"Variable localization" in an Ensemble Kalman Filter: application to the carbon cycle data assimilation

[1]Ji-Sun Kang (jskang@atmos.umd.edu),

[1]Eugenia Kalnay(ekalnay@atmos.umd.edu),

[2]Junjie Liu(jjliu@atmos.berkeley,edu),

[2]Inez Fung (ifung@berkeley.edu),

[1]Takemasa Miyoshi (miyoshi@atmos.umd.edu), and

[1]Kayo Ide (ide@atmos.umd.edu)

[1]Department of Atmospheric and Oceanic Science, University of Maryland, College Park, Maryland, 20742

[2]Dpartment of Earth and Planetary Science, University of California, Berkeley, 94720

Abstract

In Ensemble Kalman Filter (EnKF), space localization is used to reduce the impact of long-distance sampling errors in the ensemble estimation of the forecast error covariance. When two variables are not physically correlated, their error covariance is still estimated by the ensemble, and therefore it is dominated by sampling errors. We introduce a "variable localization" method, zeroing out such covariances between unrelated variables to the problem of assimilating carbon dioxide concentrations into a dynamical model using the Local Ensemble Transform Kalman Filter (LETKF) in an Observing System Simulation Experiments (OSSE) framework. A system where meteorological and carbon variables are simultaneously assimilated is used to estimate surface carbon fluxes that are not directly observed. A range of covariance structures are explored for the LETKF, with emphasis on configurations allowing non-zero error covariance between carbon variables and the wind field, which affects transport of atmospheric CO_2, but not between CO_2 and the other meteorological variables. Such "variable localization" scheme zeroes out the background error covariance among prognostic variables that are not physically related, thus reducing sampling errors. Results from the identical twin experiments show that the performance in the estimation of surface carbon fluxes obtained

using "variable localization" is much better than that using a standard full covariance approach. The relative improvement increases when the surface fluxes change with time and model error becomes significant.

1. Introduction

An Observing System Simulation Experiments (OSSEs) system for carbon cycle data assimilation has been created in parallel to a similar system that uses real meteorological and CO2 observations, and a state of the art model (Kang, 2009; Liu et al., CO_2 transport uncertainties from the uncertainties in meteorological fields, submitted to *Geophys. Res. Lett*, 2011). The ultimate goal of these parallel projects is to estimate not only atmospheric CO_2 but also surface carbon fluxes. This is a challenging problem plagued with obstacles whose origin frequently cannot be even identified using real data and without knowing the “truth”. In the course of performing OSSEs, we have found several algorithms that can substantially improve the results. The focus of this paper is on one of these algorithms, “variable localization” that reduces sampling errors and can be also applied to other problems in data assimilation.

The Local Ensemble Transform Kalman Filter [LETKF, *Hunt et al*., 2007], like other Ensemble Kalman Filter (EnKF) methods [*Evensen*, 1994; *Houtekamer and Mitchell*, 2001; *Anderson*, 2001; *Bishop et al*., 2001; *Whitaker and Hamill*, 2002; *Ott et al*., 2004, *Zupanski*, 2005, and others], produces an analysis using a multivariate background error covariance matrix that contains an estimation of the error correlation between the dynamic variables.

When the variables are physically related to each other, the multivariate background error estimation helps the analysis to efficiently correct the forecast errors. Indeed, *Liu et al.* [2009] have shown that a multivariate assimilation of AIRS (Atmospheric Infrared Sounder) humidity retrievals has lower wind analysis errors than the standard univariate assimilation used in operational numerical weather prediction (NWP) system that does not account for the error covariance between humidity and winds. However, standard multivariate EnKF also allows for error covariances among model variables even if some of those variables are not physically related to each other. In this case, the estimate of the error covariance will be solely due to sampling errors.

The direct solution to reduce sampling errors would be to increase the number of ensemble members, but this is not a practical solution because of computational and storage requirements. It is common practice in EnKF with a limited ensemble size to introduce "space localization" into the background error covariance [*Houtekamer and Mitchell*, 2001; *Hamill et al.*, 2001]. The background ensemble perturbations have error covariances that are good estimates of real covariances over relatively short distances of up to about 500-1000 km in the global NWP applications. At longer distances, the background errors are still apparently correlated, but these correlations become dominated by sampling errors, and can seriously harm the analysis. In the widely adopted technique of "space localization" to solve the problem of long-distance spurious correlations, the background error covariance terms are multiplied by an approximation of a Gaussian function that decreases with the distance between the two grid points whose error covariance is being computed and becomes negligible at distances greater than about 1000 km [*Gaspari and Cohn*, 1999].

In our carbon cycle data assimilation OSSEs, we apply a similar concept whenever two variables are not *physically* related and therefore estimates of their error covariances are spurious. In that case, we avoid spurious correlations by zeroing out these covariances due to sampling errors. For example, atmospheric CO_2 concentration is determined by the wind transport as well as by CO_2 surface fluxes. However, the evolution of the carbon variables is likely to have a much less significant dependence on some other variables such as the specific humidity or surface pressure. If this is the case, we zero out the error covariance between the atmospheric CO_2 and both of specific humidity and surface pressure in the analysis. This new methodology is denoted as "variable localization" by analogy with space localization, because the background error covariances between variables that are not physically linked in a significant way are zeroed out.

There are several previous studies that estimate surface carbon fluxes via data assimilation. *Baker et al*. [2006, 2008] have applied a four-dimensional variational (4D-Var) method to their OSSEs while *Peters et al*. [2005, 2007] and *Feng et al*. [2008] have used EnKFs. *Peters et al*. [2005, 2007] have assimilated observations from the ground-based stations whereas *Feng et al*. [2008] have used simulated observations of satellite data, Orbiting Carbon Observatory [OCO; *Crisp et al*., 2004]. Furthermore, as a part of Global and regional Earth-system (Atmosphere) Monitoring project [GEMS, *Hollingsworth et al*., 2008], a system with a two-step approach has been built for a carbon cycle data assimilation system: the first is to assimilate satellite and in situ data to monitor the atmospheric CO_2 within a 4D-Var [*Engelen et al*., 2009] and the second is a variational flux inversion system [*Chevallier et al*., 2009a and b]. These studies have shown the meaningful results in estimating surface CO_2 fluxes from the atmospheric CO_2 concentration observations of in-situ data as well as the satellite data. On the other hand, *Zupanski et al*. [2007] have applied Maximum Likelihood Ensemble Filter [*Zupanski*, 2005] to bias estimation of surface CO_2 fluxes over a local area with several tower observations.

In these studies, surface CO_2 fluxes are estimated by assimilating the observations of atmospheric CO_2 concentration, but not by any direct observations of carbon fluxes or other meteorological variables. In order to link the surface carbon fluxes with the atmospheric CO_2 concentrations, these studies have used transport models that play an important role in transferring information from the atmospheric CO_2 observations to corresponding changes in the surface flux of carbon. Also, these studies need *a priori* information about the carbon variables as an initial guess that is pre-calculated using independent observations or model simulations because the problem of determining surface carbon fluxes is otherwise ill-posed [*Enting*, 2002]. The surface CO_2 fluxes are determined by minimizing the squared normalized difference between the simulated CO_2 concentration and the observed CO_2, and *a priori* error for the atmospheric CO_2 concentrations and fluxes based on their error covariances.

So far, data assimilation studies of carbon fluxes have not yet accounted for transport errors in the atmospheric CO_2 forecast that can be caused by both the imperfections of the transport model and the uncertainty of wind analysis which drives the transport model, even though many studies [*Gurney et al*., 2004; *Baker et al*., 2008; *Stephens et al*., 2007; *Miyazaki,* 2009] have found that the accuracy of atmospheric CO_2 forecasts depends on those transport errors. The Bayesian synthesis approaches, usually referred to as inversion modeling, [*Bousquet et al*., 2000; *Gurney et al*., 2004; *Rödenbeck et al*., 2003] have been used to estimate surface carbon fluxes prior to the advent of the data assimilation studies discussed above, and also have the same limitation stemming from unresolved transport errors. Indeed, some studies [*Gurney et al.,* 2004; *Baker et al.,* 2008; *Stephens et al.,* 2007] have pointed out that the transport errors can cause biases in both the atmospheric CO_2 analysis and the surface CO_2 flux estimation. Notably, *Miyazaki* [2009] shows a significant contrast in the results of atmospheric CO_2 forecast obtained using wind fields of different accuracies. This result strongly emphasizes the importance of wind uncertainty in carbon cycle data assimilation.

As a complement of our real-data experiments (Liu et al., 2011), we present here a similar OSSE carbon cycle data assimilation system that simultaneously assimilates the observations of meteorological variables (wind, temperature, humidity, and surface pressure) and atmospheric CO_2. The system analyzes not only these meteorological variables and atmospheric CO_2, but also the surface CO_2 fluxes. Since our method generates the analysis of meteorological variables and carbon variables simultaneously, we do not need to run a transport model for carbon variables in addition to running a forecast model for meteorological variables. Besides, results of our method do not depend on *a priori information* on the initial condition of carbon variables which, like the meteorological variables, spin-up and converge even if the ensembles are started from smooth random fields [*Zupanski et al*., 2006]. Since our challenging ultimate goal is to estimate surface CO_2 fluxes from a simultaneous analysis of meteorological variables and carbon variables, we have tested new techniques to improve the ability of EnKF to reach this goal. Here we introduce one of these techniques, "variable localization" that can be usefully applied in any EnKF system.

Previous data assimilation estimates of surface carbon fluxes [*Baker et al.,* 2006, 2008; *Peters et al*., 2007; *Feng et al*., 2008; *Chevallier at al*., 2009a] can be considered as belonging to the case of "carbon-univariate" analyses where the atmospheric CO_2 concentrations and the surface CO_2 fluxes are updated by themselves, without including error correlations between these carbon variables and meteorological variables. In this study, various types of analyses including error covariances are introduced and compared, ranging from a standard "fully multivariate" analysis to a "carbon-univariate" analysis within the LETKF framework. Because this is the first test of a new methodology, this work is limited to OSSEs (in a "twin experiment" approach) in which the observations are sampled from a "nature run" assumed to be the true evolution of the system and assimilated using the LETKF. For simplicity, we also assume that the model is perfect except in the last experiments where the surface fluxes of carbon are varied in the nature run but not in the forecast model, and we focus on the impact of various "variable localization" techniques.

The paper is organized as follows. Section 2 provides a description of the model used for this study and the various "variable localization" schemes tested within the LETKF data assimilation framework. Section 3 describes the experimental design. Results are shown in section 4, and we summarize and discuss our findings in section 5.

2. Methodology

2.1. Model: SPEEDY-C

The SPEEDY model [*Molteni*, 2003] is an atmospheric global, primitive equation general circulation model (AGCM) with simplified physical parameterization schemes that is computationally efficient, but it maintains the basic characteristics of a state-of-the-art AGCM with complex physics. The version used for this study has triangular truncation T30 with 7 vertical sigma levels, and has five dynamical variables including zonal (U) and meridional wind (V) components, temperature (T), specific humidity (q), and surface pressure (Ps).

To incorporate atmospheric carbon dioxide concentration (C) and surface flux of carbon dioxide (CF), the SPEEDY model is extended to the "SPEEDY-C" model, which contains these carbon-related variables.

$$\frac{\partial(C)}{\partial t} + \Im(C) = CF \qquad (1)$$

Equation (1) shows how the tendency of atmospheric CO_2 is calculated in SPEEDY-C, where $\Im(C)$ represents the atmospheric 3-dimensional transport and mixing. In this study, chemical processes affecting atmospheric carbon dioxide are ignored since CO_2 in the atmosphere is essentially inert. Moreover, there is no feedback between the integrated CO_2 and the radiative properties of the SPEEDY-C model. Surface flux of carbon (CF) on the right-hand side of Equation (1) provides sources and sinks of CO_2. Carbon flux on the surface (CF) is converted into the atmospheric CO_2 concentrations added to the lowest layer of the model. In reality, several types of forcings make up this flux: fossil fuel emission, land surface fluxes due to vegetation and land use change, and ocean fluxes. In this OSSE study, we test the ability of our data assimilation systems with a "variable localization" to estimate surface CO_2 fluxes, so for simplicity we assume that the CF is due only to constant fossil fuel emissions for most of experiments. However, in the last experiment (shown in Figure 10), we do allow for variable fluxes associated with vegetation and ocean in the nature run but not in the forecast model run. Thus, the SPEEDY-C has six prognostic variables (U, V, T, q, Ps, C), along with either a constant or a variable forcing (CF) which is not changed by the model.

2.2. The LETKF

The LETKF is an ensemble Kalman filter method where the background error covariance $\mathbf{P}^b$ among the variables can be estimated as

$$\mathbf{P}^b = \frac{1}{K-1}\mathbf{X}^b\mathbf{X}^{b^T} \tag{2}$$

where $\mathbf{X}^b$ is the matrix whose columns contain a departure of each ensemble forecast ($\mathbf{x}^{b(i)}$) from the ensemble mean ($\overline{\mathbf{x}}^b$): the *i*-th column of $\mathbf{X}^b$ is $\mathbf{x}^{b(i)} - \overline{\mathbf{x}}^b$, $\{i = 1,2,\ldots,K\}$, K is the number of ensemble members and $\mathbf{x}$ is a state vector of dynamic variables at the model grids. The evolution of $\mathbf{P}^b$, which contains the background error covariance among the dynamic variables, is accounted for in every analysis step so that temporally and spatially varying uncertainties in the background are considered when analyzing variables.

The first step of the analysis is to compute $\mathbf{X}^b$. Then, the observation operator h is applied to the ensemble forecast $\mathbf{x}^b$ to transform the background from the model grid space to the observation space, $\mathbf{y}^{b(i)} = h(\mathbf{x}^{b(i)})$. Let $\mathbf{Y}^b = \mathbf{y}^{b(i)} - \overline{\mathbf{y}}^b$, $\{i = 1,2,\ldots,K\}$ be the background perturbations in the observation space. Then, the estimation of the background is ready to be compared with observations in the same space.

In order to produce the analysis at every grid point, the LETKF assimilates only observations within a certain distance from each grid point so that the following analysis computations are performed locally. The analysis mean, $\overline{\mathbf{x}}^a_{(l)}$, is given by

$$\overline{\mathbf{x}}^a_{(l)} = \overline{\mathbf{x}}^b_{(l)} + \mathbf{X}^b_{(l)} \overline{\mathbf{w}}_{(l)}, \tag{3}$$

where $\overline{\mathbf{w}}_{(l)}$ is the mean weighting vector calculated by

$$\overline{\mathbf{w}}_{(l)} = \widetilde{\mathbf{P}}^a_{(l)} (\mathbf{Y}^b_{(l)})^T \mathbf{R}^{-1}_{(l)} (\mathbf{y}^o_{(l)} - \overline{\mathbf{y}}^b_{(l)}). \tag{4}$$

Here, $\widetilde{\mathbf{P}}^a_{(l)} = [(\mathbf{Y}^b_{(l)})^T \mathbf{R}^{-1}_{(l)} (\mathbf{Y}^b_{(l)}) + (K-1)\mathbf{I}/\rho]^{-1}$ is the analysis error covariance in the ensemble space, $\mathbf{R}$ is the observation error covariance matrix, $\mathbf{y}^o$ is the observation vector, and ρ is the inflation factor (see section 2.3.3 for details), and the subscript (l) means a quantity defined on a local region centered at the analysis grid point l. Within a local region, *space localization* is carried out by multiplying the inverse observation error covariance matrix $\mathbf{R}^{-1}_{(l)}$ by a factor that decays from one to zero as the distance of the observations from the analysis grid point increases [*Miyoshi*, 2005, *Hunt et al.*. 2007, *Greybush et al.*, 2010].

The analysis increment, $\overline{\mathbf{x}}^a_{(l)} - \overline{\mathbf{x}}^b_{(l)}$ (Eqns. 3 and 4), is given by the background perturbation matrix multiplied by the weight vector which is a function of the innovation, $\mathbf{y}^o_{(l)} - \overline{\mathbf{y}}^b_{(l)}$, and error statistics of both background and observation. Thus, the analysis reflects observational information more than background information if the background error is greater than the observation errors, and vice versa. In addition, the ensemble perturbations of the analysis are determined by

$$\mathbf{X}^a_{(l)} = \mathbf{X}^b_{(l)}[(K-1)\widetilde{\mathbf{P}}^a_{(l)}]^{\frac{1}{2}} \tag{5}$$

With (5) we obtain the estimation of analysis uncertainty in addition to the analysis mean. The global analysis ensemble $\mathbf{x}^{a(i)}$, $\{i = 1,2,\ldots,K\}$, is formed by gathering the values obtained for $\overline{\mathbf{x}}^a_{(l)}$ and $\mathbf{X}^a_{(l)}$ at all the analysis grid points. (see *Hunt et al*. [2007] for more details and discussion on LETKF.)

2.3. LETKF application to the SPEEDY-C: variable localization

2.3.1. Motivation

In order to estimate not only the model prognostic variables (U, V, T, q, Ps, C) but also the unknown surface fluxes field (CF), we use an augmented state vector $\mathbf{x}$ consisting of (U, V, T, q, Ps, C, CF) at all model grid points, where CF, like Ps, is defined at the model surface grid points. This augmentation enables the LETKF to directly estimate the parameter like any other (unobserved) variable through the background error covariance with the observed variables [*Baek et al*., 2006; *Annan et al*., 2005].

More sophisticated schemes can be designed by taking into account that dynamical interactions of the augmented variables are not homogeneous in the SPEEDY-C. As shown in Equation (1), atmospheric CO_2 (C) is advected by (U, V) and forced by surface carbon fluxes (CF) but has no direct interaction with (T, q, Ps). In contrast, none of the meteorological variables (U, V, T, q, Ps) is dynamically affected by C or CF while CF is not affected by any of the dynamical variables (U, V, T, q, Ps), at least within the SPEEDY model formulation. When sampling the standard fully multivariate background error covariances using a finite-size ensemble (Figure 1a), however, spurious correlations may arise between the variables. This motivates us to develop analysis schemes by grouping the variables based on the idea of the localization according to the *"dynamical distance between the variables"*. This "variable localization" attempts to manage the correlations between the model variable groups, like the conventional localization attempts to suppress the spurious correlation based on the "*physical*" distance.

Various analysis methods are possible according to the method used to group variables. For example, if one groups an analysis state vector of only carbon variables (C, CF) and the other of meteorological variables (U, V, T, q, Ps) separately (Figure 1e), the analysis of carbon is determined by only assimilating atmospheric CO_2 observations *univariately for* *carbon* (Equations 3-5). In this case, the surface CO_2 fluxes are updated by the multiplication of a background perturbation matrix of surface CO_2 fluxes and the weight vector (Equation 4) as calculated from the forecast and the observations of atmospheric CO_2 concentrations. If the analysis state vector is designed to include other meteorological variables in addition to (C, CF), then the analysis can reflect the background error covariance among all those variables in order to estimate surface CO_2 fluxes *multivariately* (e.g., Figure 1a). Such an approach implies that the analysis allows error information to flow from carbon to the meteorological variables in the state vector and vice versa.

2.3.2. Different covariance structures for analyses: variable localization

In this study, we introduce variable localization and test five analysis methods characterized by the $\mathbf{P}^b$ configurations based on the "dynamical" distance between variables (Figure 1). The first method is the standard *fully multivariate* analysis (hereafter referred as ***mult***) in which the errors of all dynamic variables are coupled in the background error covariance (Fig 1a). This scheme (used in present EnKF methods) allows errors in all variables to be potentially correlated with one another. As a result, the system gives more weight to the atmospheric CO_2 observations whenever any of the dynamic variables have a larger uncertainty in the background field. On the other hand, the uncertainty of the carbon variables can also change the weight vector $\overline{\mathbf{w}}_{(l)}$ (Equations 3 and 4), which is shared among all dynamic variables. Analysis uncertainty of all variables (U, V, T, q, Ps, C, CF) is determined by Equation (5).

The second method is based on the notion that in our model the surface fluxes are only physically related to the atmospheric CO_2 but not to other dynamical variables. That is, the white areas of background error covariance matrix $\mathbf{P}^b$ in Figure 1(b) contain sampling errors rather than any useful error correlations or covariances. Thus, we zero out those white areas, the covariances between CF and all variables except atmospheric CO_2, in order to eliminate sampling errors in these correlations (*localized-multivariate* analysis: ***L-mult***, Fig 1b). This scheme has two separate analyses, one for (U, V, T, q, Ps, C) (in grey) and the other for CF (in black). Analysis of surface carbon fluxes assimilates only atmospheric CO_2 observations but not the observations of meteorological variables for computing $\overline{\mathbf{w}}_{(l)}$ and the uncertainty of CF. In contrast, the analysis of the dynamic variables except CF assimilates all available observations of (U, V, T, q, Ps, C) for computing the other $\overline{\mathbf{w}}_{(l)}$ and their analysis uncertainty. In other words, a set of Equations (4)-(5) is computed separately having $\mathbf{X}_1$=(U, V, T, q, Ps, C) and $\mathbf{X}_2$=(C, CF) to get each weight vector $\overline{\mathbf{w}}_{(l)}$ and analysis error covariance $\widetilde{\mathbf{P}}^{a}_{(l)}$ for updating (U, V, T, q, Ps, C) and CF respectively. Eliminating spurious correlations with carbon fluxes is especially important since we do not start the analysis with any *a priori* knowledge of carbon variables. Because CF is not constrained by any direct observations, it is very possible for the CF to degrade the analysis of the other variables due to bad initial values at the initial stage of the ***mult*** analysis. Thus, poor initial conditions for carbon may degrade the analysis of other meteorological variables in the ***mult*** analysis whereas ***L-mult*** prevents initial carbon from poorly influencing the analysis of all other dynamic variables.

The third method is the *1-way multivariate* analysis (***1way***) based on the notion that wind uncertainties should be able to provide useful information to update carbon variables, whereas the sampling error in the carbon variable is assumed to be too large to provide a positive impact to the wind assimilation (Fig. 1c). In the ***1way*** scheme, the atmospheric CO_2 concentrations and surface CO_2 fluxes are updated using an error covariance that includes the wind fields, while the wind and other atmospheric variables such as temperature, specific humidity, and surface pressure are updated separately and are not affected by these two carbon variables (Fig 1c). This scheme was also found useful by *Liu et al*. [2009] when assimilating AIRS moisture retrievals.

The fourth method is also based on the ***1way*** system, but zeroing out the background error covariance between surface carbon fluxes and wind fields. We refer to this as the *localized-1way multivariate* analysis (***L-1way***, Fig 1d), as in the case of the ***L-mult*** scheme. It is based on the idea that winds transport atmospheric CO_2 but not surface carbon fluxes, and thus their errors should be uncorrelated. Here, the resulting analysis of meteorological variables should be exactly same as in ***1way*** (Fig. 1c). The comparison of the ***L-1way*** method with ***1way*** provides a measure of the direct impact of wind uncertainties on the estimation of surface CO_2 fluxes.

The last method considered is the *Carbon-univariate* analysis (***C-univ***). In this method, atmospheric CO_2 concentration and surface CO_2 fluxes are updated only by these two variables themselves, unaffected by other atmospheric variables (Fig 1e). The forecasts of atmospheric CO_2 are still driven by the ensemble of wind fields. Although the ensemble transport of CO_2 provides some information about wind uncertainties to the background state of atmospheric CO_2 in ***C-univ***, the transport error term is not explicitly used for the carbon analysis.

2.3.3. Inflation of the background covariance

In practice, the ensemble forecast tends to underestimate the uncertainty in its state estimate because of limited ensemble size, model errors and nonlinearities. To compensate for this underestimation, it is necessary to inflate the background covariance (or the analysis covariance) during each data assimilation cycle. For the inflation factor, multiplicative inflation has been applied in this work [*Anderson and Anderson*, 1999]. This is carried out by multiplying the background perturbation from the ensemble mean by a factor larger than one (ρ). It is common to tune this inflation parameter manually; however, such tuning is expensive, and becomes infeasible if the inflation factor is allowed to depend on space and time, and/or the variable. Since we have found that the carbon variables require quite different inflation factors compared to the inflation for the meteorological variables, the adaptive inflation estimation introduced by *Li et al*. [2009] has been used to estimate the inflation factors for the meteorological variables, on the one hand, and the atmospheric CO_2 concentration on the other. *Li et al*. [2009] estimated simultaneously the adaptive inflation and observation errors, using the equations derived by *Desroziers et al.* [2005]. Here we assume that the observation error statistics are correct, and we calculate the inflation adaptively for each vertical layer separately. Moreover, for the atmospheric CO_2 in the lowest layer, we calculate and apply two separate inflation factors over the land and the ocean areas. The methodology of *Li et al*. [2009] compares the analysis increment (analysis minus background) and the observation increment (background minus observation) with the expected values in observation space. Thus, that methodology is only available for variables having observations, which means we need to apply a different method for the inflation of surface carbon fluxes (CF). Our approach for CF is similar to the covariance relaxation method of *Zhang et al*. [2004], except that we let the analysis perturbations maintain the same spread as the background. More details about the adaptive inflation methods can be found in *Li et al*. [2009] and *Zhang et al*. [2004].

3. Experimental design: Observing System Simulation Experiments (OSSEs)

In the OSSEs, the SPEEDY-C model with a total constant fossil fuel emission of 6PgC/yr [*Andres et al*., 1996; Figure 2a] is used to create the “nature run” assumed to be the true state in this study (but we also perform an OSSE with varying surface fluxes, obtained with a model with interactive vegetation, see Figure 10). Simulated observations are then obtained from this “nature run” by adding random observational errors. Standard deviations of the simulated observation errors are listed in Table 1. For the atmospheric variables, the observations have the spatial distribution of the rawinsonde network, with about 9% coverage of grid points globally (Figure 3a), with more observations in the Northern Hemisphere mid-latitudes.

Atmospheric CO_2 concentration is assumed to be observed from three different measurements: one comes from 18 *in situ* data locations which have continuous records of CO_2 concentration near the surface (Figure 3b: crosses); another source is from 107 flask data sites which observe CO_2 concentrations near the surface every week (Figure 3b: closed circles); lastly, GOSAT column data [*Yokota et al*., 2004] are used (Figure 3b: gray lines), with orbital return periods of three days. For simplicity, in this simulation we did not account for the impact of cloud screening. We assume that the GOSAT data have the same averaging kernel as OCO [*Wang et al*., 2009], i.e., nearly constant from the surface to the top of atmosphere. For this column data, the column observation increments are localized to each vertical level by the normalized averaging kernel for each level as follows:

$$\mathbf{y}^b = h(\mathbf{x}^b) = \mathbf{A}^T(\mathbf{H}\mathbf{x}^b) = \sum_{i=1}^{k} a_i(\mathbf{H}\mathbf{x}_i^b) \tag{6}$$

where k is the number of vertical levels, $\mathbf{H}$ the spatial interpolation operator, $\mathbf{y}^b$ the model predicted CO_2 column mixing ratio, $\mathbf{A}$ the averaging kernel, and a_i the element of $\mathbf{A}$ at the *i-th* vertical level. We localize the *j-th* ensemble forecast column CO_2 to *i-th* vertical level by the *i-th* averaging kernel element a_i as $y^b_{j,i} = a_i \times y^b_j$ and the column CO_2 observations to the *i-th* vertical level by a_i as $y^o_i = a_i \times y^o$. Then, $y^b_{j,i}$ and y^o_i are compared during the analysis.

In the data assimilation system, the same model as the "nature run" is used for the ensemble forecasts of 20 members (K=20), so that there is no model error (except for the last experiment where the nature model has variable carbon fluxes not included in the forecast model). Our goal is to test the impact of "variable localization" schemes in estimating the spatial distribution of true CF shown in Figure 2a. Since CF is a forcing term in the SPEEDY-C not changed by the forecast, it is updated only by the analysis step of data assimilation, and the updated forcing from the analysis is then used for the next forecast.

The initial ensemble members are chosen by random sampling from a long term simulation of the SPEEDY-C and a SPEEDY-C coupled with a dynamic terrestrial carbon model VEGAS [*Zeng et al.*, 2005] (hereafter referred as SPEEDY-VEGAS; *Kang,* 2009) in order to generate fields of the initial ensemble background with no *a priori* information about the nature run: 20 states of (U, V, T, q, Ps) and C are chosen randomly in time over an one-year SPEEDY-C output and a three-year SPEEDY-VEGAS run respectively, and then they are added by small random perturbations. For CF, from 20 fields of CO_2 concentration of the SPEEDY-C run in the midlevel at arbitrary times, we subtract the one-day prior state of CO_2 concentration, and then convert the units of the field from ppmv/day to $kg/m^2/s$. Figures 2c and 2d show that the initial ensemble mean of the surface carbon fluxes and the first level atmospheric CO_2 are very different from the true states in terms of both spatial patterns and intensity. Since CO_2 concentration is well-mixed in the midlevel, Figure 2d has very small values. Starting from these initial conditions without any *a priori* information, we carried out the analyses of all dynamic variables for four months using an analysis cycle of six hours.

The experimental settings described above are used for testing all schemes introduced in this study to see the impact of "variable localization" techniques. In addition to these experiments carried out in a perfect model and constant flux configuration, we have also done another set of experiments testing the impact of variable carbon fluxes in the nature model. With the same configuration of the observations and the same initial conditions, we repeated the ***L-1way*** and ***C-univ*** experiments now including terrestrial and oceanic CO_2 fluxes which evolve in time. We replace in "nature run" the CO_2 forcing every six hours by the land surface CO_2 fluxes computed by VEGAS [*Zeng et al*. 2005] that includes the vegetation impact on the carbon cycle, and the monthly prescribed oceanic fluxes [*Takahashi et al.,* 2002] in addition to the fossil fuel emission used in the previous experiments. We have produced one-year analysis and show the results for the last two-month average in Section 4.

4. Results

Table 2 contains the global RMS errors for all variables from all the analysis schemes during the last week of the 4-month data assimilation, and Figures 4 and 5 show the time evolution of the global RMS errors in zonal wind and carbon variables. Other meteorological variables have a similar pattern of RMS errors in the time series plot, compared to the zonal wind. First, the standard fully multivariate data assimilation (***mult***) has the worst results for all the variables. This is because ***mult*** allows for error covariances among all variables in the analysis even though there is no physical relationship between (C, CF) on the one hand and (T, q, Ps) on the other in the "nature" model. Therefore, the estimations of the error covariances among these variables are only due to sampling errors. Moreover, a poor representation of the initial surface carbon flux can contaminate analyses of all variables in ***mult***. As a result, the ***mult*** analysis has larger errors and eventually undergoes filter divergence, i.e., the feedback from the sampling errors makes the analysis of meteorological variables so poor that the diagnosis of the model variables in the forecast fails after 40 days. In theory, this problem of the ***mult*** system could be resolved by using much larger ensemble size so that sampling errors are reduced, but in practice this approach is not computationally feasible.

By eliminating the unphysical relationship between the carbon flux CF and (U, V, T, q, Ps), ***L-mult*** prevents a poor initial representation of CF from degrading the analysis of the other variables. Also, the analysis of carbon variables benefits from better states of other variables (without a contamination of the surface carbon flux). As a result, the ***L-mult*** analysis is improved significantly for all dynamic variables and filter divergence is avoided. Still, there is unnecessary feedback between C and (T, q, Ps), which is negligible in nature. Thus, ***L-mult*** is not the optimal method and can be improved further by additional variable localization (Figure 4, 5 and Table 2).

In ***1way***, we zero out the background error covariance between (C, CF) and (T, q, Ps). Furthermore, ***1way*** does not allow any changes in the meteorological variables due to the CO_2 variables. Compared with ***mult***, this does not allow any feedback between carbon variables (C, CF) and (T, q, Ps), but, in contrast with ***L-mult***, it does include the covariance between CF and wind fields. Carbon variables from ***1way*** analysis are improved significantly while the analyses of meteorological variables are, as expected, comparable with the results of ***L-mult***. Mean RMS errors (Table 2) show that the differences between ***1way*** and ***L-mult*** are only on the order of 1% for (U, V, T, q, Ps) whereas ***1way*** improves the estimates of (C, CF) by 30-35%.

Figure 6 compares maps of the analysis errors in the zonal and meridional wind fields obtained with ***1way*** and with ***L-mult***. Due to the distribution of the rawinsonde network sites (Figure 3a), errors are large over the oceans and polar regions in both ***L-mult*** and ***1way***. The analysis of wind in ***1way*** has similar error patterns but smaller error amplitudes than in ***L-mult***. By contrast, ***1way*** results in a major improvement in the atmospheric CO_2 analysis as shown in Figure 7. Since ***L-mult*** considers the background of (T, q, Ps) in addition to (U, V, C, CF) for analyzing the atmospheric CO_2, the background uncertainties of (T, q, Ps) can influence the weight between the background and the observations of atmospheric CO_2. Although a large uncertainty of temperature can be related to the wind uncertainty so that the carbon dioxide concentration could be affected by those wind errors, this is not a first-order effect and does not need to be considered during the analysis. This is what the result in Figure 7 shows: ***L-mult*** has larger errors overall and the spatial pattern is not as smooth as the nature run or the results from ***1way***. Because the ***L-mult*** analysis reflects more strongly the observations of atmospheric carbon whenever there are large background uncertainties of (T, q, Ps) in addition to (U, V, C, CF), atmospheric CO_2 observations are over-weighted for the case of ***L-mult*** producing an analysis with additional noise (Figure 7a) compared with the case of ***1way*** (Figure 7c). Over the ocean, where there are few observations of meteorological variables, their estimated error, given by the background spread, is large. Thus, the ***L-mult*** tends to give more weights to the atmospheric CO_2 observations than it should because it considers the joint background uncertainties of (U, V, T, q, C, Ps) altogether.

We further localize the variables in ***L-1way*** by zeroing out the correlation between CF and (U,V) from the ***1way*** system. The analysis can still include the uncertainties in the wind field to assist the analysis of atmospheric CO_2, but the error of surface carbon flux is coupled with only the atmospheric CO_2 uncertainty reflecting the fact that carbon flux is only related to low level atmospheric CO_2 and not with the wind. Again, the meteorological variables are not affected by (C, CF), so that the analysis of (U, V, T, q, Ps) are exactly the same as in ***1way***, also true for the ***C-univ*** analysis for the same reason. From Table 2 and Figure 5, we find that ***L-1way*** has the best performance of five schemes for estimating surface CO_2 fluxes, while the result for atmospheric CO_2 concentration is comparable with that from ***1way*** (Figure 4). This implies that the surface carbon fluxes should not be linked to the wind fields in the background error covariance matrix. As a result, the spatial distribution of the analysis from ***L-1way*** in Figure 8 also shows a promising performance in estimating surface carbon fluxes, capturing well the major source regions in the Northern Hemisphere.

The last method considered, ***C-univ***, has stable results in the analysis of the carbon variables, but the surface carbon flux is slightly worse than that of ***L-1way*** (Figure 4, 5 and Table 2). Interestingly, the RMS error of surface carbon analysis grows with time whereas ***L-1way*** keeps reducing the errors (Figure 5). Since these two systems differ only in whether the transport error is considered when analyzing the atmospheric CO_2 concentrations, the gradual increase of RMS error in ***C-univ*** can be seen as a result of neglecting transport errors.

Figure 9 displays global maps of analysis errors in surface CO_2 flux analyses resulting from the ***L-mult, 1way, L-1way,*** and ***C-univ*** experiments (recall that the standard multivariate LETKF without any variable localization blew up after 40 days). ***L-mult*** has a broad area of overall errors (Fig 9a). It is apparent that the presence of an error covariance among all of the atmospheric variables is not helpful for the analysis of carbon, since it just introduces sampling errors. By removing the irrelevant error covariance between carbon and temperature, humidity, and surface pressure from ***L-mult***, the results in ***1way*** show improvement overall (Fig 9b) compared to the multivariate analyses. ***L-1way*** provides further localization between the surface carbon flux and wind fields, compared to ***1way***, and hence obtains the smallest errors in carbon flux analysis. This technique clearly has less error, especially over the oceans, than ***L-mult*** or ***1way***.

The approach embodied in ***C-univ*** has lost the error information contained in the relationship between wind and atmospheric CO_2 uncertainties and hence has somewhat worse results than ***L-1way***. Indeed, over the polar regions (Figure 9d), ***C-univ*** has spurious estimates of surface carbon fluxes in areas where there are large errors in the wind analysis (Figure 6), whereas ***L-1way*** does not have those errors. We also note that the error in ***C-univ*** over the polar region grows with time, and this leads to RMS error increases in Figure 5. Thus, we can conclude that the reason for increasing RMS error in the surface carbon fluxes is that transport errors are not accounted for in ***C-univ***. In addition, experiments with an imperfect model [*Kang*, 2009] indicate that the perfect model assumption underestimates the impact of this deficiency of ***C-univ*** since transport errors are also underestimated in this scenario.

When we allow for time-varying surface CO_2 forcing, the estimation problem becomes more difficult because we are not anymore under a perfect model scenario, since the forecast model does not change the surface fluxes, only changed by the analysis cycle. Thus, the overall errors for both schemes become larger and require further research on potential improvements in the data assimilation techniques (see below). Nevertheless, the relatively small advantage of ***L-1way*** compared to ***C-univ*** observed with constant fluxes (Fig. 9c and Fig 9d) becomes much more significant (Figure 10) indicating that, for this scenario, the estimation of surface CO_2 fluxes from ***L-1way*** is significantly better than that from ***C-univ***. It is important to note that ***L-1way*** outperforms ***C-univ*** especially over the ocean and the Southern Hemisphere where the wind uncertainties are dominant due to the lack of rawinsonde observations. Since the analysis cycle updates surface CO_2 fluxes which in turn force the atmospheric CO_2 forecast for the next six hours, unresolved transport errors when assimilating CO_2 in ***C-univ*** can degrade the analysis of carbon variables more in the case with the time-varying forcing than in the case with a constant forcing.

We note that the adaptive inflation estimation has relatively large changes during the first ten days of the analysis when the errors in the initial conditions of the background states are very large compared to the observation errors (not shown). The adaptive inflation of the background covariance for the meteorological variables, which is estimated initially to be about 35%, settles after spin-up at about 5% ($\rho \approx 1.05$ in Equation 4). The inflation factor estimated for the atmospheric CO_2 concentration also decreases with time: the inflation factor is estimated at about 50% during the first week and then converges in time to less than 10%. These adaptive inflation factors are similar for all the variable localization schemes that we have examined in this study. The inflation for the surface carbon fluxes is estimated to be small, less than 2%, as could be expected for a variable that is not observed [*Anderson*, 2009]. If instead, we allow the inflation for the carbon flux to be the same as for atmospheric CO_2, there is filter divergence in the estimation of the surface carbon flux analysis. Thus, the adaptive inflation estimation algorithm [*Li et al.*, 2009; Zhang et al., 2004] appears to work quite well in the carbon cycle data assimilation system.

5. Summary and Discussion

We have developed a method to estimate surface carbon fluxes via an EnKF data assimilation analyzing the meteorological variables and the carbon variables simultaneously. The method is fairly efficient in terms of computational cost since it does not require an additional run of the transport model as the observation operator during the analysis, a step generally used in previous studies. In addition, simultaneous analyses allow accounting for the important day-to-day wind uncertainties when analyzing CO_2 variables. Atmospheric CO_2 observations are assimilated from a simulated network of *in situ* (continuous record), flask (weekly record), and satellite-based measurements with realistic resolution. The results of this study, although far from perfect, are promising especially considering that no *a priori* information about carbon has been used.

The focus of this paper is a comparison of several "variable localization" schemes that reduce sampling errors in the ensemble estimation of the covariance between physically uncorrelated variables by zeroing out the background error covariance among these variables. Since carbon variables in the nature run do not have a physical relationship with temperature, specific humidity and surface pressure, the standard EnKF approach of coupling errors of all variables in ***mult*** analysis induces sampling error into the system. As a result, the accuracy of ***mult*** analyses for all dynamic variables gets progressively worse together with surface carbon flux estimation until about 40 days, when filter divergence takes place. Of the five new methods introduced here, the localized one-way approach, ***L-1way***, has the best performance in the estimation of surface carbon fluxes. The atmospheric CO_2 analysis includes the error covariance of CO_2 and surface carbon flux as well as the wind transport error, which is strongly related to the forecast of atmospheric CO_2. This approach excludes the non-physical error covariance between the wind field and surface CO_2 flux and among the carbon variables and temperature, humidity, and surface pressure, which are dominated by sampling errors. Moreover, the carbon variables are not allowed to influence the analysis of meteorological variables because CO_2 is poorly observed, and thus would increase the sampling errors in the better observed winds and temperatures [*Liu et al.*, 2009].

The results from ***L-1way*** can be contrasted with ***C-univ***, which is closer to previous studies in a sense that transport error covariances are not considered during the carbon analysis. Nevertheless, the ***C-univ*** approach within EnKF does allow for information on transport uncertainties because the different ensemble members have different winds, and therefore different CO_2 transports. As a result, the carbon univariate approach gives quite good results when we use constant surface fluxes, although slightly worse than those obtained with the ***L-1way*** approach. The improvement of ***L-1way*** over ***C-univ*** becomes much larger when the imperfection of CO_2 forecast becomes important. The advantages of ***L-1way*** results compared to ***C-univ*** results demonstrate that it is necessary to resolve transport error for the analysis of atmospheric CO_2.

We note that the variable localization design of the most successful method in this paper, the ***L-1way***, is based on our OSSE experimental setting since, in our nature run, atmospheric CO_2 is only transported and mixed by the wind fields and the varying CO_2 has no radiative impact and thus no temperature dependence. In a more realistic model, assimilating real observations, the variable localization technique we have introduced needs to be adapted by considering the *"dynamical distance"* between each pair of variables in a real nature and model. If the background error covariance is dominated by sampling errors, it will be beneficial to zero out the covariance as we did here, even if the two variables are, to some extent, physically related. For example, biospheric and air-sea carbon fluxes have diurnal, seasonal, and interannual variabilities that are modulated by precipitation, temperature, cloud cover, relative humidity, and wind speeds. Only if the atmospheric carbon model is realistic enough to represent well the covariability of two of these variables, should the corresponding error covariance be retained. Furthermore, a study with more realistic settings such as using a realistic model and an imperfect model assumption is required as a next step, in order to further examine the impact of this new method on assimilating real observations.

We point out that, in this paper, we introduced the methodology of constraining the unobserved surface CO_2 fluxes by assimilating atmospheric CO_2 observations simultaneously with atmospheric observations allowing transport errors to be considered during the analysis step. In principle, this methodology could be extended to the estimation of surface moisture/heat fluxes from the assimilation of observations of humidity/temperature in the atmosphere, another major challenge in current models.

Finally, we note that the results of these new techniques such as variable localization and adaptive inflation have clearly improved our ability to estimate the surface fluxes, so that these techniques can be used in other Ensemble Kalman Filter data assimilation problems. Nevertheless, since our ultimate goal is to estimate as well as possible not only the atmospheric CO_2 but also the surface carbon fluxes, it is clear that significant more progress is needed, especially in the imperfect model scenario. We are doing research with several promising additional new techniques, including the estimation of the model bias, and the restructuring of ensemble perturbations that in time tend to align themselves too much along the most unstable direction (leading local Lyapunov vectors). The difficulty of the problem makes clear the need to perform OSSEs as well as real data experiments in order to understand what can be achieved with real data and what techniques should be tested.

Acknowledgments

We are grateful to the US Department of Energy for the support of the research project, "Carbon data assimilation with coupled Ensemble Kalman filter", under DOE Grant DEFG0207ER64437. Support was also received from NASA Grants NNX08AD4oG, NNX07AM97G, NOAA Grant NA09OAR4310178, and ONR Grant N000141010557. The SPEEDY model was kindly provided by Franco Molteni and Fred Kucharski. The very constructive suggestions of Andy Jacobson and two anonymous reviewers improved the paper content and presentation.

References

Anderson J. L. (2001), An Ensemble Adjustment Kalman Filter for Data Assimilation. *Mon. Wea. Rev. 129*, 2884-2903.

Anderson, J. L. (2009), Spatially and temporally varying adaptive covariance inflation for ensemble filters. *Tellus*, *61A*, 72–83.

Anderson, J. L. and S.L. Anderson (1999), A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Mon. Wea. Rev. 127***,** 2741–2758.

Andres, R. J., G. Marland, I. Fung, and E. Matthews (1996), A 1° x 1° distribution of carbon dioxide emissions from fossil fuel consumption and cement manufacture, 1950-1990, *Global Biogeochem. Cycles*, *10***,** 419-429.

Annan, J. D., D. J. Lunt, J. C. Hargreaves, and P. J. Valdes (2005), Parameter estimation in an atmospheric GCM. *Nonlinear processes in geophysics*, **12**, 363–371.

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

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

Baker, D. F., Bösch, H., Doney, S. C., O'Brien, D., and Schimel, D. S.: Carbon source/sink information provided by column CO_2 measurements from the Orbiting Carbon Observatory, *Atmos. Chem. Phys.*, *10,* 4145-4165, doi:10.5194/acp-10-4145-2010.

Bishop, C. H., B. J. Etherton, and S. J. Majumdar (2001), Adaptive Sampling with the Ensemble Transform Kalman Filter. Part I: Theoretical Aspects. *Mon. Wea. Rev. 129*, 420-436.

Bousquet, P., P. Peylin, P. Ciais, C. Le Quere, P. Friedlingstein, and P. P. Tans (2000), Regional changes in carbon dioxide fluxes of land and oceans since 1980. *Science*, *290,* 1342-1346.

Chevallier, F., R. J. Engelen, C. Carouge, T. J. Conway, P. Peylin, C. Pickett-Heaps, M. Ramonet, P. J. Rayner, and I. Xueref-Remy (2009a), AIRS-based versus flask-based estimation of carbon surface fluxes, *J. Geophys. Res.*, *114*, D20303, doi:10.1029/2009JD012311.

Chevallier, F., S. Maksyutov, P. Bousquet, F.-M. Bréon, R. Saito, Y. Yoshida, and T. Yokota (2009b), On the accuracy of the CO_2 surface fluxes to be estimated from the GOSAT observations, *Geophys. Res. Lett.*, *36*, L19807, doi:10.1029/2009GL040108.

Crisp, D., R. M. Atlas, F.-M.Breon, L. R. Brown, J. P. Burrows, P. Ciais, B. J. Connor, S. C. Doney, I. Y. Fung, D. J. Jacob, E. Miller D. O'Brien, S. Pawson, J. T. Randerson, P. Rayner, R. J. Salawitch, S. P. Sander, B. Sen, G. L. Stephens, P. P. Tans, G. C. Toon, P. O. Wennberg, S. C. Wofsy, Y. L. Yung, Z. Kuang, B. Chudasama, G. Sprague, B. Weiss, R. Pollock, D. Kenyon, and S. Schroll (2004), The Obiting Carbon Observatory (OCO) mission. *Advances in Space Research*, *34*, 700-709

Desroziers G., L. Berre, B. Chapnik, and P. Poli (2005), Diagnosis of observation, background and analysis error statistics in observation space. *Quart. J. Roy. Meteor. Soc.*, *131*, 3385-3396.

Engelen, R. J., S. Serrar, and F. Chevallier (2009), Four-dimensional data assimilation of atmospheric CO_2 using AIRS observations, *J. Geophys. Res.*, *114*, D03303, doi:10.1029/2008JD010739.

Enting, I. G. (2002), Inverse Problems in Atmospheric Constituent Transport, Cambridge University Press, N. Y.

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

Feng, L., Palmer, P. I., Bösch, H., and Dance, S.: Estimating surface CO_2 fluxes from space-borne CO_2 dry air mole fraction observations using an ensemble Kalman Filter, *Atmos. Chem. Phys., 9,* 2619-2633, doi:10.5194/acp-9-2619-2009, 2009.

Gaspari, G., and S. E. Cohn (1999), Construction of correlation functions in two and three dimensions. *Quart. J. Roy. Meteor. Soc., 125*, 723-757.

Greybush, S., E. Kalnay, T. Miyoshi, K. Ide, and B. Hunt (2010), Balance and Ensemble Kalman Filter Localization Techniques, *Mon. Wea. Rev*., in press, doi: 10.1175/2010MWR3328.1.

Gurney, K. R., R. M. Law, A. S. Denning, P. J. Rayner, B. C. Pak, D. Baker, P. Bousquet, L. Bruhwiler, Y. H. Chen, P. Ciais, I. Y. Fung, M. Heimann, J. John, T. Maki, S. Maksyutov, P. Peylin, M. Prather, and S. Taguchi (2004), Transcom 3 inversion intercomparison: Model mean results for the estimation of seasonal carbon sources and sinks. *Global Biogeochemical Cycles*, *18(1),* 10.1029/2003GB002111.

Hamill, T. M., J. S. Whitaker, and C. Snyder (2001), Distance-Dependent Filtering of Background Error Covariance Estimates in an Ensemble Kalman Filter. *Mon. Wea. Rev.*, *129*, 2776-2789.

Hollingsworth, A., et al. (2008), The Global Earth-system Monitoring using Satellite and in-situ data (GEMS) Project: Towards a monitoring and forecasting system for atmospheric composition, *Bull. Am. Meteorol. Soc., 89,* 1147–1164, doi:10.1175/2008BAMS2355.1.

Houtekamer, P. L. and H. L. Mitchell (2001), A Sequential Ensemble Kalman Filter for Atmospheric Data Assimilation. Monthly Weather Review: Vol. 129, pp. 123–137.

Hunt, B. R., E. Kostelich, and I. Szunyogh (2007), Efficient Data Assimilation for Spatiotemporal Chaos: a Local Ensemble Transform Kalman Filter, *Physica D*, *230*, 112-126.

Kang, J. (2009), Carbon Cycle Data Assimilation Using a Coupled Atmosphere-Vegetation Model and the Local Ensemble Transform Kalman Filter, Ph.D Thesis, University of Maryland

Li, H., E. Kalnay, and T. Miyoshi, (2009), Simultaneous estimation of covariance inflation and observation errors within ensemble Kalman filter. *Quart. J. Roy. Meteor. Soc., 135,* 523-533.

Liu, J., E. Kalnay, I. Fung, M. Chahine, and E. T. Olsen (2011), Assimilation of AIRS CO2 observations with an EnKF in a Carbon-Climate Model, paper presented at 91st American Meteorlolgical Society Annual Meeting, Seattle, WA, USA.

Liu, J., H. Li, E. Kalnay, E.J. Kostelich, and I. Szunyogh (2009), Univariate and Multivariate Assimilation of AIRS Humidity Retrievals with the Local Ensemble Transform Kalman Filter. *Mon. Wea. Rev.*, *137*, 3918–3932.

Miyazaki, K. (2009), Performance of a local ensemble transform Kalman filter for the analysis of atmospheric circulation and distribution of long-lived tracers under idealized conditions, *J. Geophys. Res.*, *114*, D19304, doi:10.1029/2009JD011892.

Miyoshi, T (2005), Ensemble Kallman Filter Experiments with a Primitive-Equation Global Model, Ph.D Thesis, University of Maryland

Molteni, F. (2003), Atmospheric simulations using a GCM with simplified physical parametrizations. I: Model climatology and variability in multi-decadal experiments. *Climate Dyn.*, *20*, 175-191.

Ott, E., B. R. Hunt, I. Szunyogh, A. V. Zimin, E. J. Kostelich, M., Corazza, E. Kalnay, D. J. Patil, and J. A. Yorke (2004), Estimating the state of large spatio- temporally chaotic systems. *Phys. Lett. A., 330***,** 365- 370.

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. USA.*, *104*, 18,925– 18,930.

Rödenbeck, C., S. Houweling, M. Gloor, 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.

Stephens, B. B., et al. (2007), Weak northern and strong tropical land carbon ptake from vertical profiles of atmospheric CO_2, *Science, 316*, 1732– 1735.

Takahashi, T., S. C. Sutherland, C. Sweeney, A. Poisson, N. Metzl, B. Tilbrook, N. Bates, R. Wanninkhof, R. A. Feely, C. Sabine, J. Olafsson, Y. Nojiri, 2002: Global sea-air CO_2 flux based on climatological surface ocean pCO_2, and seasonal biological and temperature effects, *Deep-Sea Research II*, *49*, 1601-1622.

Wang, H., D. J. Jacob, M. Kopacz, D. B. A. Jones, P. Suntharalingam, J. A. Fisher, R. Nassar, S. Pawson, and J. E. Nielsen (2009), Error correlation between CO_2 and CO as constraint for CO_2 flux inversions using satellite data, *Atmos. Chem. Phys.,* 9, 7313-7323.

Yokota, T., H. Oguma, I. Morino, and G. Inoue (2004), A nadir looking SWIR FTS to monitor CO_2 column density for Japanese GOSAT project, *Proc. Twenty-fourth Int. Sympo. on Space Technol. and Sci (Selected Papers), JSASS and Organizing Comm. of the 24th ISTS*, 887–889.

Zeng, N., A. Mariotti, and P. Wetzel, 2005: Terrestrial mechanisms of interannual CO_2 variability, *Global Biogeochemical Cycles*, *19*, GB1016, doi:10.1029/2004GB002273.

Zhang F., C. Snyder, and J. Sun (2004), Impacts of initial estimate and observation availability on convective-scale data assimilation with an ensemble Kalman filter. *Mon. Wea. Rev.*, *132*, 1238–1253.

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

Zupanski, M., 2005: Maximum Likelihood Ensemble Filter: Theoretical Aspects. *Mon. Wea. Rev.*, *133*, 1710–1726. doi: 10.1175/MWR2946.1

Zupanski, M., S. J. Fletcher, I. M. Navon, B. Uzunoglu, R. P. Heikes, D. A. Randall, T. D. Ringlee and D. Daescu, 2006: Initiation of ensemble data assimilation. *Tellus*, *58A*, 159-170.

Figure Captions

Figure 1. Schematic plots of background error covariance matrices ($P^b=(x^b)(x^b)^T/(K-1)$) for (a) *mult*, (b) *L-mult*, (c) *1way*, (d) *L-1way*, and (e) *C-univ* analysis systems. Here, C indicates atmospheric CO_2 concentration and CF indicates surface carbon fluxes. The colors of the variable names are matched with the system used for their updates. White areas with “no” indicate the error correlation between variables is assumed to be zero during the analysis while areas with “yes” indicate that the errors are allowed to be correlated. For example, in 1(d), the errors of the standard atmospheric variables are coupled, the atmospheric CO_2 errors are coupled with those of the wind but the wind errors are not coupled with the CO_2 errors (1-way coupling), and the surface carbon flux errors are only coupled with the CO_2 errors.

Figure 2. True state of (a) surface CO_2 fluxes (6 PgC/yr) and (b) atmospheric CO_2 concentrations in the lowest layer at the initial time, as well as initial ensemble mean of (c) surface CO_2 fluxes and (d) atmospheric CO_2. Units for atmospheric CO_2 concentration are ppmv, units for surface CO_2 fluxes are 10^{-9} kg/m^2/s.

Figure 3. The simulated observational coverage of (a) meteorological variables (black dots) and (b) atmospheric CO_2 concentration (gray lines: GOSAT column data, crosses: continuous *in situ* data, closed circles: weekly flask data).

Figure 4. Time series of global RMS error of (a) U (m/s) and (b) atmospheric CO_2 concentration in the lowest layer (ppmv) for four months of analysis. (solid gray: *mult*, solid black: *L-mult*, dashed gray: *1way*, dashed black: *L-1way*, dotted light gray: *C-univ*)

Figure 5. Same as Figure 4, except for the surface CO_2 fluxes.

Figure 6. Analysis error (unit: m/s) of (a) zonal wind and (b) meridional wind from the localized multivariate analysis (*L-mult*) for the last three months of data assimilation. (c) and (d): The same as in (a) and (b) except from *1way*. Shading indicates positive errors and contours indicate negative errors with the same color scale as the shading.

Figure 7. Analysis (left column) of atmospheric CO_2 concentration in the lowest layer and its error (right column) after four months of analysis. (a) and (b) results from *L-mult*, (c) and (d) from *1way*. Units are ppmv. Shading indicates positive errors and contours indicate negative errors with the same color scale as the shading.

Figure 8. (a) True state of surface CO2 fluxes and (b) the analysis after four months of the L-1way (localized 1-way multivariate) data assimilation. (Units are 10^{-9} kg/m^2/s.)

Figure 9. Analysis errors of surface CO_2 fluxes after four months of analysis. (a) results from *L-mult,* (b) from *1way*, (c) from *L-1way* and (d) from *C-univ*. Units are 10^{-9} kg/m^2/s. (Shading indicates positive errors and contours indicate negative errors with the same color scale as the shading.)

Figure 10. (a) True state of surface CO_2 fluxes from a time-varying terrestrial and oceanic forcing and a fossil fuel emission, and the estimated surface CO_2 fluxes from (b) *L-1way*, and (c) *C-univ* data assimilation for the last two months (November-December) of one-year analysis

Table 1. Standard deviation of errors used in creating the simulated observations.

Variable	Std. dev. of error
U	1.0 m/s
V	1.0 m/s
T	1.0 K
q	0.1 g/kg
Ps	1.0 hPa
C	1.0 ppmv

Table 2. RMS error of variables from different localization schemes for the last week of four-month analysis: one-week time average of every six hour values of $\sqrt{\sum_{i}^{n}(x_i^a - x_i^t)^2 \big/ n}$, where x_i^a / x_i^t is the analysis/the truth at i-th point, and n is the total number of grid points (units: CF=10^{-9}kg/m^2/s, C=ppmv, U and V=m/s, T=K, q=g/kg, Ps=hPa). The errors corresponding to Ensemble Kalman Filter divergence are symbolically represented as “infinite”.

<table>
<tr><th></th><th>mult</th><th>L-mult</th><th>1way</th><th>L-1way</th><th>C-univ</th></tr>
<tr><td>CF</td><td>∞</td><td>8.65</td><td>6.10</td><td>5.65</td><td>5.79</td></tr>
<tr><td>C</td><td>∞</td><td>1.05</td><td>0.68</td><td>0.71</td><td>0.67</td></tr>
<tr><td>U</td><td>∞</td><td>1.32</td><td colspan="3">1.32</td></tr>
<tr><td>V</td><td>∞</td><td>1.22</td><td colspan="3">1.20</td></tr>
<tr><td>T</td><td>∞</td><td>0.53</td><td colspan="3">0.54</td></tr>
<tr><td>q</td><td>∞</td><td>0.34</td><td colspan="3">0.35</td></tr>
<tr><td>Ps</td><td>∞</td><td>1.15</td><td colspan="3">1.14</td></tr>
</table>

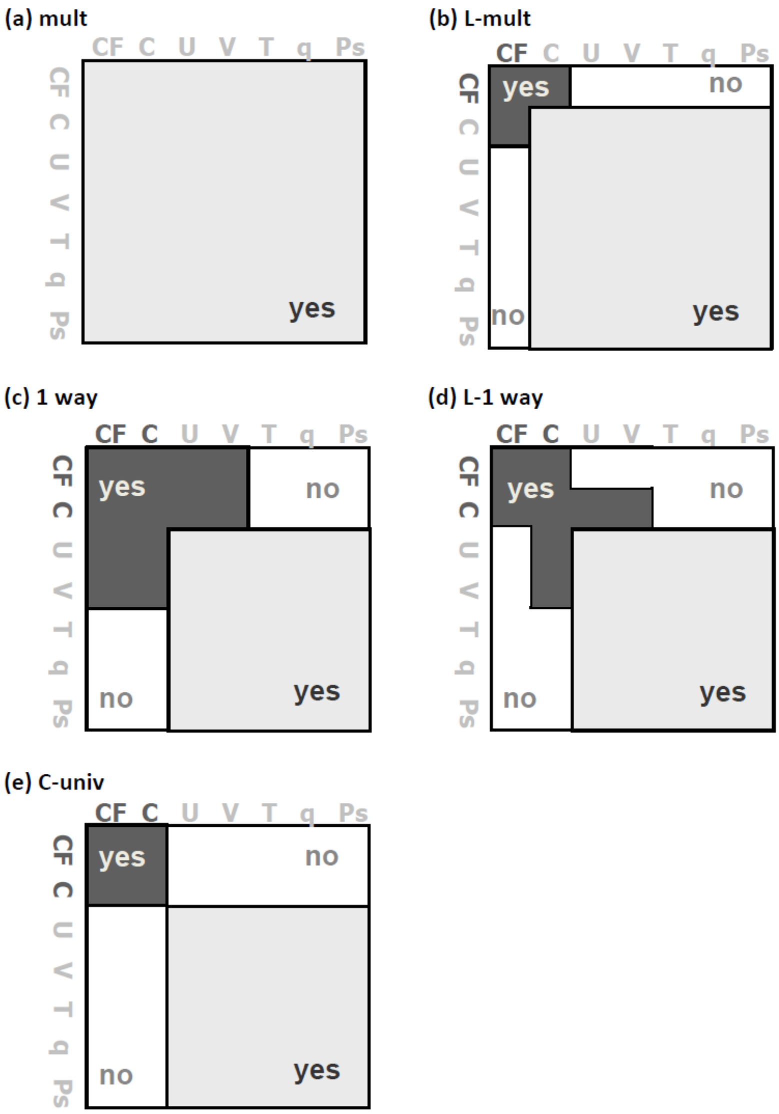

Figure 1. Schematic plots of background error covariance matrices ($P^b=(x^b)(x^b)^T/(K-1)$) for (a) *mult*, (b) *L-mult*, (c) *1way*, (d) *L-1way*, and (e) *C-univ* analysis systems. Here, C indicates atmospheric CO_2 concentration and CF indicates surface carbon fluxes. The colors of the variable names are matched with the system used for their updates. White areas with "no" indicate the error correlation between variables is assumed to be zero during the analysis while areas with "yes" indicate that the errors are allowed to be correlated. For example, in 1(d), the errors of the standard atmospheric variables are coupled, the atmospheric CO_2 errors are coupled with those of the wind but the wind errors are not coupled with the CO_2 errors (1-way coupling), and the surface carbon flux errors are only coupled with the CO_2 errors.

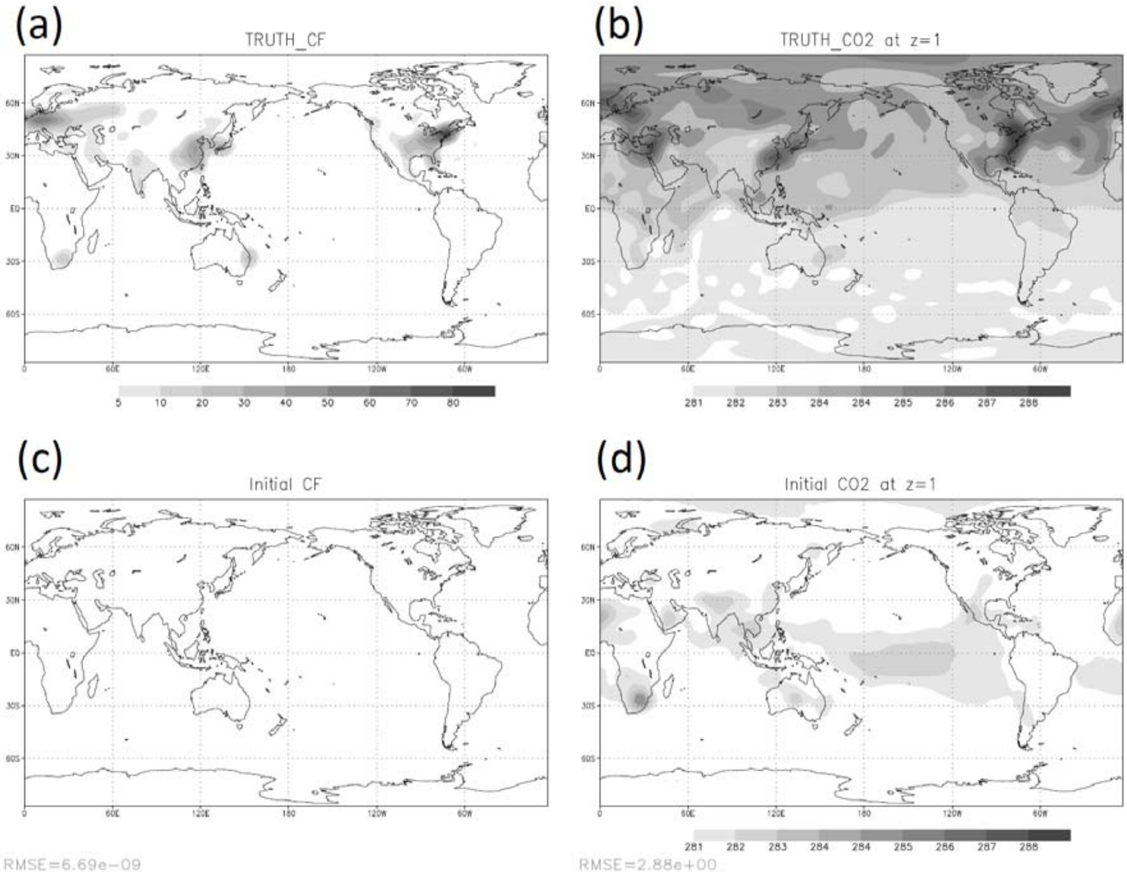

Figure 2. True state of (a) surface CO_2 fluxes (6 PgC/yr) and (b) atmospheric CO_2 concentrations in the lowest layer at the initial time, as well as initial ensemble mean of (c) surface CO_2 fluxes and (d) atmospheric CO_2. Units for atmospheric CO_2 concentration are ppmv, units for surface CO_2 fluxes are 10^{-9} kg/m^2/s.

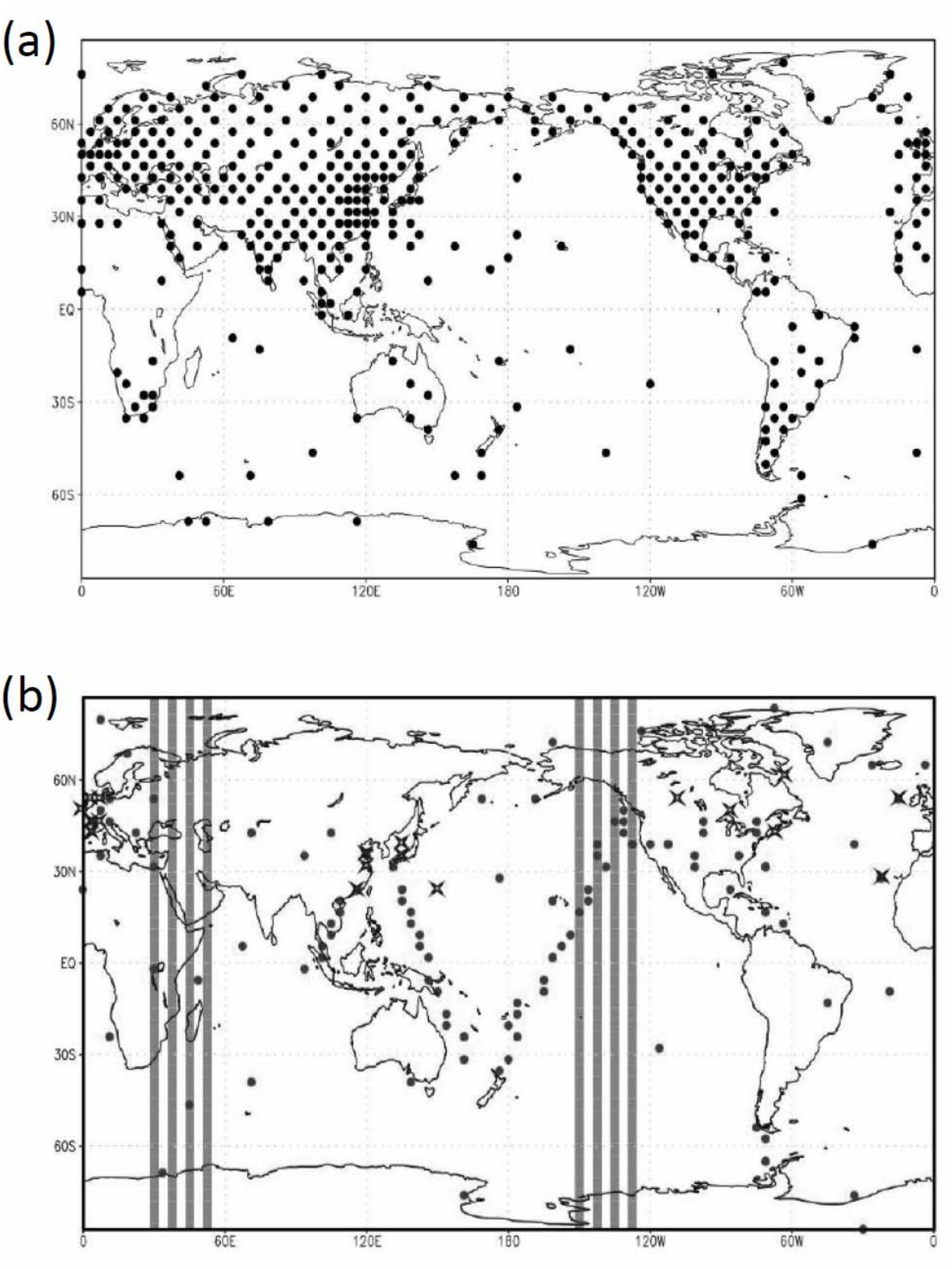

Figure 3. The simulated observational coverage of (a) meteorological variables (black dots) and (b) atmospheric CO_2 concentration (gray lines: GOSAT column data, crosses: continuous *in situ* data, closed circles: weekly flask data).

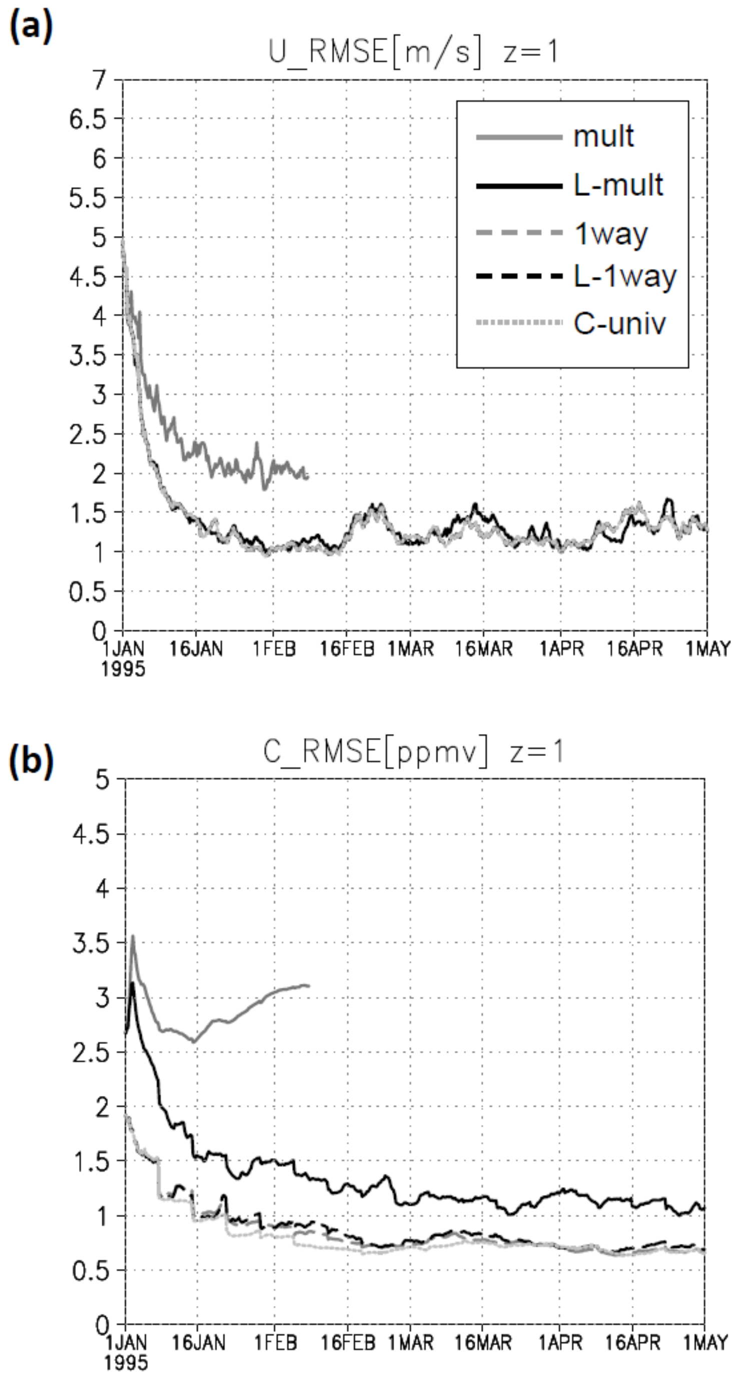

Figure 4. Time series of global RMS error of (a) U (m/s) and (b) atmospheric CO_2 concentration in the lowest layer (ppmv) for four months of analysis. (solid gray: *mult*, solid black: *L-mult*, dashed gray: *1way*, dashed black: *L-1way*, dotted light gray: *C-univ*)

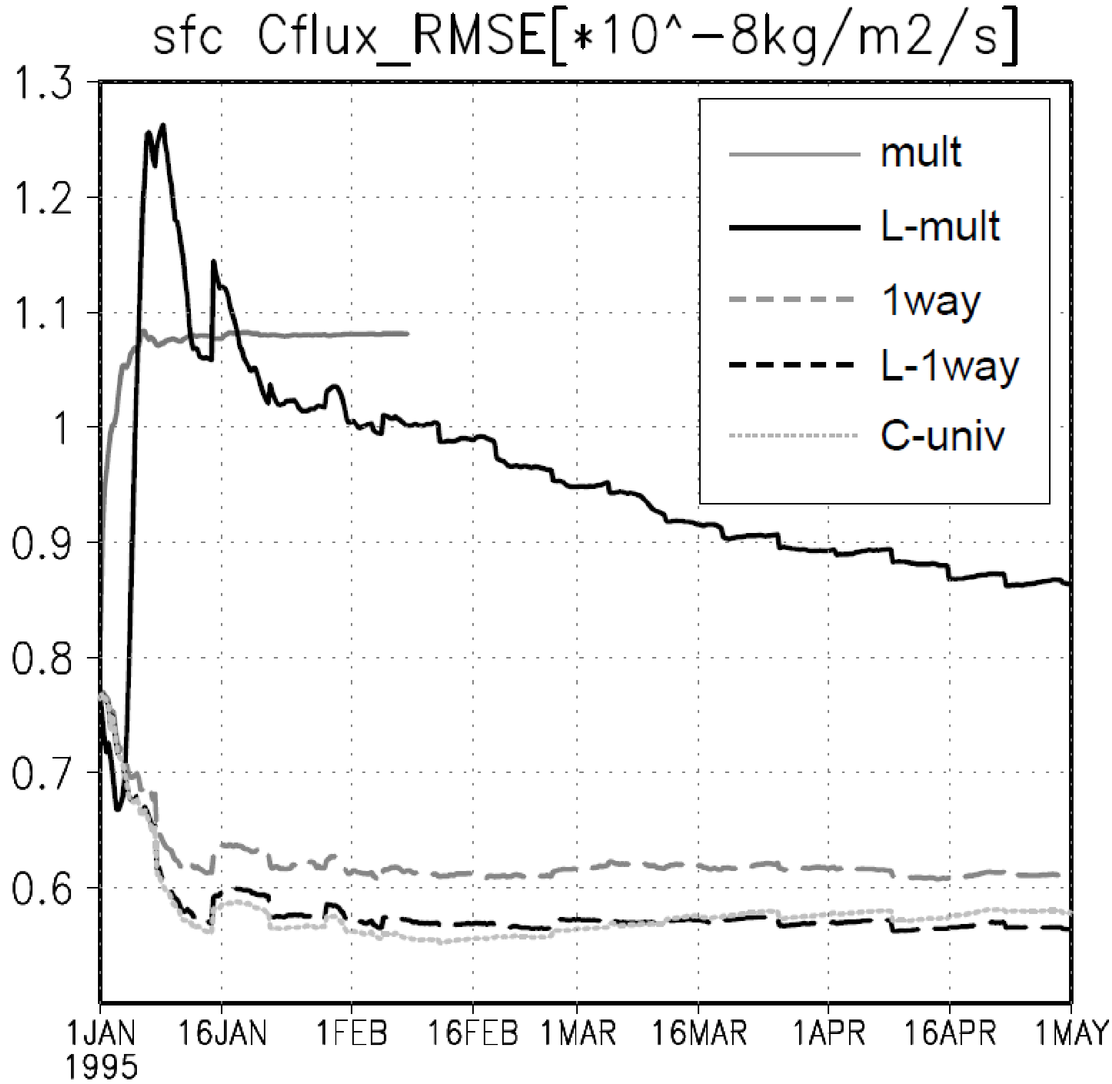

Figure 5. Same as Figure 4, except for the surface CO_2 fluxes.

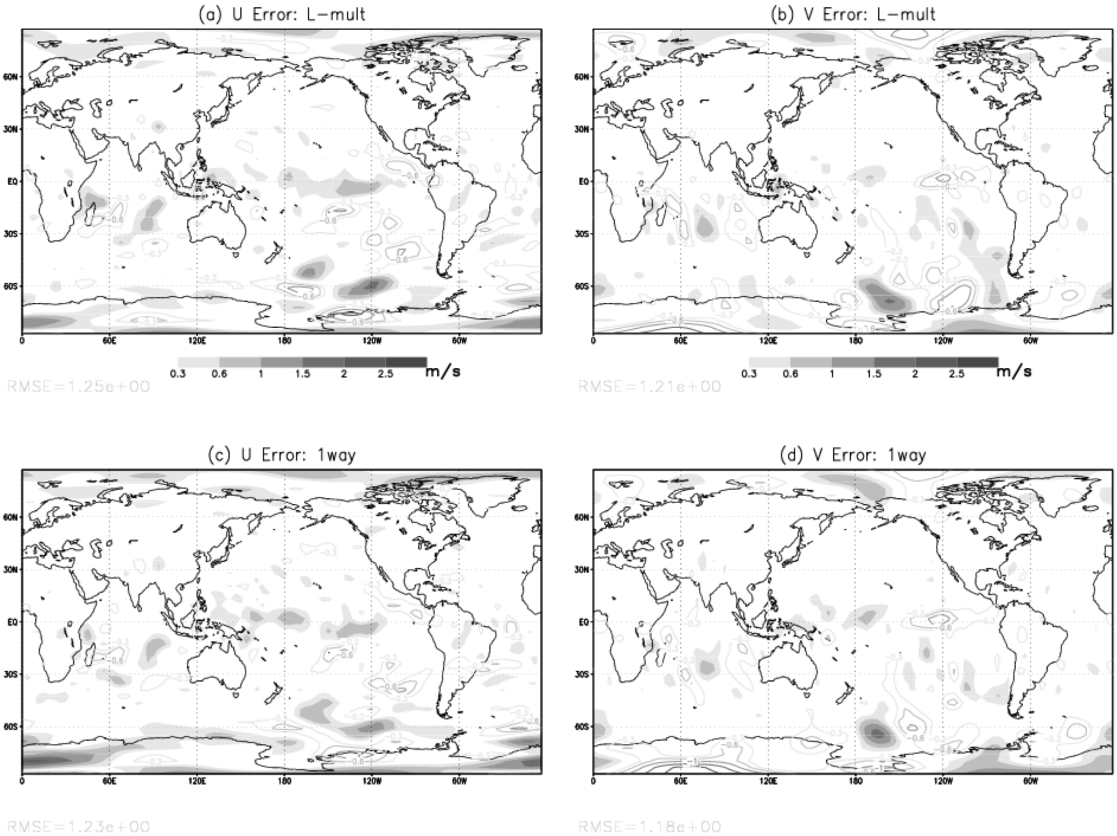

Figure 6. Analysis error (unit: m/s) of (a) zonal wind and (b) meridional wind from the localized multivariate analysis (*L-mult*) for the last three months of data assimilation. (c) and (d): The same as in (a) and (b) except from *1way*. Shading indicates positive errors and contours indicate negative errors with the same color scale as the shading.

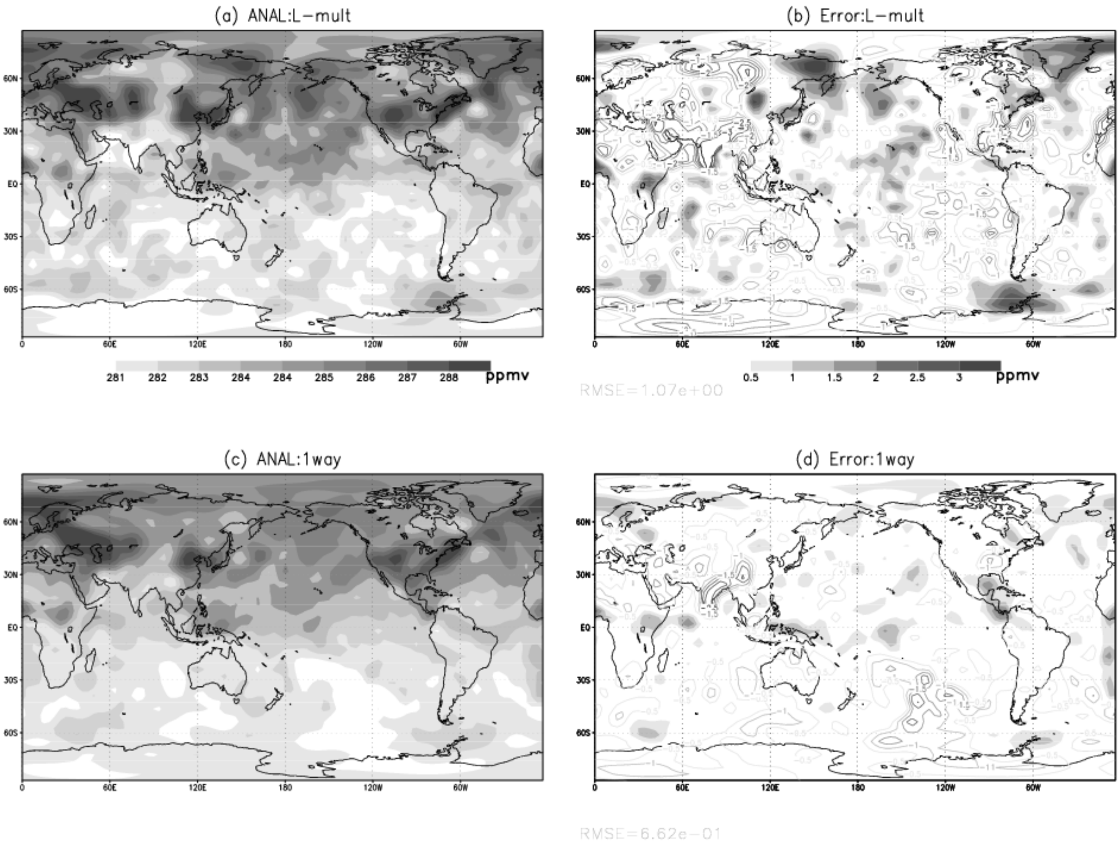

Figure 7. Analysis (left column) of atmospheric CO_2 concentration (ppmv) in the lowest layer and its error (right column) after four months of analysis. (a) and (b) results from *L-mult*, (c) and (d) from *1way*. Units are ppmv. Shading indicates positive errors and contours indicate negative errors with the same color scale as the shading.

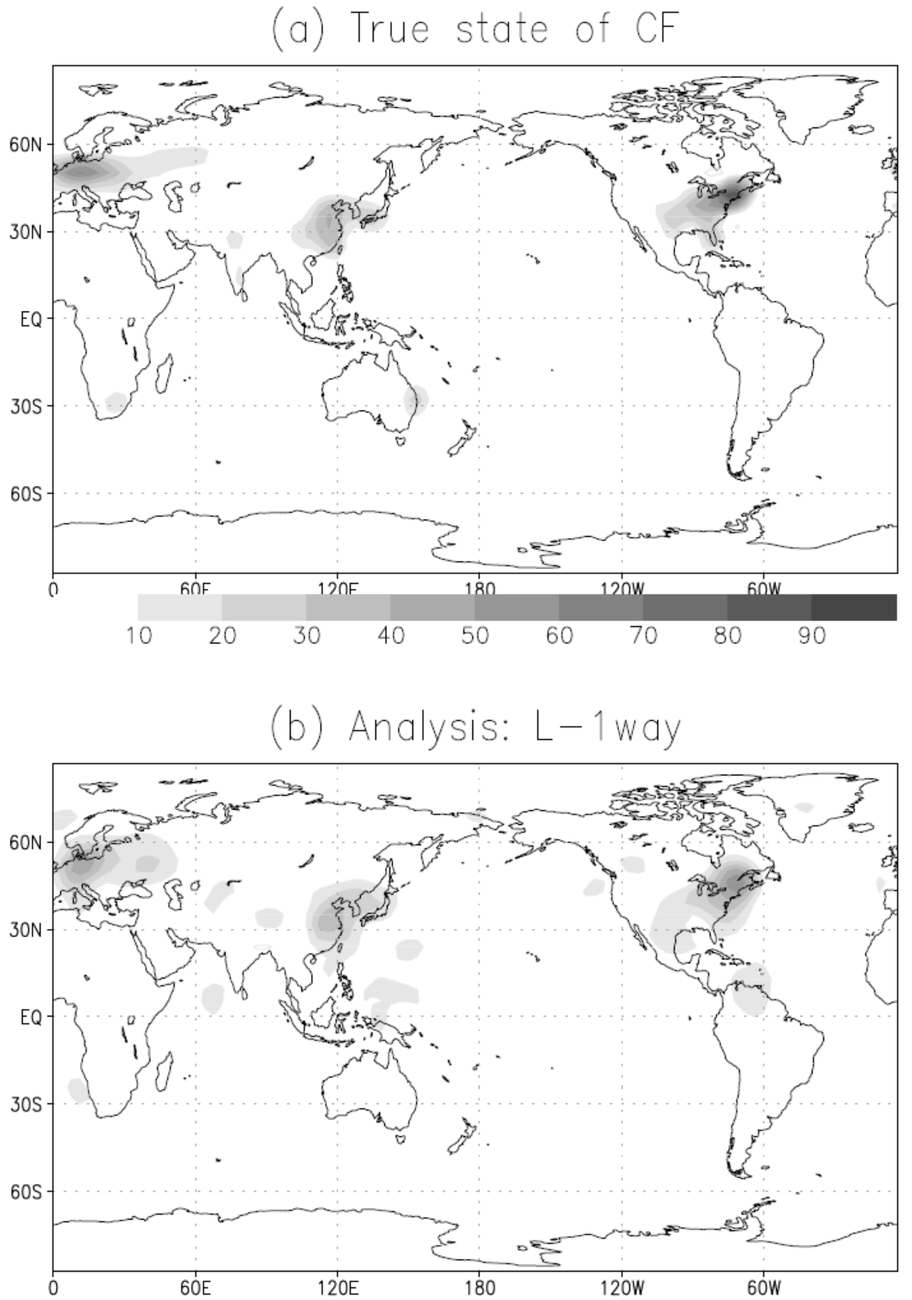

Figure 8. (a) True state of surface CO2 fluxes and (b) the analysis after four months of the L-1way (localized 1-way multivariate) data assimilation. (Shading indicates positive errors and contours indicate negative errors with the same contour scale as the shading. Units are 10^{-9} kg/m^2/s)

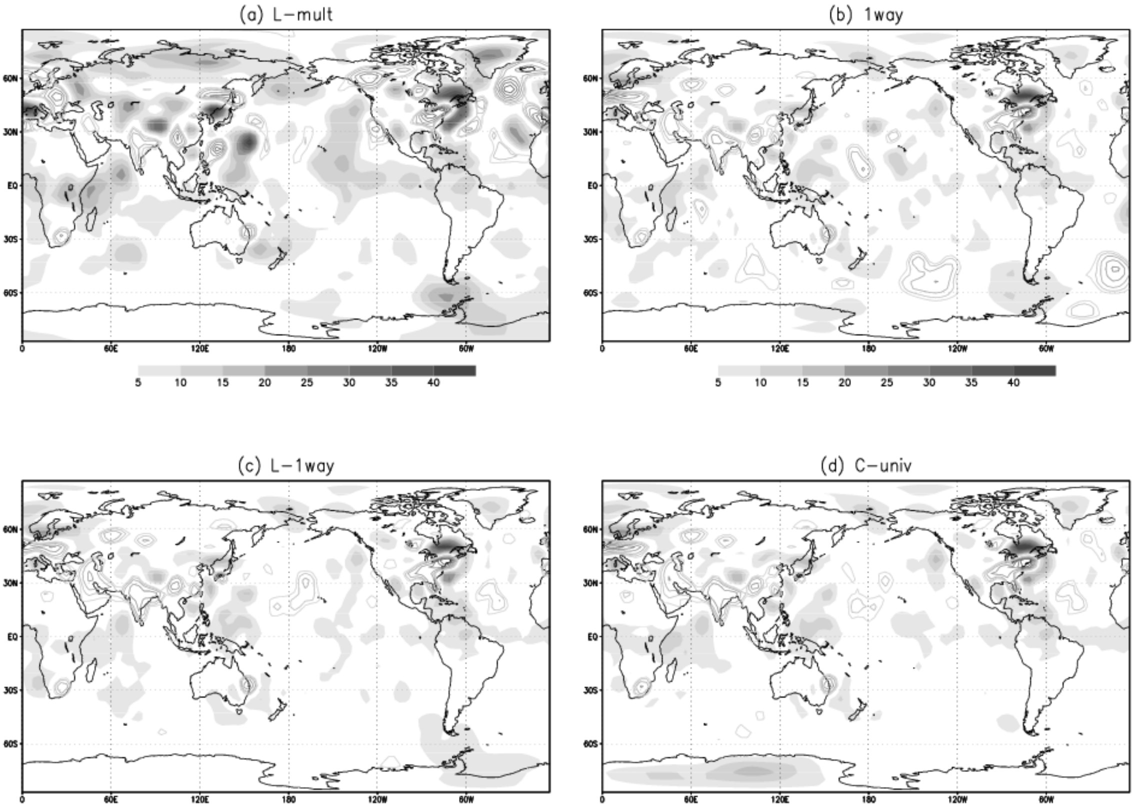

Figure 9. Analysis errors of surface CO_2 fluxes after four months of analysis. (a) results from *L-mult,* (b) from *1way*, (c) from *L-1way* and (d) from *C-univ*. Units are 10^{-9} kg/m^2/s. (Shading indicates positive errors and contours indicate negative errors with the same contour scale as the shading.)

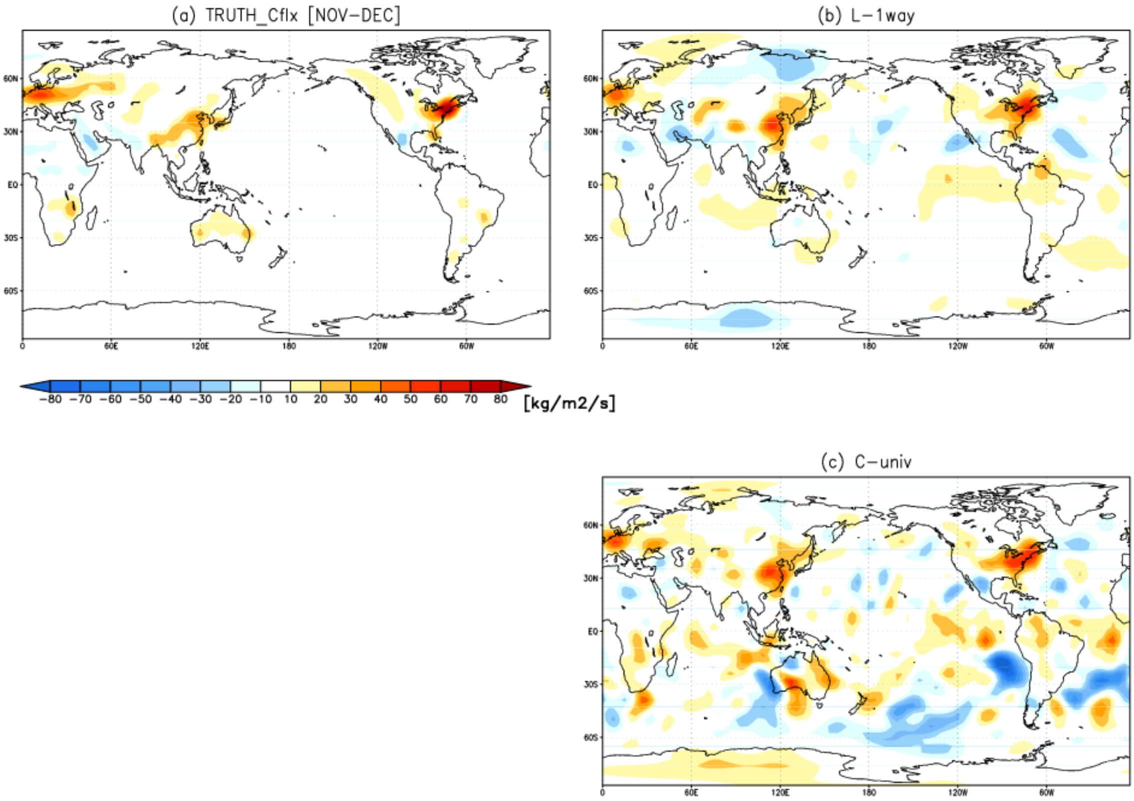

Figure 10. (a) True state of surface CO_2 fluxes from a time-varying terrestrial and oceanic forcing and a fossil fuel emission, and the estimated surface CO_2 fluxes from (b) *L-1way*, and (c) *C-univ* data assimilation for the last two months (November-December) of one-year analysis.