FRS		49.88	−81.57	210
HRV	Harvard Forest, Massachusetts, USA	42.90	−72.30	340
HUN	Hegyhatsal, Hungary	46.95	16.65	Multiple
LEF	Park Falls, Wisconsin, USA	45.92	−90.27	Multiple
NGL		53.17	13.30	65
ORL		47.80	2.50	Multiple
PLR		45.93	7.70	3480
PLS		67.97	24.12	565
SGP	Southern Great Plains, Oklahoma, USA	36.61	−97.49	60
SOBS	Canada	53.98	−105.12	25
SSL		47.92	7.92	1205
WKT	Moody, Texas, USA	31.32	−97.32	Multiple

Table 1. (continued)

Code	Name	Latitude (deg)	Longitude (deg)	Altitude/Height (m)
WPL		55.00	−112.50	550
ZEP	Ny-Alesund, Svalbard, Norway and Sweden	78.90	11.88	475
	Rowley, Iowa, USA	42.40	−91.84	400
	Homer, Illinois, USA	40.07	−87.92	400
	Tower s.carol	33.41	−81.83	300
	Martha's Vineyard	41.33	−70.52	10
	Sable Island	43.93	−60.02	20
	Boreas NOBS	55.88	−98.48	30
	Canaan Valy, West Virginia, USA	39.06	−79.42	30
	Chestnut Ridge, Tennessee, USA	35.93	−84.33	30
	Mead, Nebraska, USA	41.16	−96.47	10
	Morgan Mon., Indiana, USA	39.32	−86.41	30
	Fort Peck, Montana, USA	48.31	−105.10	30
	Ozark, Missouri, USA	38.74	−92.20	30
	Storm Peak Lab	40.45	−106.73	9
	Fraser Exp For	39.90	−105.88	18
	Niwot Ridge, T-Van	40.05	−105.58	5
	Hidden Peak, Utah, USA	40.56	−111.64	18
	Roof Butte Lookout, Navajo Reservation, USA	36.46	−109.10	20
	Pajarito Mt, New Mexico, USA	35.89	−106.39	20
	Jackson Hole Summit, Wyoming, USA	43.59	−110.85	10
	Fir (summit)	44.65	−123.55	15
	Metolius	44.45	−121.56	15
	Yaquina Head	44.67	−124.07	15
	Mary's Peak	44.50	−123.55	15
	NGBER (Burns)	43.45	−119.72	15
	Carbo Europe (CE) Towers	53.32	−8.12	
	Carbo Europe (CE) Towers	78.90	11.88	
	Carbo Europe (CE) Towers	35.52	12.63	
	Carbo Europe (CE) Towers	45.75	3.00	
	Carbo Europe (CE) Towers	47.92	7.92	
	Carbo Europe (CE) Towers	44.18	10.70	
	Carbo Europe (CE) Towers	45.93	7.70	
	Carbo Europe (CE) Towers	46.55	7.98	
	Carbo Europe (CE) Towers	53.38	6.37	
	Carbo Europe (CE) Towers	54.93	8.32	
	Carbo Europe (CE) Towers	49.23	19.93	
	Carbo Europe (CE) Towers	67.97	24.12	
	Carbo Europe (CE) Towers	82.45	−61.48	
	Carbo Europe (CE) Towers	51.97	4.92	
	Carbo Europe (CE) Towers	47.97	2.10	
	Carbo Europe (CE) Towers	50.15	4.87	
	Carbo Europe (CE) Towers	53.33	23.25	
	Carbo Europe (CE) Towers	55.95	−2.78	
	Carbo Europe (CE) Towers	46.95	16.65	
	Carbo Europe (CE) Towers	43.80	11.20	
	Carbo Europe (CE) Towers	60.08	17.47	
	Carbo Europe (CE) Towers	41.58	−0.17	
	Carbo Europe (CE) Towers	44.20	0.90	
	Carbo Europe (CE) Towers	43.90	0.95	
	Aircraft Profiles			
CAR	Briggsdale, Colorado, USA	40.90	−104.80	Multiple
HAA	Molokai Island, Hawaii, USA	21.23	−158.95	Multiple
PFA	Poker Flat, Alaska, USA	65.07	−147.29	Multiple
RTA	Rarotonga, Cook Islands	−21.25	−159.83	Multiple
AIA	Bass Strait/Cape Grim, Australia	−40.53	144.30	Multiple

$\beta_{GPP} \sim N(1,0.3^2)$, $\beta_{RESP} \sim N(1,0.3^2)$, and $\beta_{Ocean} \sim N(1,0.2^2)$. Finally, we introduced kernel smoothing to create large features assuming different scaling factors over land and ocean. At a given grid box, kernel smoothing computes a weighted average of the surrounding grid boxes such that the weights are inversely proportional to the distances. Here we considered the Gaussian distribution as the kernel or the weighting function. The final maps have means approximately equal to 1.0 for all 3 biases and standard deviations approximately equal to 0.2 and 0.06 for land and ocean biases, respectively.

[20] We created pseudo-observations (CO_2 concentrations) by running the transport model forward for four years with the biased fluxes, holding the biases constant throughout the years. In the fourth year, CO_2 concentrations were sampled at the observation stations. Each observation was randomly perturbed by an error according to the specified uncertainty level at the given station (see section 2.2). At the end of the third year, the 3-D model state (CO_2 concentra-

Table 2. Observation Errors in ppm (σ_{Obs}) at the Continuous Sites According to the Local Time and Station Height

	Station Height in Meters (h)		
Local Time (t)	$h < 50$	$50 \leq h < 200$	$h \geq 200$
0	20	10	1
1	20	10	1
2	20	10	1
3	20	10	1
4	20	10	1
5	20	10	1
6	20	10	1
7	15	10	1
8	10	5	1
9	5	5	1
10	5	1	1
11	1	1	1
12	1	1	1
13	1	1	1
14	1	1	1
15	1	1	1
16	1	1	1
17	5	1	1
18	5	5	1
19	10	5	1
20	15	10	1
21	20	10	1
22	20	10	1
23	20	10	1

tion) was saved, which was then used as the starting point for the assimilation scheme.

2.4. Transport Model

[21] Inverse modeling methods require a transport model, which serves as the observation operator (H in equation (1)) in the assimilation scheme. The observation operator performs the necessary interpolations and transformations from the state variable to the observation space. In the carbon problem, the transport model converts CO_2 fluxes on the Earth's surface to CO_2 concentrations in the atmosphere. In this study, we used the Parameterized Chemistry Transport Model (PCTM) as the observation operator [*Kawa et al.*, 2004]. The core of the PCTM code consists of the semi-Lagrangian advection scheme developed by *Lin and Rood* [1996]. Subgrid-scale transport processes such as convection and boundary layer turbulence have been included. The model is driven by assimilated weather data from the GEOS-4 (Goddard Earth Observation System, version 4) reanalyses.

[22] In this study, PCTM was run at 10° longitude by 6° latitude horizontal resolution with 25 vertical levels. The model integration time step was 1 hour, which was consistent with the assumed spatial resolution. The transport was run at such a coarse spatial resolution for testing purposes because it speeds up the forward run, which is the most time consuming part in the assimilation scheme. The number of degrees of freedom are also reduced in the state vector. However, at coarse resolution, the transport becomes unrealistic. In real life problems, much finer scale horizontal resolution would be used in order to get a better match with the real observations. In a real data assimilation experiment, biases could still be estimated at coarser resolution (e.g., 10° longitude by 6° latitude), while running the transport in high resolution (2.5° longitude by 2° latitude).

2.5. MLEF

[23] We minimize the cost function given in equation (1) via an iterative conjugate-gradient algorithm, which converges in a single iteration to the Kalman filter (KF) solution given in (2), when the observation operator H in equation (1) is linear and the ensemble size is equal to the size of the control variable (theoretical proof given in *Zupanski* [2005], Appendix A). Here the control variable is the vector of unknowns. However, in the experiments presented, the ensemble size is considerably smaller than the size of the control variable, which might result in somewhat degraded MLEF solution, compared to the KF solution. As demonstrated by *Zupanski et al.* [2007a] the MLEF solution smoothly converges to the KF solution as the ensemble size approaches the size of the control vector, which provides a justification for using smaller ensemble sizes (in this experiment 200 ensemble members relative to the 1097 unknowns), with the benefit of reduced computational cost.

[24] There are two major steps involved in a data assimilation cycle: (1) the analysis step, and (2) the forecast step. In the analysis step, we find an optimal state by minimizing the cost function (equation (1)), provided the observations. In the forecast step, a prior for the next cycle is found by applying a forecast model to the optimal state found at the analysis step (see equations (8) and (9)). At the first assimilation cycle (cold start), the background or initial guess serves as the forecast.

[25] The prior and the posterior uncertainties of the MLEF solution are defined in ensemble subspace as square roots of the forecast error covariance $\mathbf{P}_f^{\frac{1}{2}}$ and the analysis error covariance $\mathbf{P}_a^{\frac{1}{2}}$:

$$\boldsymbol{P}_a^{\frac{1}{2}} = \boldsymbol{P}_f^{\frac{1}{2}}(\boldsymbol{I} + \boldsymbol{A})^{-\frac{1}{2}}, \tag{6}$$

where

$$\boldsymbol{A} = \boldsymbol{P}_f^{T/2}\boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{H}\boldsymbol{P}_f^{1/2} \tag{7}$$

[26] The matrix $\boldsymbol{A}$ of dimension $N_{ens} \times N_{ens}$ (N_{ens} being the ensemble size) is the so-called information matrix in ensemble subspace [*Zupanski et al.*, 2007b] and is used in this study as a guidance when determining the necessary ensemble size. Selection of ensemble size is crucial in an ensemble-based assimilation technique. Ensembles that are too small deteriorate the quality of the final solution whereas too large ensembles increase the computational cost. In order to determine adequate ensemble size we evaluated an information measure referred to as Degrees of Freedom for Signal - DFS [e.g., *Purser and Huang*, 1993; *Rodgers*, 2000; *Rabier et al.*, 2002; *Fisher*, 2003; *Zupanski et al.*, 2007b]. The DFS, being a positive integer number limited by the ensemble size and the number of observations, was considered a good indicator of whether selected ensemble size was appropriate. We considered the selected ensemble size appropriate if further increase in the ensemble size did not result in the further increase of the DFS. In this study, we selected 200 ensemble members based on this measure.

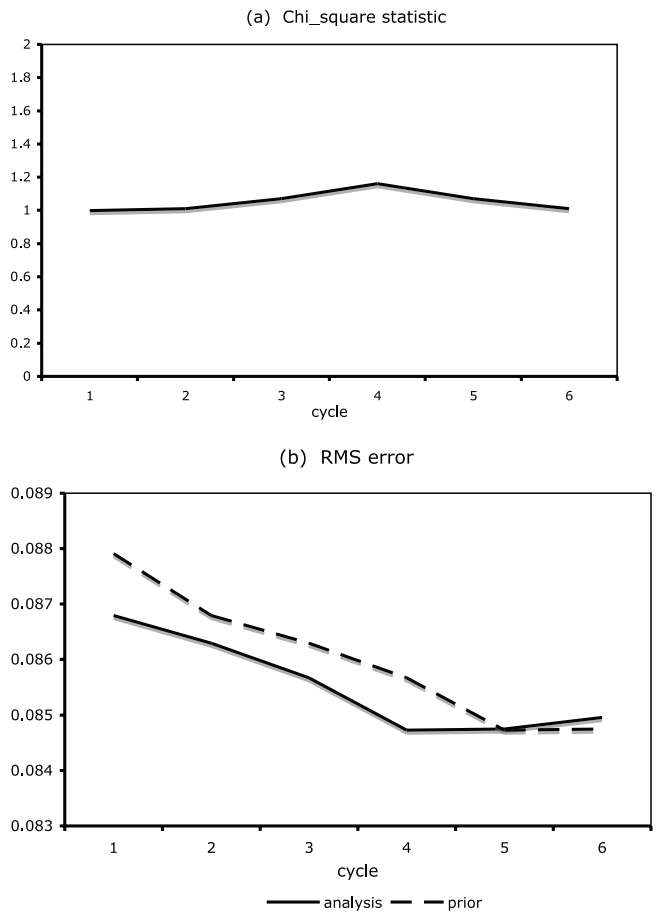

Figure 2. Chi-square statistic for each cycle is given in Figure 2a. A plot of root mean square (RMS) error with respect to the analysis and the prior is shown in Figure 2b.

[27] We have assumed that the forecast model for the bias is an identity model $M = \boldsymbol{I}$. The forecast model explains the evolution of state vector over time defined as

$$\beta_f(t+1) = M[\beta_a(t)], \tag{8}$$

and similarly evolution of covariance matrix is given by

$$\boldsymbol{P}_f(t+1) = M\boldsymbol{P}_a(t)M^T + \boldsymbol{Q}, \tag{9}$$

where **Q** represents the errors induced by imperfect forecast model, which is usually neglected. Hence according to our assumption, the previous analysis state becomes the prior state for the next cycle. Biogeochemically, the use of the identity matrix as a forecast means that we assume persistent biases in modeled fluxes, but allow observations to correct the assumption in each successive assimilation cycle.

2.6. Covariance Smoothing and Localization

[28] Strong covariance smoothing in the first data assimilation cycle and covariance localization in all cycles are required in this problem because of the sparseness of the observing network. Previous large region inversions did not require such schemes because the inversions were carried out by prescribing perfectly correlated spatial structures for several large regions, which were over-determined by the observations. Large region inversions can be considered as an extreme case of covariance smoothing.

2.6.1. Smoothing

[29] We introduce spatial correlations by assuming that grid points are correlated according to an exponential decay function [*Rödenbeck et al.*, 2003; *Michalak et al.*, 2004; *Peters et al.*, 2005]. Hence given any two grid points i and j, we can formulate the state covariance function as

$$\boldsymbol{P}_f = \begin{bmatrix} \sigma_g^2 & \sigma_g^2 e^{-\frac{d_{ij}}{L_g}} & 0 & 0 & 0 & 0 \\ \sigma_g^2 e^{-\frac{d_{ij}}{L_g}} & \sigma_g^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_r^2 & \sigma_r^2 e^{-\frac{d_{ij}}{L_r}} & 0 & 0 \\ 0 & 0 & \sigma_r^2 e^{-\frac{d_{ij}}{L_r}} & \sigma_r^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_o^2 & \sigma_o^2 e^{-\frac{d_{ij}}{L_o}} \\ 0 & 0 & 0 & 0 & \sigma_o^2 e^{-\frac{d_{ij}}{L_o}} & \sigma_o^2 \end{bmatrix}, \tag{10}$$

where d_{ij} is the great circle distance between ith and jth grid points; σ_g^2, σ_r^2, and σ_o^2 are variances of GPP, respiration, and ocean biases respectively; L_g, L_r, and L_o are de-correlation length scales for GPP, respiration, and ocean biases respectively.

[30] Unlike previous studies [*Michalak et al.*, 2004; *Peters et al.*, 2005, 2007], smoothing is introduced only at the first cycle (cold start) of the MLEF. In MLEF, the cross-correlations among the biases are not exactly zero as in equation (10). We define the covariance matrices in ensemble space (reduced rank space) as square roots of the forecast error covariance $\boldsymbol{P}_f^{\frac{1}{2}}$. In the first cycle, each column of the matrix is generated by drawing a random realization from the normal distribution with mean 0 and variance σ_β^2 (variance of the bias parameter) and then by adding exponential spatial smoothing. Since we never explicitly use the full covariance matrix $\boldsymbol{P}_f = \boldsymbol{P}_f^{\frac{1}{2}}\boldsymbol{P}_f^{\frac{T}{2}}$, we cannot apply smoothing to it. Thus the full covariance will indicate non-zero correlations between the distant grid points and the cross-correlations between GPP and respiration. Eventually, in the assimilation process, correlations and cross-correlations that are absorbed from the observations are added. Because it reduces the number of degrees of freedom, covariance smoothing also helps in reducing the number of ensemble members required for the analysis.

[31] The actual length scales of persistent biogeochemical model biases are not known. One could make a reasonable guess by observing the spatial correlations of the auxiliary variables such as soil moisture, soil temperature, leaf area index, or nutrient levels, which may vary similarly to the biases. In this study, we chose to smooth initial covariance with an e-folding length of 800 km over the land points and 1600 km over the ocean points. These parameters were selected according to previous studies [*Michalak et al.*, 2004; *Peters et al.*, 2005, 2007] even though those studies

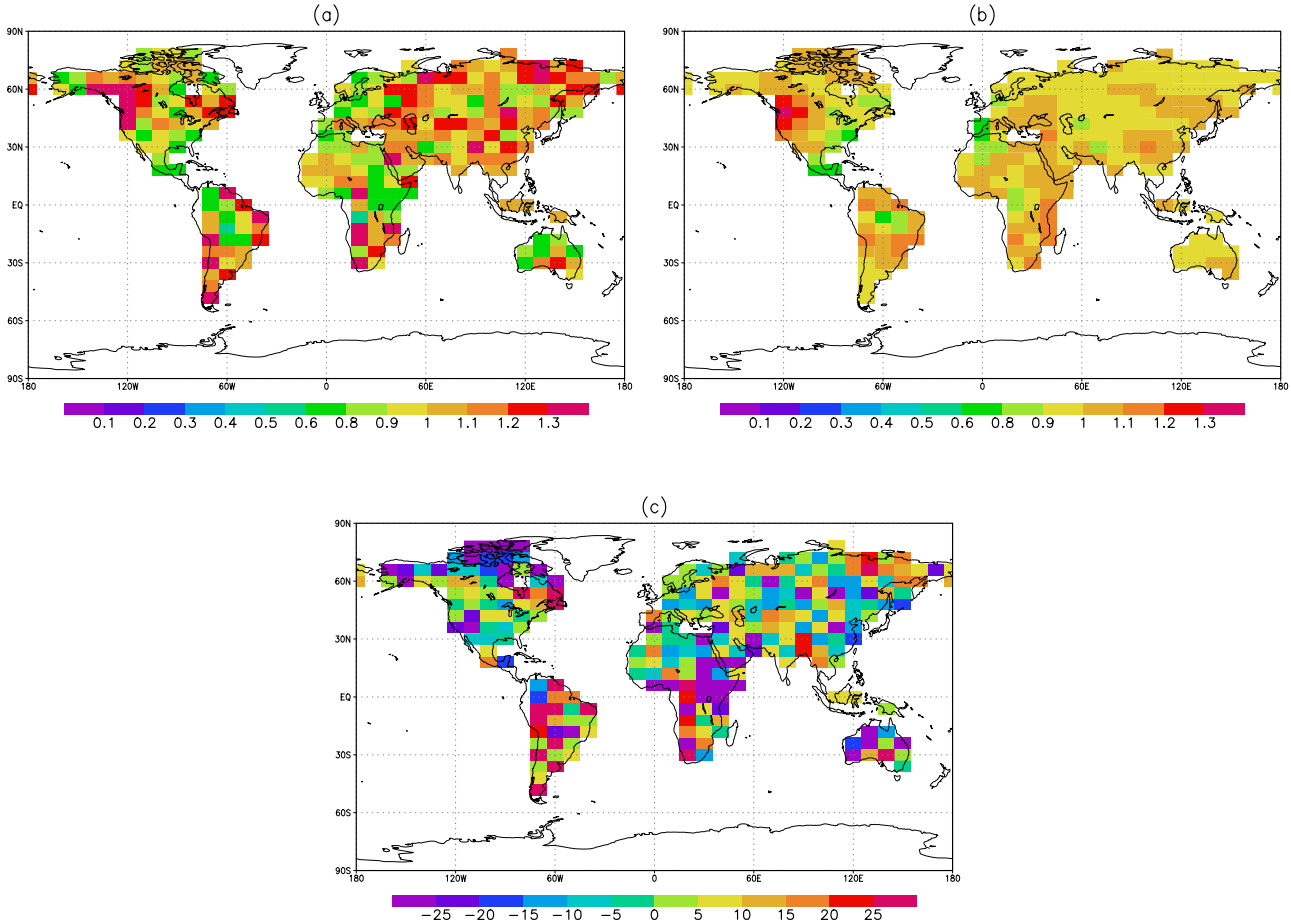

Figure 3. (a) Pseudotruth, (b) recovered, and (c) percentage difference relative to the pseudotruth for GPP bias (β_{GPP}). Note that the prior equals to 1 everywhere (yellow color).

were focused on flux estimation. For the first assimilation cycle only, we assumed that the spatial correlations are isotropic (does not depend on the direction), which may not be true in the real situation. However, in subsequent cycles, the assimilation system "learns" about spatial covariance in the biases of the component fluxes from the data in non-isotropic ways.

2.6.2. Localization

[32] Covariance localization helps to constrain the data assimilation problem with either limited number of observations or limited ensemble members [*Houtekamer and Mitchell*, 1998, 2001; *Peters et al.*, 2005; *Zupanski et al.*, 2007a]. Usually the ensemble size is of the order of hundreds because of the computational limitations, whereas the number of degrees of freedom could be in the order of thousands or millions. Thus the approximation of the state covariance matrix by a limited number of ensemble members may cause large sampling errors in the covariance between distant points. Localization is important in ensemble data assimilation because it prevents sampling errors at large distances and thereby reduces the ensemble size required for the analysis dramatically.

[33] *Peters et al.* [2005] introduced a localization scheme based on an exponential decay function, which imposed circular regions of flux covariance centered at the observation sites. However the flux patterns may be neither exactly circular nor centered at observation sites in carbon problem due to the transport. In some occasions, an influence region may not even contain the observation location because of patterns of advection of CO_2 fields. For example, the distance over which NEE influences CO_2 is much greater in the direction of the wind than across wind, and varies strongly with the degree of vertical mixing. We introduced a localization scheme, which is sensitive to dynamical changes in the analysis and forecast uncertainties [*Zupanski et al.*, 2007a]. To define a "distance" for covariance localization, we employed the ratio $\boldsymbol{r}$ between the forecast and the analysis uncertainty [or in other words, the ratio between the prior ($\boldsymbol{\sigma}_{Prior}$) and the posterior ($\boldsymbol{\sigma}_{Posterior}$) uncertainty] defined as

$$\boldsymbol{r} = \left[\frac{\sigma_{Prior}}{\sigma_{Posterior}}\right], \tag{11}$$

where the prior uncertainty is defined as $\sigma_{Prior} = \left[diag\left(\boldsymbol{P}_f\right)\right]^{\frac{1}{2}}$ and the posterior uncertainty is defined as $\sigma_{Posterior} = \left[diag\left(\boldsymbol{P}_a\right)\right]^{\frac{1}{2}}$.

[34] Note that both the prior and posterior uncertainty correspond to the same time (same data assimilation cycle). The problem here is that posterior uncertainty $\sigma_{Posterior}$ is

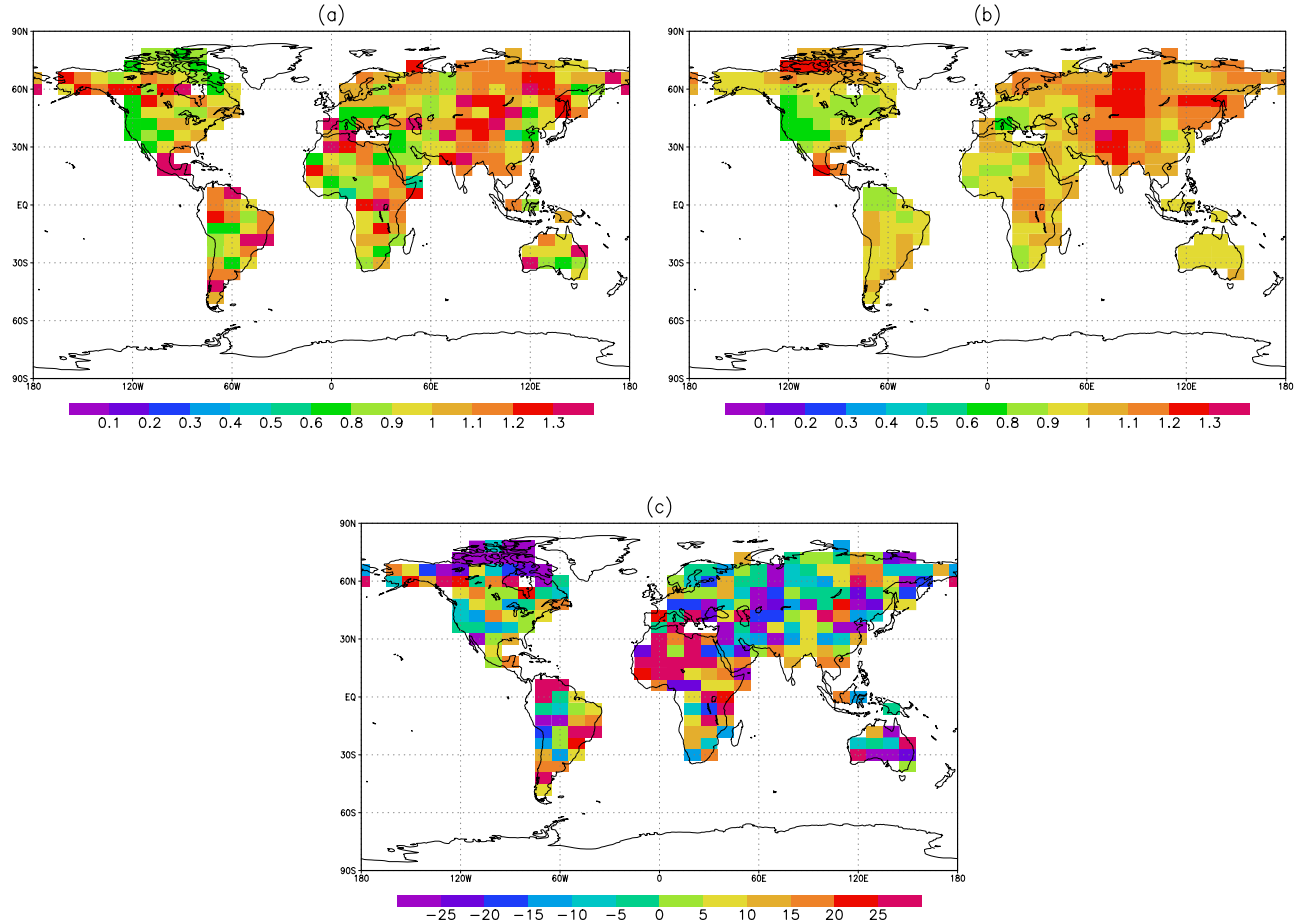

Figure 4. (a) Pseudotruth, (b) recovered, and (c) percentage difference relative to the pseudotruth for respiration bias (β_{RESP}). Note that the prior equals to 1 everywhere (yellow color).

only available after the "optimal" solution has already been obtained, but we need it earlier, at the beginning of the minimization of the cost function. However, an estimate of $\sigma_{Posterior}$ is obtained as a by-product of the Hessian preconditioning [e.g., *Zupanski*, 2005], thus it is available for calculating the ratio (11). The first estimate of $\sigma_{Posterior}$ is identical to the final $\sigma_{Posterior}$ for linear models, while for non-linear models it only approximates the final $\sigma_{Posterior}$. In this application, the model (PCTM) is very close to linear, thus the initial and final estimates of $\sigma_{Posterior}$ are identical.

[35] According to information theory [e.g., *Rodgers*, 2000], and also equation (6) of this paper, the ratio between the prior and the posterior error covariance matrices measures the information content of the assimilated observations. Thus one can interpret the ratio $\boldsymbol{r}$ as a "transport-weighted distance" defined in the space of the information measures. Note that this distance is different from the geodesic distance, which is used in most covariance localization approaches in the current literature. For example $\boldsymbol{r}$ decreases more slowly with distance in the direction of transport than in other directions.

[36] The ratio $\boldsymbol{r}$ is always greater than or equal to one. The greater values of the ratio represent the areas with the greater influence from the observations. We selected the influence regions based on the distributions of the ratio. The land and the ocean regions were selected separately because of the different magnitudes of the flux fields. In this study, we selected 60% of land points and 10% of ocean points based on the upper tail values of the ratio distribution.

2.7. Defining Priors

[37] In real situations, we do not have any prior information about the biases. As a starting point we assume that the carbon fluxes are unbiased, thus we use $\boldsymbol{\beta}_{GPP} = \boldsymbol{\beta}_{RESP} = \boldsymbol{\beta}_{Ocean} = 1$ at every grid point as priors in the first data assimilation cycle. In all subsequent cycles, the estimated biases from the previous cycle are used as priors. Selection of the prior uncertainties for biases is crucial in data assimilation. Choosing too tight or too loose uncertainties may prevent reaching a reasonable solution. Because we know the (simulated) "true" biases in this particular experiment, we could have prescribed uncertainties as the difference between the truth and the prior. However, in a real situation, such an estimator does not exist. Hence we assumed constant prior standard deviations for the biases at each grid point ($\sigma_{GPP} = 0.1$, $\sigma_{RESP} = 0.1$, and $\sigma_{Ocean} = 0.03$) to be consistent with a real experiment.

[38] According to Bayesian theory, the posterior distribution lies in between the prior and the observed distributions. However, if little information is available, the posterior

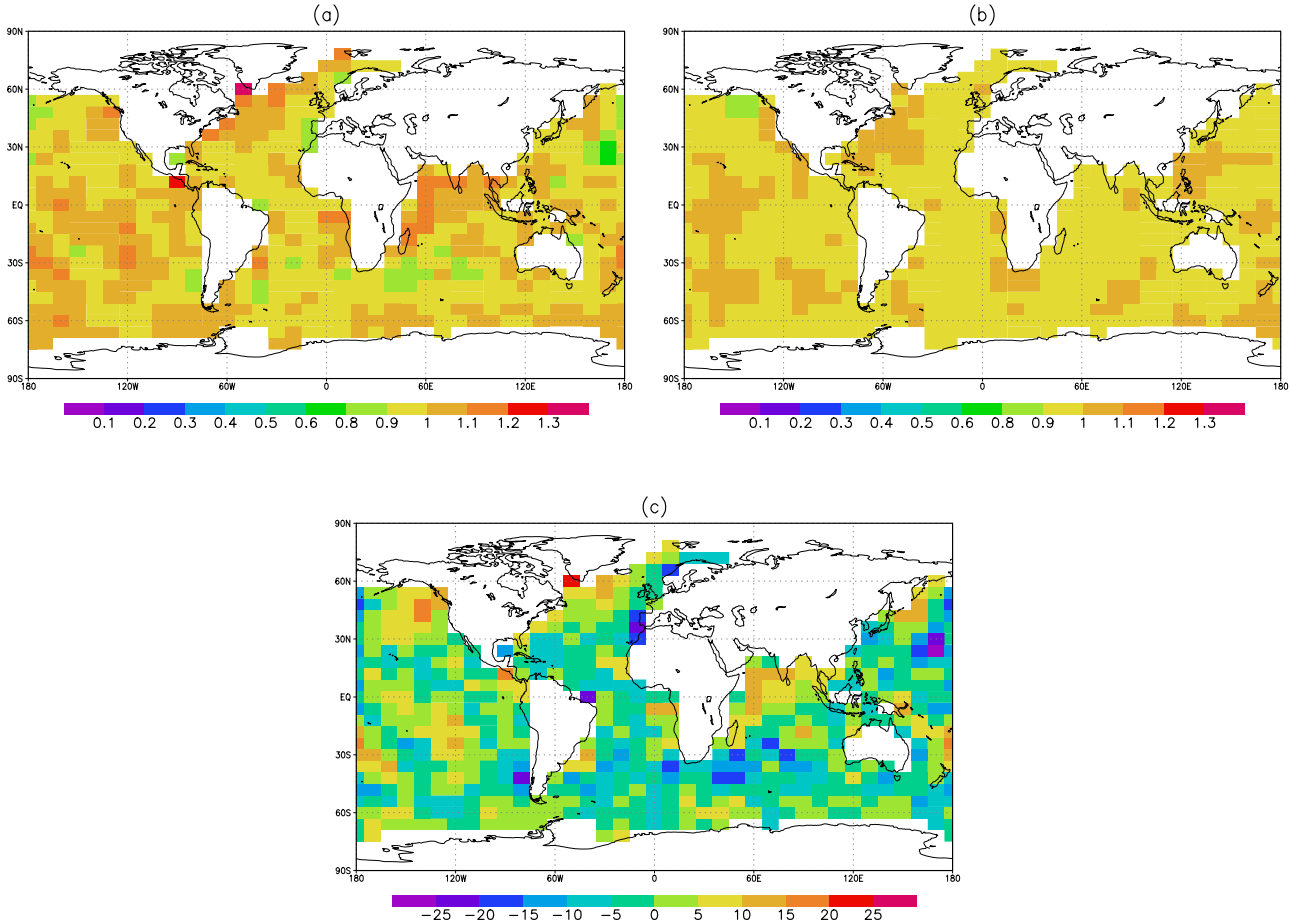

Figure 5. (a) Pseudotruth, (b) recovered, and (c) percentage difference relative to the pseudotruth for ocean bias (β_{Ocean}). Note that the prior equals to 1 everywhere (yellow color).

distribution tends to depend strongly on prior knowledge, thus, the solution we reach strongly depends on the prior distribution. Hence a reasonable prior distribution needs to be defined at the beginning of the assimilation cycle in order to achieve a realistic solution. In the case of a satellite experiment, we would have more freedom to choose a loose prior because of the tremendous amount of data.

3. Results and Discussion

[39] The assimilation cycle in this experiment lasts eight weeks, so each year of optimization requires 6.5 assimilation cycles. To test the performance of the model two measures of Root Mean Square (RMS) errors are calculated: One is based on the distance between the truth and the analysis (rms_analysis); the other one is based on the distance between the truth and the prior (rms_prior). In any cycle, lower rms_analysis relative to the rms_prior indicates that the solution is closer to the truth and hence shows an impact from the observations. The χ^2 diagnostic statistic evaluates the correctness of the innovation (observed minus forecast) covariance matrix that employs the predefined observation error covariance matrix $\boldsymbol{R}$, and the MLEF-computed forecast error covariance $\boldsymbol{P}_f$ [*Zupanski*, 2005]. Under the Gaussian assumption and for a linear observation operator $\boldsymbol{H}$, this statistic should be equal to one. But, in reality, it is approximately (not exactly) equal to one because of the statistically small samples (i.e., relatively few observations per cycle). Very large χ^2 values indicate too loose a fit to the observations and very small χ^2 values indicate too tight a fit to the observations. Figure 2a shows that the χ^2 statistic in each cycle is close to one, which indicates that the errors are consistent. This also indicates that the forecast error variance is estimated reasonably well on a global scale, however, as will be shown later, underestimation of the forecast error variance occurs in data void regions. RMS errors plotted in Figure 2b indicate an impact from the observations in each cycle except the last two cycles. One possible explanation for the slight increase of the rms_prior is that the filter has achieved convergence to the constant bias solution, given the available information in the observations. This is possible in our pseudodata experiment because the prescribed true biases are assumed to be constant throughout the year. With the current data coverage, we expect that up to 4 cycles (32 weeks) might be needed to recover a constant bias. One could prescribe a new bias after the first 32 weeks and perform the same procedure as in the first 32 weeks. We would expect that the MLEF would perform similarly, however, it might take shorter or longer to recover this new bias, since the atmospheric and biological conditions are different. These additional experiments would be more appropriate with real data, where the data would indicate what time scale would be most realistic for the bias. These issues are beyond the

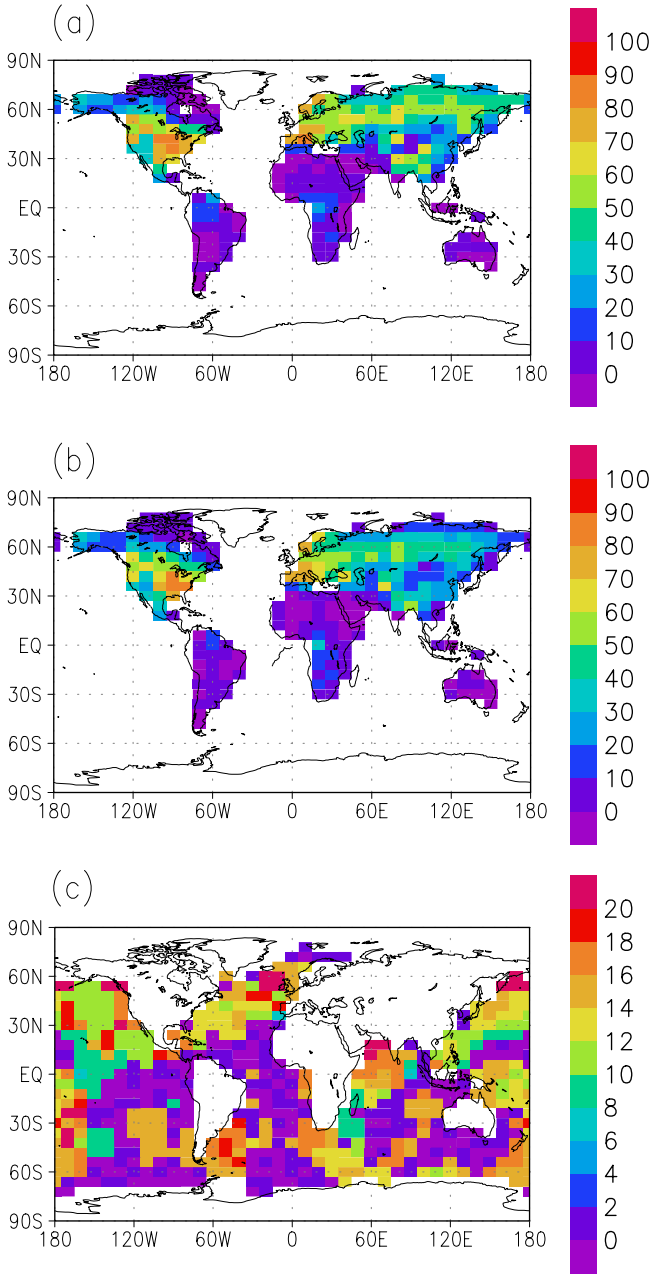

Figure 6. Percentage uncertainty reduction for recovered biases after six assimilation cycles for (a) GPP (β_{GPP}), (b) respiration (β_{RESP}), and (c) ocean (β_{Ocean}).

scope of this paper. Figure 2b shows relatively small improvement over the priors because the recovered biases show very little improvement over the ocean and the under represented land regions and hence the RMS errors are dominated by these regions.

[40] Results of the experiment with artificially generated biases of GPP, respiration, and air-sea flux are shown in Figures 3, 4, and 5. The Figures 3c–5c show the relative error, (truth − recovered) × 100/truth, which indicates the accuracy of the estimate. According to these plots, the coupled model generally recovered the pseudotruth within about 5–10% at most places after 6 assimilation cycles. Poorly recovered areas are shown by red or purple colors. The model performs well in the densely observed northern latitude regions (North America, Europe) compared to the sparsely observed southern latitude regions. Uncertainty reduction with respect to the prescribed background uncertainty for each bias component is given in Figure 6. The maximum reduction appears where observation sites are abundant. However, hotspots are not necessarily centered on the observation sites, because of the nature of the localization scheme we considered here, which takes into account the dynamics of CO_2 concentration fields. Uncertainty reduction for the bias in air-sea flux is minimal because the ocean fluxes are much weaker compared to the land fluxes. Thus the observation sites tend to see a much weaker signal from the ocean fluxes. Uncertainty reduction is much smaller in the sparsely observed land regions, as well.

[41] Figure 7 shows the estimated global annual mean fluxes for GPP, respiration, and NEE. The left column ((a), (c), and (e)) in the figure shows the true fluxes and the right column ((b), (d), and (f)) shows the recovered fluxes. Recovered fluxes show better agreement with the truth. Poorly recovered grid boxes in the bias estimators (see Figures 3c–4c) correspond to weaker or zero flux regions; these regions have minimal impact on observable CO_2, hence they minimally or do not contribute to the flux estimation.

[42] Usually data assimilation requires a dynamic model or a forecast model, which propagates the state vector and state covariance matrix from one cycle to another. In this experiment, we lack predictive equations to predict flux biases β from one time window to the next. Rather we have used the identity operator as the forecast model, as was done by *Peters et al.* [2005, 2007]. Hence the analysis state from a given cycle becomes the prior for the next cycle. However, the identity model cannot simulate error growth, so over many assimilation cycles, error covariance can be reduced to unreasonably small values. In such a case, the perturbations used to generate each ensemble member would become very small and the map of β's could converge prematurely to incorrect values. To avoid this problem, covariance inflation was applied to avoid/reduce covariance underestimation. At the end of each cycle, we inflated the covariance matrix uniformly by 30% for land points and by 5% for ocean points. Different magnitudes were assumed because the covariance minimization is much weaker for the ocean points compared to the land points. This inflated covariance will be the prior covariance for next cycle. *Peters et al.* [2005] introduced perturbations by generating random numbers from the prescribed covariance structure in each assimilation cycle, so that every cycle was treated as a "cold start". One can reach a reasonable solution with this technique, but it sacrifices the ability of the filter to "learn" about the covariance. Since the filter does not update the forecast error covariance, it does not have an ability to learn about it from the observations.

[43] Once optimized biases are available, one can easily estimate the CO_2 fluxes and their uncertainties. For example, if we wish to estimate the monthly average fluxes at a

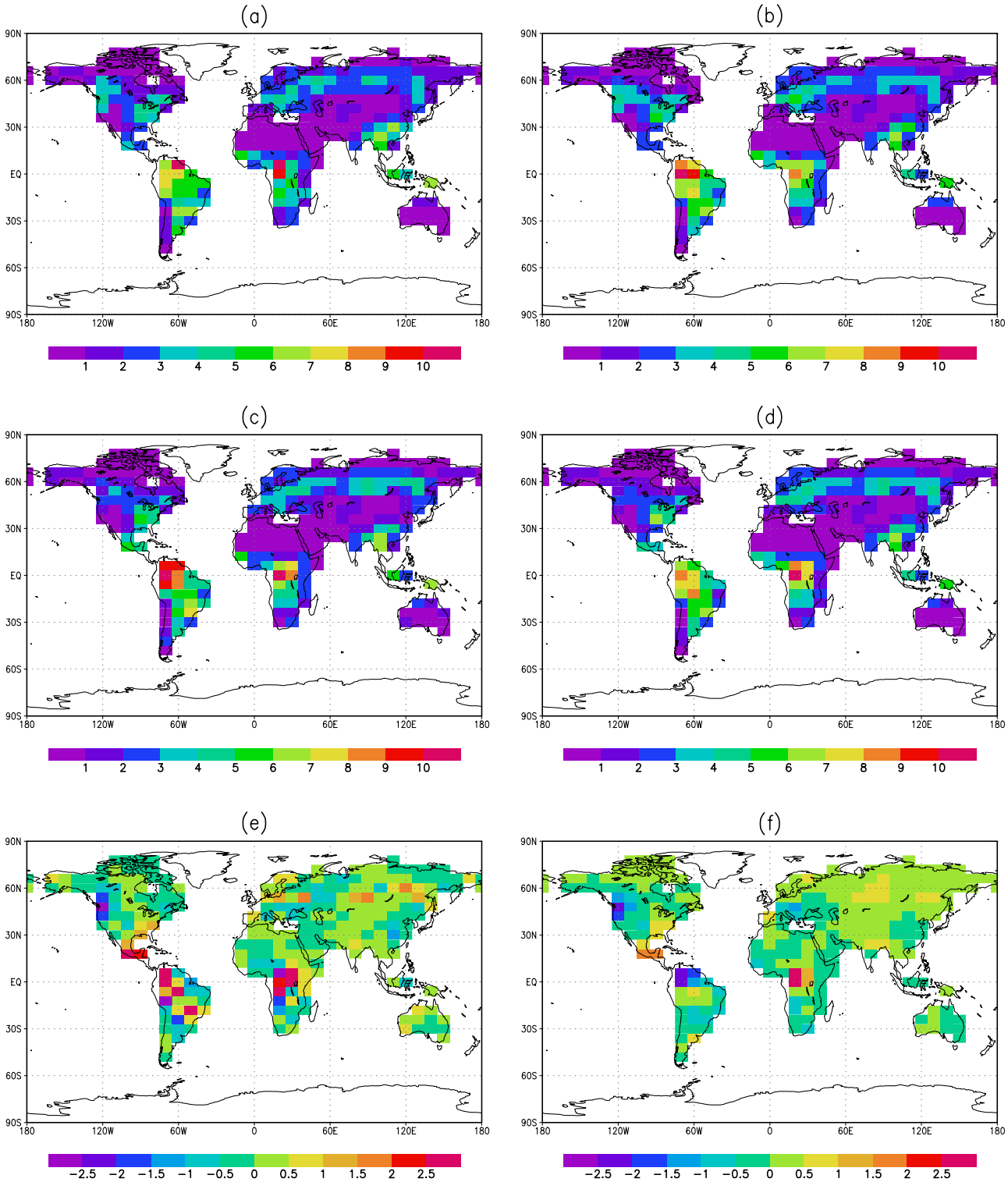

Figure 7. Mean annual fluxes for (a) GPP truth, (b) GPP recovered, (c) respiration truth, (d) respiration recovered, (e) NEE truth, and (f) NEE recovered. Units are in 10^{-8} kgC/m^2/s.

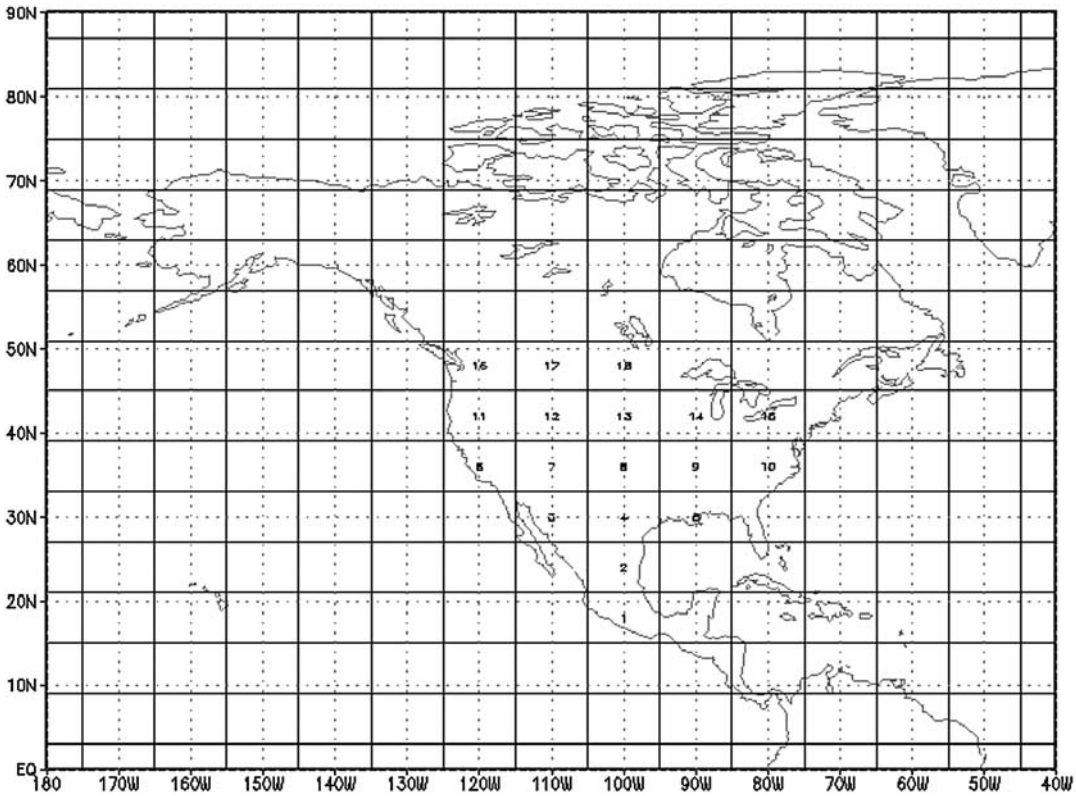

Figure 8. Grid boxes that are numbered from 1 to 18 indicate the North America Temperate (TC-2) region. Note that these numbers correspond to the covariance matrices defined in Figures 9 and 10.

given land location, we can simply calculate this quantity as

$$F = \beta_{RESP}\overline{RESP} - \beta_{GPP}\overline{GPP}, \tag{12}$$

where $\overline{(\ldots)}$ represents the monthly average. The corresponding uncertainty estimate is given by

$$\sigma_F^2 = \overline{(RESP)}^2\sigma_{\beta_{RESP}}^2 + \overline{(GPP)}^2\sigma_{\beta_{GPP}}^2 - 2 \times \overline{(RESP)} \times \overline{(GPP)} \times Cov(\beta_{RESP}, \beta_{GPP}), \tag{13}$$

where $\sigma_{\beta_{RESP}}^2 = Var(\beta_{RESP})$ and $\sigma_{\beta_{GPP}}^2 = Var(\beta_{GPP})$.

[44] The mean flux over a region can be estimated as

$$\bar{F}_{\mathrm{Region}} = \frac{1}{n}\sum_i F_i = \frac{1}{n}\sum_i \beta_{RESP,i}\overline{RESP}_i - \frac{1}{n}\sum_i \beta_{GPP,i}\overline{GPP}_i, \tag{14}$$

and the corresponding uncertainty in regional monthly NEE is given by

$$\begin{aligned}\sigma_{F_{\mathrm{Region}}}^2 = {} & \frac{1}{n^2}\sum_i\sum_j \overline{(RESP)}_i \times \overline{(RESP)}_j \times Cov(\beta_{RESP,i}, \beta_{RESP,j}) \\ & + \frac{1}{n^2}\sum_i\sum_j \overline{(GPP)}_i \times \overline{(GPP)}_j \times Cov(\beta_{GPP,i}, \beta_{GPP,j}) \\ & - \frac{2}{n^2}\sum_i\sum_j \overline{(RESP)}_i \times \overline{(GPP)}_j \times Cov(\beta_{RESP,i}, \beta_{GPP,j}),\end{aligned} \tag{15}$$

where i and j indicate the grid boxes within the region and n indicates the total number of grid boxes in the region.

[45] Note that even though the prescribed covariance matrix is defined by assuming that the GPP and respiration are uncorrelated, after smoothing cross-correlations will be added. Also, during the assimilation processes, the filter develops these cross-correlations from the data. Hence in the flux uncertainty estimation, these cross-correlations need to be taken into account.

[46] As an example, we extracted the covariance matrices of the biases for the North America Temperate region (TransCom3 region = 2), which consists of 18 grid boxes (see Figure 8). Figures 9a–9c indicate the prescribed error covariance matrices (after smoothing) of β_{GPP}, β_{RESP}, and their cross-covariance. Note that the grid points have been correlated according to the exponential covariance function and some noise is added to the cross-covariance, as described in section 2.6.1. Figures 10a–10c show the corresponding analysis error covariance matrices, after 6 assimilation cycles.

[47] All covariance matrices are diagonally dominant (Figure 10). Even though we introduced strong smoothing at the first cycle, these correlations were minimized through the learning process. Since this is a well-observed region (see Figure 2), learning from the observations was much stronger and most correlations were relaxed by the 6th assimilation cycle, making matrices (a) and (b) in Figure 10 diagonally dominant. In the process of assimilation, the variances also were greatly minimized with respect to the prescribed error covariance (compare the corresponding

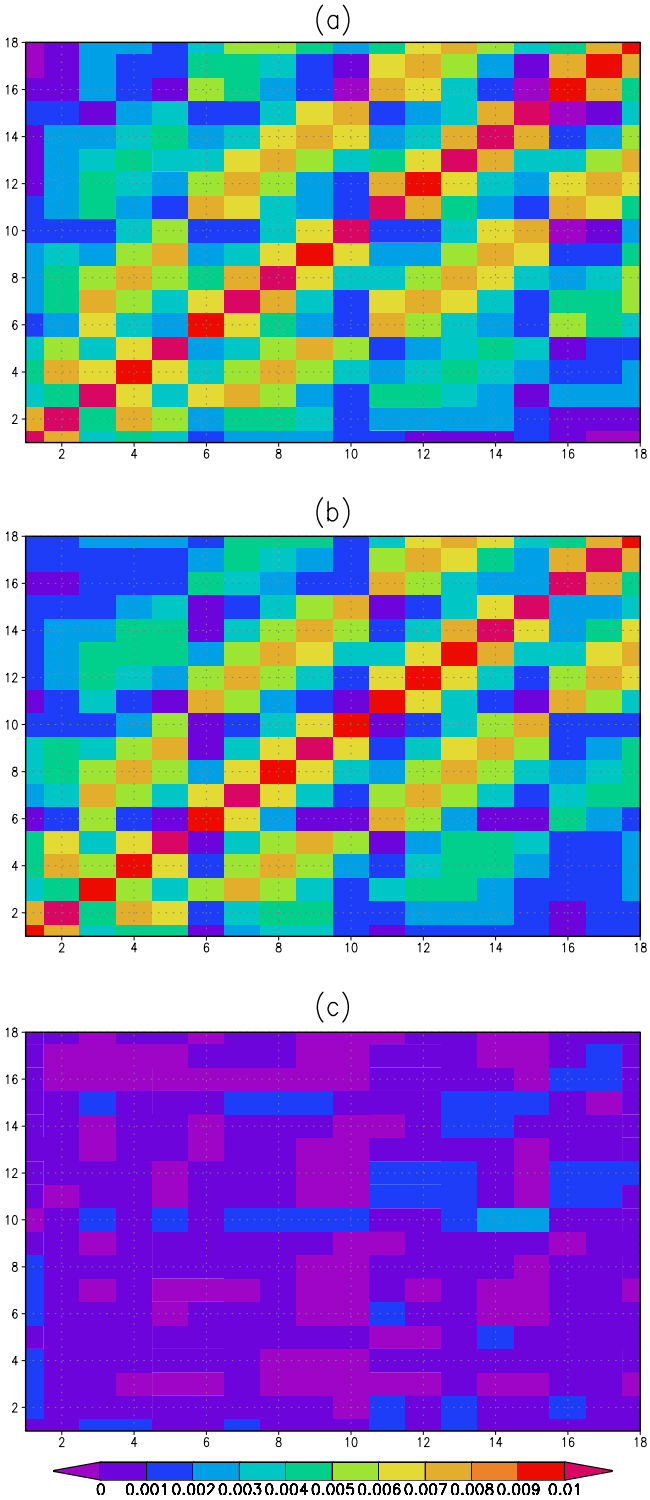

Figure 9. Prescribed error covariance matrices (after smoothing) of (a) β_{GPP}, (b) β_{RESP}, and (c) their cross-covariance for Temperate North America region.

diagonal values of (a) and (b) in Figure 9 with Figure 10). These plots also show that the minimization was much stronger in the areas where observations are abundant. The cross-covariance matrix has a more pronounced diagonal in Figure 10c, indicating that the two bias components are correlated at a given location, and almost negligible correlations further apart. The average correlation along the diagonal is about 0.39 and all correlations are positive. It is possible that these weak correlations are coming from the component fluxes and hence may affect the interpretation of

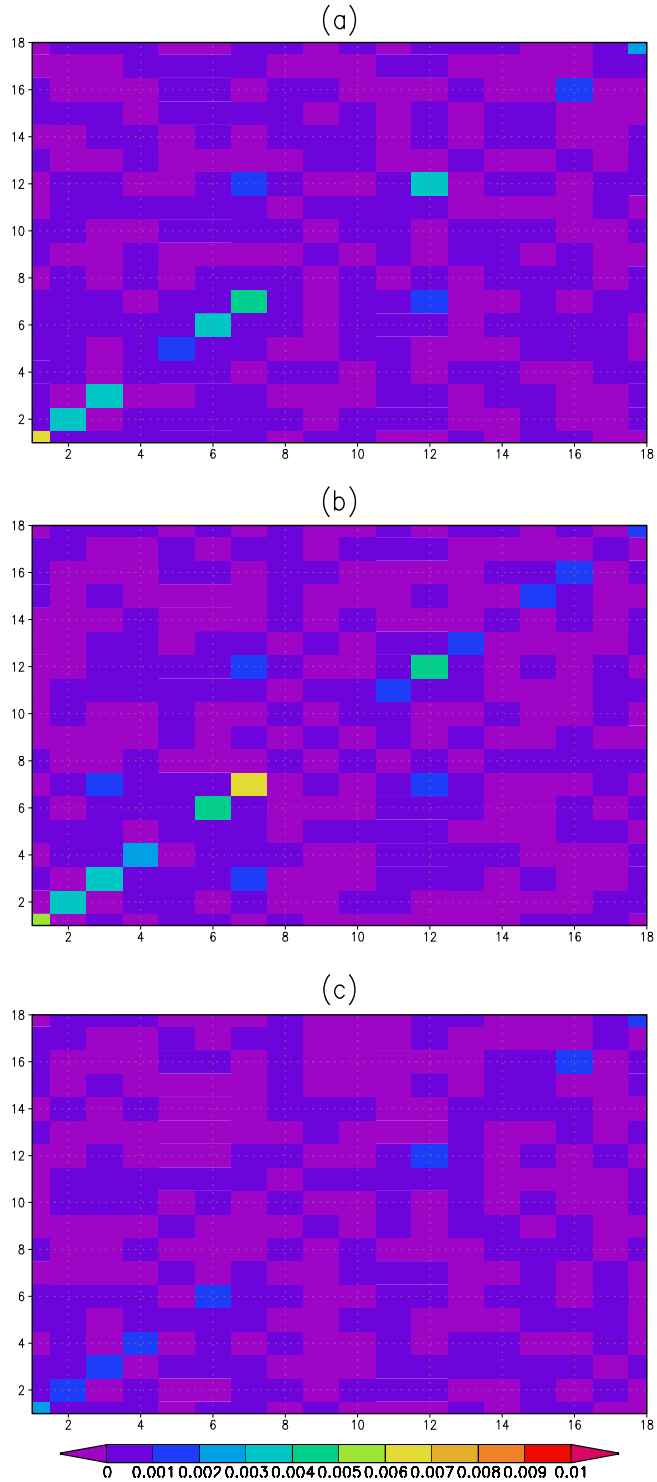

Figure 10. Analysis error covariance ($\boldsymbol{P_a}$) matrices of (a) β_{GPP}, (b) β_{RESP}, and (c) their cross-covariance for Temperate North America region after six assimilation cycles.

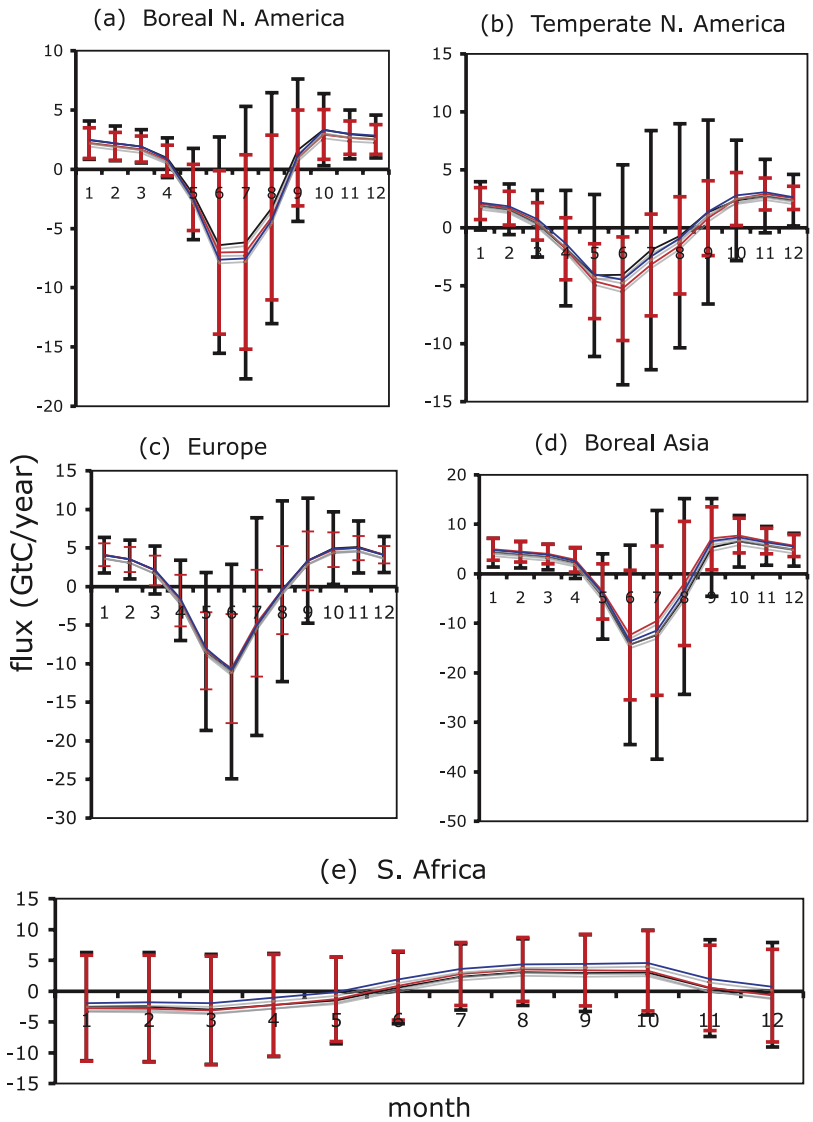

Figure 11. Mean NEE flux estimates with uncertainty for Boreal North America, Temperate North America, Europe, Boreal Asia, and South African regions (red—recovered, black—prior, and blue—truth).

the biases. We do not interpret these results quantitatively because this is a pseudodata experiment.

[48] Using equations (14) and (15), one can easily compute the mean flux along with the uncertainty for a given region, provided the monthly averages of GPP and respiration. We estimated the mean NEE along with their uncertainties for three well observed land regions and two sparsely observed land regions (see Figure 11). North American and European regions, which contain a large number of observation sites, were satisfactorily recovered with substantial uncertainty reductions (Figures 11a–11c). The Boreal Asia region was moderately recovered (see Figure 11d). In the South African region, very little change occurred in recovered NEE from the assumed prior because of under representation of the observation sites (see Figure 11e).

[49] Figure 12 shows the ocean flux estimates for 2 Transcom oceanic regions; North Pacific and Southern ocean. In the Southern Ocean region, the recovered flux showed very little improvement from the prior (see Figure 12b). Our method satisfactorily recovered the fluxes in the North Pacific oceanic region (see Figure 12a). In the oceanic regions, error minimization was minimal compared to the land regions. The percentage minimization with respect the prior for North Pacific and Southern Ocean were 13% and 8%, respectively. Previous studies also show that the uncertainty reduction over the oceanic regions was smaller compared to land regions [*Peters et al.*, 2005; *Gurney et al.*, 2004]. In comparison to the previous batch mode inversion results by *Gurney et al.* [2004], our method showed a great uncertainty reduction over well-observed land regions such as Boreal North America, Temperate North America, and Europe. This is due to the capability of assimilating large observation vectors with this method. However, the results over the oceanic regions did not show much improvement.

[50] The magnitudes of the error bars depend on the season. For example, in North America and Europe, mag-

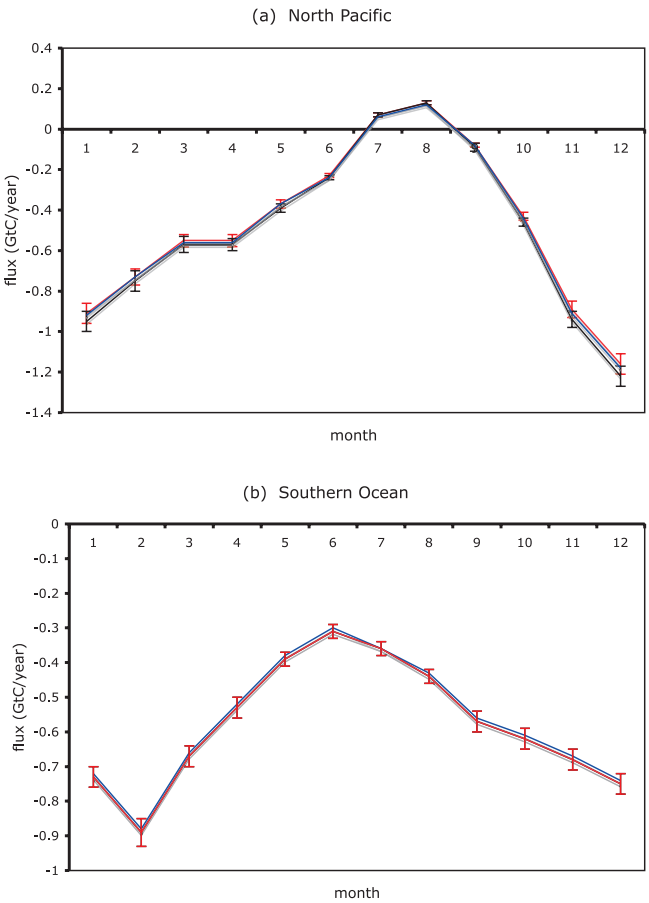

Figure 12. Mean flux estimates with uncertainty for North Pacific and Southern oceans (red—recovered, black—prior, and blue—truth).

nitudes of the error bars during the winter were much smaller than those of during the summer (see Figures 11a–11c). This happened because the NEE uncertainty is a function of the component (GPP and respiration) fluxes, which depend on the season (see equation (15)).

[51] Even though the mean fluxes were satisfactorily recovered, uncertainties were underestimated by the method, because of the lack of a better forecast model. The identity operator satisfactorily propagates the mean biases. However, covariance inflation, which serves as the covariance propagation model, is not sufficient for this particular case. We inflated both land and ocean biases uniformly (land by 30% and ocean by 5%). This forecast model may be suitable for a densely observed system as in an inversion using satellite observations. However, with the current observation network, densely observed regions were satisfactorily recovered but the forecast error variance was underestimated. On the other hand, sparsely observed regions were poorly recovered and the forecast error variance was overestimated. This problem can be avoided by eliminating the forecast error covariance update and perturbing the prior mean by prescribed covariance matrix in each cycle [*Peters et al.*, 2005, 2007], however, at the price of losing the capability to "learn" about the forecast error covariance, even in the densely observed areas. Unlike in the study of *Peters et al.* [2005] we retained the update of the forecast error covariance over the entire global domain, however, our filter was able to "learn" about the forecast error covariance only in the data rich regions. We are currently examining ways to improve the filter performance in the data sparse regions, by employing a variable inflation scheme based on the sigma ratio defined in equation (11).

4. Conclusions

[52] In this study, we have introduced an ensemble-based technique to estimate biases of photosynthesis, respiration, and ocean fluxes at each model grid point over a global scale. Some of the unique characteristics of this technique are the use of the MLEF approach, application of a computationally inexpensive processing of observations (compared to serial processing of observation in the previous studies), and the use of a "flow-dependent" distance function for covariance localization. The method performs satisfactorily with the existing observation network of flask measurements, continuous measurements and aircraft profiles and shows that it is capable of handling very large state vectors. However, it requires strong smoothing in the first data assimilation cycle and covariance localization in all cycles in order to get a reasonable solution.

[53] The current observing network is much more dense over North America and Europe than in other regions. Hence fluxes in these regions were recovered quite well with the assimilation scheme. Fluxes on the more sparsely observed southern continents were poorly recovered. More spatial coverage of observation stations is essential in grid-scale global inversions rather than accumulating more sites in already well represented areas.

[54] Ocean biases were poorly recovered with our assimilation scheme. In this case, the percentage uncertainty reduction was also minimal. Ocean fluxes are approximately ten times weaker than the land fluxes. Hence signals at the observation sites are dominated by the land fluxes. Little improvement to the ocean bias can be introduced by considering separate influence functions for land and ocean. In future work, this method will be implemented with real observations (and compared with other approaches).

[55] **Acknowledgments.** This research constitutes a part of the North American Carbon Program. We are thankful to David Baker for providing the low-resolution version of the PCTM model. We would also like to thank Wouter Peters for his valuable comments. This research was funded by NASA grants (NNX06AC75G and NNG05GF41G 02).

References

Baker, I. T., A. S. Denning, N. Hanan, L. Prihodko, P.-L. Vidale, K. Davis, and P. Bakwin (2003), Simulated and observed fluxes of sensible and latent heat and CO_2 at the WLEF-TV Tower using SiB2.5, *Global Change Biol.*, *9*, 1262–1277.

Baker, D. F., S. C. Doney, and D. S. Schimel (2006), Variational data assimilation for atmospheric CO_2, *Tellus, Ser. B*, *58*(5), 359–365.

Baker, I. T., et al. (2007), Global net ecosystem exchange (NEE) fluxes of CO_2, Oak Ridge Natl. Lab. Distrib. Active Arch. Cent., Oak Ridge, Tenn. (Available at http://www.daac.ornl.gov)

Bruhwiler, L. P., A. M. Michalak, W. Peters, D. F. Baker, and P. Tans (2005), An improved Kalman Smoother for atmospheric inversions, *Atmos. Chem. Phys.*, *5*, 1891–1923.

Bruhwiler, L. M. P., A. M. Michalak, and P. P. Tans (2007), Spatial and temporal resolution of carbon flux estimates for 1983–2002 EGU, *Biogeosci. Discuss.*, *4*, 4697–4756.

Burgers, G., P. J. V. Leeuwen, and G. Evensen (1998), Analysis scheme in the ensemble Kalman filter, *Mon. Weather Rev.*, *126*, 1719–1724.

Chevallier, F., M. Fisher, P. Peylin, S. Serrar, P. Bousquet, F.-M. Bréon, A. Chédin, and P. Ciais (2005), Inferring CO_2 sources and sinks from satellite observations: Methods and application to TOVS data, *J. Geophys. Res.*, *110*, D24309, doi:10.1029/2005JD006390.

Cohn, S. E. (1997), An introduction to estimation theory, *J. Meteorol. Soc. Jpn.*, *75*, 257–288.

Crisp, D., and C. Johnson (2005), The orbiting carbon observatory mission, *Acta Astronaut.*, *56*(1–2), 193–197.

Denman, K. L., et al. (2007), Coupling between changes in the climate system and biogeochemistry, in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group 1 to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by S. Solomon et al., Cambridge Univ. Press, New York.

Denning, A. S., J. G. Collatz, C. Zhang, D. A. Randall, J. A. Berry, P. J. Sellers, G. D. Colello, and D. A. Dazlich (1996), Simulations of terrestrial carbon metabolism and atmospheric CO_2 in a general circulation model. part 1: Surface carbon fluxes, *Tellus*, *48B*, 521–542.

Engelen, R. J., A. S. Denning, K. R. Gurney, and TransCom3 modelers (2002), On error estimation in atmospheric CO_2 inversions, *J. Geophys. Res.*, *107*(D22), 4635, doi:10.1029/2002JD002195.

Evensen, G. (1994), Sequential data assimilation with a nonlinear quasi-geostropic model using Monte Carlo methods to forecast error statistics, *J. Geophys. Res.*, *99*(C5), 10,143–10,162.

Fisher, M. (2003), Estimation of entropy reduction and degrees of freedom for signal for large variational analysis systems, *ECMWF Tech. Memo. No. 397*, 18 pp, European Center for Medium-Range Weather Forecasts (ECMWF), Reading, UK.

Fletcher, S. J., and M. Zupanski (2006), A data assimilation method for lognormally distributed observational errors, *Q. J. R. Meteorol. Soc.*, *132*, 2505–2520.

Gurney, K. R., et al. (2002), Towards robust regional estimates of CO_2 sources and sinks using atmospheric transport models, *Nature*, *415*, 626–630.

Gurney, K. R., et al. (2004), Transcom 3 inversion intercomparison: Model mean results for the estimation of seasonal carbon sources and sinks, *Global Biogeochem. Cycles*, *18*, GB1010, doi:10.1029/2003GB002111.

Houtekamer, P. L., and H. L. Mitchell (1998), Data assimilation using an ensemble Kalman filter technique, *Mon. Weather Rev.*, *126*, 796–811.

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

Kaminski, T., M. Heimann, and R. Giering (1999), A coarse grid three-dimensional global inverse model of the atmospheric transport—2. Inversion of the transport of CO_2 in the 1980s, *J. Geophys. Res.*, *104*(D15), 18,555–18,581.

Kaminski, T., P. J. Rayner, M. Heimann, and I. G. Enting (2001), On aggregation errors in atmospheric transport inversions, *J. Geophys. Res.*, *106*(D5), 4703–4715.

Kawa, S. R., D. J. Erickson III, S. Pawson, and Z. Zhu (2004), Global CO_2 transport simulations using meteorological data from the NASA data assimilation system, *J. Geophys. Res.*, *109*, D18312, doi:10.1029/2004JD004554.

Lin, S. J., and R. B. Rood (1996), Multidimensional flux-form semi-Lagrangian transport schemes, *Mon. Weather Rev.*, *124*, 2046–2070.

Michalak, A. M., L. Bruhwiler, and P. P. Tans (2004), A geostatistical approach to surface flux estimation of atmospheric trace gases, *J. Geophys. Res.*, *109*, D14109, doi:10.1029/2003JD004422.

Montzka, S. A., P. Calvert, B. D. Hall, J. W. Elkins, T. J. Conway, P. P. Tans, and C. Sweeney (2007), On the global distribution, seasonality, and budget of atmospheric carbonyl sulfide (COS) and some similarities to CO_2, *J. Geophys. Res.*, *112*, D09302, doi:10.1029/2006JD007665.

Peters, W., J. B. Miller, J. Whitaker, A. S. Denning, A. Hirsch, M. C. Krol, D. Zupanski, L. Bruhwiler, and P. P. Tans (2005), An ensemble data assimilation system to estimate CO_2 surface fluxes from atmospheric trace gas observations, *J. Geophys. Res.*, *110*, D24304, doi:10.1029/2005JD006157.

Peters, W., et al. (2007), An atmospheric perspective on North American carbon dioxide exchange: Carbon tracker, *Proc. Natl. Acad. Sci. U.S.A.*, *104*(48), 18,925–18,930.

Purser, R. J., and H.-L. Huang (1993), Estimating effective data density in a satellite retrieval or an objective analysis, *J. Appl. Meteorol.*, *32*, 1092–1107.

Rabier, F., N. Fourrie, C. Djalil, and P. Prunet (2002), Channel selection methods for Infrared Atmospheric Sounding Interferometer radiances, *Q. J. R. Meteorol. Soc.*, *128*, 1011–1027.

Rödenbeck, C., S. Houweling, M. Gloor, and M. Heimann (2003), CO_2 flux history 1982–2001 inferred from atmospheric data using a global inversion of atmospheric transport, *Atmos. Chem. Phys.*, *3*, 1919–1964.

Rodgers, C. D. (2000), *Inverse Methods for Atmospheric Sounding: Theory and Practice*, 238 pp., World Scientific Publication Co. Ltd., Singapore.

Schaefer, K., A. S. Denning, N. Suits, J. Kaduc, I. Baker, S. Los, and L. Prihodko (2002), The effect of climate on inter-annual variability of CO_2 fluxes, *Global Biogeochem. Cycles*, *16*(4), 1101, doi:10.1029/2004GB001928.

Takahashi, T., et al. (2002), Global sea-air CO_2 flux based on climatological surface ocean pCO_2 and seasonal biological and temperature effects, *Deep-Sea Res., Part II*, *49*(9–10), 1601–1622.

Tarantola, A. (1987), *The Least-Squares (12-norm) Criterion in Inverse Problem Theory: Methods for Data Fitting and Parameter Estimation*, pp. 187–287, Elsevier Sci., New York

Whitaker, J. S., and T. M. Hamill (2002), Ensemble data assimilation without perturbed observations, *Mon. Weather Rev.*, *130*, 1913–1924.

Zupanski, M. (2005), Maximum likelihood ensemble filter: Theoretical aspects, *Mon. Weather Rev.*, *133*, 1710–1726.

Zupanski, D., and M. Zupanski (2006), Model error estimation employing an ensemble data assimilation approach, *Mon. Weather Rev.*, *134*, 1337–1354.

Zupanski, D., A. S. Denning, M. Uliasz, M. Zupanski, A. E. Schuh, P. J. Rayner, W. Peters, and K. D. Corbin (2007a), Carbon flux bias estimation employing maximum likelihood ensemble filter (MLEF), *J. Geophys. Res.*, *112*, D17107, doi:10.1029/2006JD008371.

Zupanski, D., A. Y. Hou, S. Q. Zhang, M. Zupanski, C. D. Kummerow, and S. H. Cheung (2007b), Information theory and ensemble data assimilation, *Q. J. R. Meteorol. Soc.*, *133*, 1533–1545.

A. S. Denning and R. S. Lokupitiya, Department of Atmospheric Science, Colorado State University, Fort Collins, CO 80523-1371, USA. (ravi@atmos.colostate.edu)

K. R. Gurney, Department of Earth and Atmospheric Science, 550 Stadium Mall Dr, Purdue University, West Lafayette, IN 47907, USA.

S. R. Kawa, NASA Goddard Space Flight Center, 8800 Greenbelt Road, Greenbelt, MD 20771, USA.

D. Zupanski and M. Zupanski, Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, CO 80523-1375, USA.