An ensemble data assimilation system to estimate CO_2 surface fluxes from atmospheric trace gas observations

W. Peters,[1,2] J. B. Miller,[1,2] J. Whitaker,[1,2] A. S. Denning,[3] A. Hirsch,[1,2] M. C. Krol,[4] D. Zupanski,[5] L. Bruhwiler,[1] and P. P. Tans[1]

Received 29 April 2005; revised 25 August 2005; accepted 20 October 2005; published 23 December 2005.

[1] We present a data assimilation system to estimate surface fluxes of CO_2 and other trace gases from observations of their atmospheric abundances. The system is based on ensemble data assimilation methods under development for Numerical Weather Prediction (NWP) and is the first of its kind to be used for CO_2 flux estimation. The system was developed to overcome computational limitations encountered when a large number of observations are used to estimate a large number of unknown surface fluxes. The ensemble data assimilation approach is attractive because it returns an approximation of the covariance, does not need an adjoint model or other linearization of the observation operator, and offers the possibility to optimize fluxes of chemically active trace gases (e.g., CH_4, CO) in the same framework. We assess the performance of this new system in a pseudodata experiment that resembles the real problem we will apply this system to. The sensitivity of the method to the choice of several parameters such as the assimilation window size and the number of ensemble members is investigated. We conclude that the system is able to provide satisfactory flux estimates for the relatively large scales resolved by our current observing network and that the loss of information in the approximated covariances is an acceptable price to pay for the efficient computation of a large number of surface fluxes. The full potential of this data assimilation system will be used for near–real time operational estimates of North American CO_2 fluxes. This will take advantage of the large amounts of atmospheric data that will be collected by NOAA-CMDL in conjunction with the implementation of the North American Carbon Program (NACP).

Citation: 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.

1. Introduction

[2] Studies of the carbon cycle based on observations of atmospheric concentration patterns have been ongoing for several decades [e.g., *Tans et al.*, 1990; *Conway et al.*, 1994; *Ciais et al.*, 1995; *Denning et al.*, 1995; *Francey et al.*, 1995; *Keeling et al.*, 1996; *Fan et al.*, 1998; *Gurney et al.*, 2002]. One branch of these studies is atmospheric transport inversions, in which net CO_2 exchange across the Earth's surface is deduced ''top-down'' from CO_2 concentration measurements in the atmosphere (see *Enting* [2002] for a discussion of these methods). Tracer transport models with varying degrees of sophistication provide a link between observations and net CO_2 fluxes. The majority of in situ observations are from NOAA CMDL's Cooperative Air Sampling Network, in which flasks are filled at a large number of sites and analyzed in the laboratory to determine concentrations of CO_2, CH_4, and several other species. Currently, this network is undergoing rapid expansion specifically across North America in support of the North American Carbon Program (NACP) [*Wofsy and Harriss*, 2002]. NACP aims to provide detailed knowledge on the North American carbon cycle, and atmospheric transport inversions using surface, airborne, and tall tower observations of CO_2 and related trace species are an important component of this program. However, new methods to combine models and observations are needed to optimally exploit the large number of new observations and provide detailed estimates of carbon fluxes and their uncertainties.

[3] Most previous atmospheric transport inversions of CO_2 aimed to solve a problem with several years (~10–20) of monthly fluxes at a limited number (22–100) of

[1]NOAA Earth Systems Research Lab, Boulder, Colorado, USA.
[2]Cooperative Institute for Research in Environmental Sciences, Boulder, Colorado, USA.
[3]Department of Atmospheric Science, Colorado State University, Fort Collins, Colorado, USA.
[4]Netherlands Institute for Space Research, Utrecht, Netherlands.
[5]Cooperative Institute for Research in the Atmosphere, Fort Collins, Colorado, USA.

0148-0227/05/2005JD006157$09.00

large regions as unknowns [e.g., *Rayner and Law*, 1999; *Bousquet et al.*, 2000; *Peylin et al.*, 2002; *Gurney et al.*, 2002; *Law et al.*, 2003; *Maksyutov et al.*, 2003]. The solution was sought as an improvement of existing flux estimates that was optimally consistent, in a Bayesian sense, with the atmospheric observations. The flux estimates to be improved were derived "bottom-up" from several sources such as oceanic pCO_2 measurements [*Takahashi et al.*, 2002], fossil fuel burning estimates [*Andres et al.*, 1996], and biosphere flux estimates from process models [*Randerson et al.*, 1997]. The atmospheric constraints in most of these studies were in the form of the GlobalView data product [*Masarie and Tans*, 1995]; a gap-filled, time-smoothed representation of the real CO_2 observations that reflects mostly slow seasonal variations in the well-mixed background atmosphere away from strong sources. As a result, atmospheric transport inversions of this kind were quite successful at delineating continental-scale flux variations at seasonal to decadal timescales, but lacked the temporal and spatial detail needed to study the carbon cycle at regional scales.

[4] Notable exceptions to these large-region approaches are the geostatistical inversion presented by *Michalak et al.* [2004], and the grid-scale inversions of *Kaminski et al.* [1999], *Houweling et al.* [1999], *Rödenbeck et al.* [2003], and *Peylin et al.* [2005]. All these studies estimated fluxes at a spatial scale of several degrees. Even though the number of observations did not increase over previous inversions, *Michalak et al.* [2004] and *Rödenbeck et al.* [2003] were able to solve for a greatly increased number of unknown fluxes by using a prespecified covariance structure as an extra constraint on the solution. In the work by *Michalak et al.* [2004], a Bayesian system without prespecified mean fluxes was solved resulting in a top-down CO_2 flux estimate. The comparison of such an independent estimate to bottom-up estimates of CO_2 fluxes lends credibility to both methods. *Peylin et al.* [2005] were the first to use continuous CO_2 measurements from six sites in Europe to estimate daily CO_2 fluxes at the model grid scale. This study spanned only a one month period though, and the authors describe their technique as suitable for intensive campaigns.

[5] Both in the grid-scale and large-region inversions, the resulting system of linear equations was solved in one large effort usually employing singular value decomposition to invert large matrices. However, the most computationally expensive step in these inversions was to establish the linear relationship between the unknown surface CO_2 fluxes and the atmospheric CO_2 observations. These relationships are sometimes called Green's functions, source-receptor relationships, base functions, or observation operators. This last term will be used throughout this paper. The observation operators were constructed prior to the actual inversion using a tracer transport model. This precalculation requires multiple simulations with expensive tracer transport models, equal to the number of unknown fluxes or the number of observations depending on which one is smallest and whether an adjoint of the tracer transport algorithm is available. Each simulation spans a year or more since the "atmospheric memory" for CO_2 fluxes is quite long. This is due to the large distance between the location of the emissions and the location of most measurement sites, the slow decrease of CO_2 gradients in the atmosphere, and because CO_2 is inert in the atmosphere except at very high altitudes.

[6] To answer the specific questions outlined in the NACP program [*Wofsy and Harriss*, 2002], fluxes need to be estimated in more detail thus increasing the number of unknowns by at least two orders of magnitude compared to most previous studies. This is only viable with the planned increase in measurement frequency and density under NACP. As a result, the effort of precalculating the observation operators becomes too large even for today's supercomputers, and the resulting set of equations cannot be solved by traditional batch methods because of the sheer size of the matrices involved. However, extensive experience in optimizing such a large (and even several orders of magnitude larger) number of unknowns using many observations is available from Numerical Weather Prediction (NWP) research. Methods commonly used in that branch of research were first applied to assimilate trace gas concentrations [*Lyster et al.*, 1997; *Miller et al.*, 1999; *Menard et al.*, 2000; *Khattatov et al.*, 2000; *Eskes et al.*, 2003; *Stajner and Wargan*, 2004], and are now making their way into the trace gas flux estimation problem [*Kleiman and Prinn*, 2000; *Pétron et al.*, 2004; *Yudin et al.*, 2004].

[7] One such innovation from NWP methods was used in *Bruhwiler et al.* [2005], where instead of solving the Bayesian system in one large operation, smaller subsets of unknowns were optimized in a time stepping approach called a fixed lag Kalman smoother [*Cohn et al.*, 1994; see also *Hartley and Prinn*, 1993]. This reduced the effort of precalculating the observation operators to 6–9 months per unknown instead of more than a year, and greatly reduced the size of the matrices involved in the inversion. *Bruhwiler et al.* [2005] demonstrated that the fixed lag Kalman smoother approach gives the same result as the traditional batch approach. However, the targeted scales in the NACP program would still yield covariance matrices that are too large to handle even in the fixed lag Kalman smoother. Moreover, the Kalman smoother is not suited to assimilate (quasi-)continuous observations because of the expensive precalculation of observation operators.

[8] In this work, we will expand on the fixed lag Kalman smoother by introducing three further innovations taken from NWP methods: (1) the representation of covariances by an ensemble instead of by a full covariance matrix [*Houtekamer and Mitchell*, 1998], (2) propagation of the state by a dynamical model, which precludes the need for precalculcated prior flux estimates, and (3) replacement of precalculated observation operators by a forward operator working on the ensemble. Together, these three innovations bring the targeted scales within our reach at acceptable computational costs.

[9] The resulting fixed lag Ensemble Kalman Smoother will be referred to as SEAT-A (System for Ensemble Assimilation of Tracers in the Atmosphere) and is the first application of ensemble data assimilation techniques in the CO_2 flux estimation problem. Our goals in this paper are to (1) explain how this assimilation system works, (2) show its accuracy in solving the targeted optimization problem, and (3) test the sensitivity of the system to a few important choices of parameters like the number of ensemble mem-

Table 1. Reference List of Mathematical Symbols Used in This Work, as Well as Their Name, Unit, and Dimensions[a]

Symbol	Name	Unit	Dimension
$\mathbf{x}$	state vector	$kgC/m^2/s$	s
$\mathbf{x}'_i$	state vector deviations	$kgC/m^2/s$	s
$\mathbf{P}$	state covariance matrix	$(kgC/m^2/s)^2$	s × s
$\mathbf{X}$	state deviation matrix	$kgC/m^2/s$	s × N
$\mathbf{y}^\circ$	observation vector	ppm	m
$\mathbf{R}$	observation-error covariance matrix	ppm^2	m × m
$\mathcal{H}$	observation operator	$kgC/m^2/s \rightarrow ppm$	s → m
$\mathbf{H}$	linear observation operator in matrix form	$kgC/m^2/s \rightarrow ppm$	s × m
$\mathcal{M}$	dynamic model	$kgC/m^2/s \rightarrow kgC/m^2/s$	s → s
$\mathbf{M}$	linear dynamic model in matrix form	$kgC/m^2/s \rightarrow kgC/m^2/s$	s → s
$\mathbf{Q}$	dynamical model error matrix	$(kgC/m^2/s)^2$	s × s
η	dynamical model error vector	$kgC/m^2/s$	s
$CO_{2i}(x,y,z,t)$	background CO_2 concentrations	ppm	TM5 grid × N

[a]In this work s = 14,400 (1200 × 12), N = 1500, and m ≈ 50 for each cycle of the assimilation. The TM5 grid has several resolutions because of the two-way nesting. Arrows indicate the mapping from one dimension to another by a matrix or operator.

bers. We will describe the system in detail in section 2 including a brief comparison to related ensemble techniques in NWP, followed by a description of the test problem configuration (section 3). The results of several pseudodata experiments are presented in section 4. In section 5, we will discuss the benefits and weaknesses of our ensemble method and the way SEAT-A will be applied in the near future.

2. Method

[10] Generally, data assimilation systems progress with two distinct steps in one assimilation cycle: (1) the analysis step and (2) the forecast step. The first step can be described as finding the state of a system that is optimally consistent (''optimally'' as yet undefined) with observations, whereas the second step describes the evolution in time of the optimal state to a point in time where new observations are available. At that point, the forecast state serves as the first guess, or ''background'' for the next analysis step. Note that the use of the term ''background'' throughout this work is similar to the use of ''prior'' in most CO_2 related inverse modeling studies. We reserve the word ''prior'' in this work though to refer to fluxes that are created before the inversion and are therefore fixed, whereas our ''background'' fluxes result from the assimilation process and therefore contain information drawn from previous analysis cycles.

[11] State vector analysis in a Bayesian least squares or maximum likelihood framework is common to all inversion methods described in the introduction. We describe the algorithm for state analysis (1) in the ensemble data assimilation system in section 2.1. The forecast step (2) is a new concept for the CO_2 flux estimation problem that was not employed in the traditional batch inversions, nor in the fixed lag Kalman smoother of *Bruhwiler et al.* [2005]. Instead, bottom-up or climatological CO_2 fluxes served as background states for the next cycle. The forecast model plays an important role in data assimilation, as we will explain in section 2.2. The analysis and forecast steps combined form a full data assimilation cycle described in section 2.3. Once the system is initialized (described in section 2.4), the succession of cycles requires little further input to the system besides observations. User intervention through a process called ''covariance localization'' described in section 2.5, can be beneficial though. Section 2.6 presents a brief comparison of SEAT-A and data assimilation systems as employed in NWP methods. A short description of the TM5 model can be found in Appendix A. The notation in this work will follow the suggestions by *Ide et al.* [1997] and is summarized in Table 1.

2.1. State and Covariance Analysis

[12] The starting point of our discussion is the general cost function:

$$\mathrm{J} = (\mathbf{y}^\circ - \mathcal{H}(\mathbf{x}))^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y}^\circ - \mathcal{H}(\mathbf{x})) + \left(\mathbf{x} - \mathbf{x}^{\mathbf{b}}\right)^{\mathrm{T}}\mathbf{P}^{-1}\left(\mathbf{x} - \mathbf{x}^{\mathbf{b}}\right) \quad (1)$$

of a system in which the maximum likelihood solution of unknown variables in state vector $\mathbf{x}$ [dimension s] is found as a balance between information drawn from observations $\mathbf{y}^\circ$ [m] with covariance $\mathbf{R}$ [$m \times m$] and a priori knowledge contained in the background state variable $\mathbf{x}^b$ [m] with covariance $\mathbf{P}$ [$s \times s$]. The observation operator $\mathcal{H}$ samples the state vector $\mathbf{x}$ [s] and returns a vector [m] to be compared to the observations. The state vector $\mathbf{x}$ (and its covariance $\mathbf{P}$) that minimizes J can be shown [*Tarantola*, 2004] to be:

$$\mathbf{x}^{\mathrm{a}}_{\mathrm{t}} = \mathbf{x}^{\mathrm{b}}_{\mathrm{t}} + \mathbf{K}\left(\mathbf{y}^\circ_{\mathrm{t}} - \mathcal{H}\left(\mathbf{x}^{\mathrm{b}}_{\mathrm{t}}\right)\right) \quad (2)$$

$$\mathbf{P}^{\mathrm{a}}_{\mathrm{t}} = (\mathbf{I} - \mathbf{KH})\mathbf{P}^{\mathrm{b}}_{\mathrm{t}} \quad (3)$$

in which t is a subscript for time, superscript b refers to background quantities and a to analyzed ones, $\mathbf{H}$ is the linear(ized) matrix form of the observation operator $\mathcal{H}$, and $\mathbf{K}$ [$s \times m$] is the Kalman gain matrix defined as:

$$\mathbf{K} = \left(\mathbf{P}^{\mathrm{b}}_{\mathrm{t}}\mathbf{H}^{\mathrm{T}}\right)\left(\mathbf{H}\mathbf{P}^{\mathrm{b}}_{\mathrm{t}}\mathbf{H}^{\mathrm{T}} + \mathbf{R}\right)^{-1} \quad (4)$$

[13] In the atmospheric CO_2 inversion we discuss in this work the state vector $\mathbf{x}$ holds unknown surface fluxes [units of $kgC/m^2/s$] to be optimized with atmospheric observations [units of ppm], linked together through operator $\mathcal{H}$ which is an (usually but not necessarily fully linear) atmospheric transport model. This transport model (not to be confused with the forecast model discussed in the next section!) takes an initial distribution of CO_2 concentrations [ppm] and propagates it forward in time using offline stored meteorological wind fields, while altering the CO_2 concentrations at

the surface by the CO_2 fluxes we are trying to optimize. The observation vector $\mathbf{y}^\circ$ contains the observed CO_2 mixing ratios minus a set of atmospheric background CO_2 mixing ratios (from now on denoted as CO_2(x,y,z,t)) as we are only trying to account for changes in mixing ratio since the start of our inversion, not for all of the ~375 ppm of CO_2 found in the atmosphere today.

[14] In batch approaches of atmospheric CO_2 inversions, the subscript t is dropped and the state vector $\mathbf{x}$ includes multiple years of fluxes at once. The background state variable $\mathbf{x}^b$ then holds several years of previously calculated bottom-up flux estimates that are usually referred to as "prior fluxes." The same system was solved in the fixed lag Kalman smoother of *Bruhwiler et al.* [2005], but with only a few months of fluxes in the state variable $\mathbf{x}$, leading to a much more computationally efficient algorithm. The most expensive parts of batch and regular Kalman smoother methods are the precalculation of observation operators ($\mathbf{H}$ matrices in equations (3) and (4)), and solving the covariance analysis equation (3), as matrix $\mathbf{P}$ [$s \times s$] with s $\mathcal{O}(10^5)$ quickly becomes too large even for today's powerful computers.

[15] In an ensemble Kalman filter [see *Evensen*, 1994; *Houtekamer and Mitchell*, 1998], the information in the covariance matrix $\mathbf{P}$ (both background and analyzed) is represented in fewer dimensions N by an ensemble of state vectors $\mathbf{x}_i$ composed of a mean state, and deviations from the mean state:

$$\mathbf{x}_\mathrm{i} = \overline{\mathbf{x}} + \mathbf{x}'_\mathrm{i} \tag{5}$$

The deviations $\mathbf{x}'_i$ are created such that the normalized ensemble of deviations define the columns of a matrix $\mathbf{X}$ [$s \times N$]:

$$\begin{aligned}\mathbf{X} &= \frac{1}{\sqrt{N-1}}\left(\mathbf{x}'_1, \mathbf{x}'_2, \ldots, \mathbf{x}'_N\right)^T \\ &= \frac{1}{\sqrt{N-1}}\left(\mathbf{x}_1 - \overline{\mathbf{x}}, \mathbf{x}_2 - \overline{\mathbf{x}}, \ldots, \mathbf{x}_N - \overline{\mathbf{x}}\right)\end{aligned} \tag{6}$$

which is the square root of the covariance matrix:

$$\mathbf{P} = \mathbf{X}\mathbf{X}^\mathrm{T} \tag{7}$$

In the limit of N $\rightarrow \infty$ this representation of $\mathbf{P}$ is exact, while in an ensemble Kalman filter with a finite number of members $\mathbf{P}$ is approximated. The ensemble of state vectors thus defines the Gaussian probability density function (PDF) of the state vector $\mathbf{x}$ with covariance $\mathbf{P}$. The variance of an individual state vector element is simply calculated from the spread in the corresponding elements in the ensemble. Note that equation (7), together with the factor $1/\sqrt{(N-1)}$ in equation (6), represents an average over all the ensemble members. Vectors $\mathbf{x}'_i$ can easily be created as unconditional realizations of matrix $\mathbf{P}$, for instance through a Cholesky decomposition (see *Michalak et al.* [2004] for an example). We will elaborate on the structure of $\mathbf{P}$ in section 3.

[16] *Whitaker and Hamill* [2002] describe an efficient algorithm to calculate an analyzed ensemble with the correct covariance structure from the background ensemble. They called this the ensemble square root filter (EnSRF), and we follow their formulation for our system. The EnSRF algorithm is particularly efficient when all the available observations in a certain time step of the filter are processed one at a time, which is possible without loss of accuracy when observation errors are uncorrelated (diagonal matrix $\mathbf{R}$). Since this is often (but not necessarily correctly) assumed in the CO_2 problem, we will limit ourselves here to a description of that case and refer to *Whitaker and Hamill* [2002] for the description of an EnSRF with correlated observation errors. Note that for consistency with *Whitaker and Hamill* [2002] and the notation in equations (1)–(7), we denote all ensemble derived quantities in equations (8)–(13) as matrices even though their size reduces them to vectors or scalar values.

[17] In the sequential EnSRF algorithm, the batch of observations belonging to one time step of the filter are processed one at a time which reduces the size of the Kalman gain matrix $\mathbf{K}$ in each sequential analysis step to [s $\times$ 1], a vector the size of the number of unknowns. The Kalman gain matrix is calculated from the ensemble of state vectors and equation (4) using the approximations:

$$\begin{aligned}\mathbf{H}\mathbf{P}\mathbf{H}^\mathrm{T} \approx \frac{1}{N-1}&\left(\mathcal{H}\left(\mathbf{x}'_1\right), \mathcal{H}\left(\mathbf{x}'_2\right), \ldots, \mathcal{H}\left(\mathbf{x}'_N\right)\right) \\ &\cdot \left(\mathcal{H}\left(\mathbf{x}'_1\right), \mathcal{H}\left(\mathbf{x}'_2\right), \ldots, \mathcal{H}\left(\mathbf{x}'_N\right)\right)^T\end{aligned} \tag{8}$$

$$\mathbf{P}\mathbf{H}^T \approx \frac{1}{N-1}\left(\mathbf{x}'_1, \mathbf{x}'_2, \ldots, \mathbf{x}'_N\right)\left(\mathcal{H}\left(\mathbf{x}'_1\right), \mathcal{H}\left(\mathbf{x}'_2\right), \ldots, \mathcal{H}\left(\mathbf{x}'_N\right)\right)^T \tag{9}$$

Where each entry N denotes one column of ensemble state vectors or ensemble modeled CO_2 values as in equation (6). In the case of one observation, equation (8) thus simply describes a "dot product" of two vectors and $\mathbf{H}\mathbf{P}\mathbf{H}^T$ becomes a [1 $\times$ 1] scalar value, while $\mathbf{P}\mathbf{H}^T$ is a [s $\times$ 1] vector. Through equations (8)–(9), the Kalman gain matrix $\mathbf{K}$ linearly maps observed quantities to state vector elements as an average over all the ensemble members.

[18] The Kalman gain matrix is used to update the mean state vector with equation (2), whereas the deviations from the mean state vector are updated independently using:

$$\mathbf{x}'^\mathrm{a}_\mathrm{i} = \mathbf{x}'^\mathrm{b}_\mathrm{i} - \tilde{\mathbf{k}}\mathcal{H}\left(\mathbf{x}'^\mathrm{b}_\mathrm{i}\right) \tag{10}$$

Where the [s $\times$ 1] vector $\tilde{\mathbf{k}}$ [s] is related to the [s $\times$ 1] Kalman gain matrix $\mathbf{K}$ by a scalar quantity α calculated as:

$$\begin{aligned}\tilde{\mathbf{k}} &= \mathbf{K} \cdot \alpha \\ \alpha &= \left(1 + \sqrt{\frac{\mathbf{R}}{\mathbf{H}\mathbf{P}^b\mathbf{H}^T + \mathbf{R}}}\right)^{-1}\end{aligned} \tag{11}$$

This independent update of the state and its ensemble of deviations prevents systematic underestimation of $\mathbf{P}^a$ that was shown to occur previously when perturbed observations were used to update the ensemble deviations [*Whitaker and Hamill*, 2002]. The calculation of α requires the evaluation of [1 $\times$ 1] scalars $\mathbf{R}$ and $\mathbf{H}\mathbf{P}\mathbf{H}^T$ only.

[19] The analyzed mean and ensemble state from one observation will serve as the background state for the next until all observations are processed. They will also go into the calculation of the next observations' Kalman gain matrix through equations (8)–(9). Before calculating the

next Kalman gain matrix though, we also need to update the ensemble of sampled CO_2 concentrations $\mathcal{H}(\mathbf{x}_i'^b)$ to reflect the new information in the updated state vectors. One way to accomplish this is to simply reapply operator $\mathcal{H}$ to the new ensemble of background state vectors. However, this would be very expensive in our problem since we would have to rerun our transport model for each observation. Therefore we update the sampled CO_2 concentration vector in a way similar to the state vector update, using the ensemble averaged information in the Kalman gain matrix. Each modeled CO_2 concentration corresponding to an observation m that has yet to be assimilated (denoted $\mathcal{H}(\mathbf{x_t})_m$ here) is updated using the equation:

$$\mathcal{H}\left(\mathbf{x}_\mathbf{t}^\mathbf{a}\right)_\mathbf{m} = \mathcal{H}\left(\mathbf{x}_\mathbf{t}^\mathbf{b}\right)_\mathbf{m} + \mathbf{H_m K}\left(\mathbf{y}_\mathrm{t}^\circ - \mathcal{H}\left(\mathbf{x}_\mathrm{t}^\mathrm{b}\right)\right) \tag{12}$$

whereas the deviations are updated using:

$$\mathcal{H}\left(\mathbf{x}_\mathbf{i}'^\mathbf{a}\right)_\mathbf{m} = \mathcal{H}\left(\mathbf{x}_\mathbf{i}'^\mathbf{b}\right)_\mathbf{m} - \mathbf{H_m}\tilde{\mathbf{k}}\mathcal{H}\left(\mathbf{x}_\mathrm{i}'^\mathrm{b}\right) \tag{13}$$

Where we have replaced the operator $\mathcal{H}_m$ with its matrix equivalent $\mathbf{H}_m$ in the right hand side term. In these equations only the term $\mathbf{H_m K}$ has to be calculated, which is easily accomplished realizing (from equation (4)) that this multiplication has a term $\mathbf{H_m P^b H^T}$ in its numerator. This is again a scalar value calculated from equation (8) where the first right hand side term contains an ensemble of modeled CO_2 concentration yet to be optimized while the second right hand side term refers to a model ensemble of the CO_2 observation currently being optimized. Note that these equations are analogous to equations (2) and (10) except for the operator $\mathcal{H}$ (or its matrix equivalent $\mathbf{H}$) in each term. After the update of the ensemble of modeled CO_2 values the algorithm continues with the next observation until all observations are processed to reach the final analyzed ensemble. It is important to note that in all the analysis equations, we have replaced each occurrence of the linear matrix $\mathbf{H}$ either by the full operator $\mathcal{H}$ working on each ensemble member, or a quantity that can be derived from the ensemble. This renders the linear matrix $\mathbf{H}$ obsolete in our implementation.

[20] So what are the implications for the CO_2 flux estimation problem? If we create an ensemble of N CO_2 flux fields that has a mean $\overline{\mathbf{x}}$ and spans the covariance structure $\mathbf{P}$, we are able to find optimized fluxes using a set of CO_2 observations with covariance $\mathbf{R}$ simply by running an atmospheric tracer transport model (operator $\mathcal{H}$) forward N times and sampling it consistently with the observations to create first $\mathcal{H}(\overline{\mathbf{x}})$ and $\mathcal{H}(\mathbf{x}_i)$, then $\mathbf{PH}^T$, $\mathbf{HPH}^T$, $\mathbf{K}$, and finally α, $\tilde{\mathbf{k}}$, $\mathbf{x}^a$, $\mathbf{x}_i'^a$, and $\mathbf{P}^a$. Thus we can solve the analysis equations without the need to precalculate $\mathbf{H}$ (base functions) and without explicitly forming and inverting the large covariance matrix $\mathbf{P}$. This is the property of an Ensemble Kalman Filter that allows large state vectors to be optimized without losing the ability to calculate its covariance, and to assimilate a large number of observations without having to precalculate observation operators. Moreover, observation operator $\mathcal{H}$ can be a fully nonlinear forward calculation that includes chemistry, making these equations suitable to solve for fluxes of, for instance, CO or CH_4. We note though that this framework still assumes Gaussian errors on the observations and background state vector and returns an analyzed state vector with a Gaussian error. These Gaussian assumptions might not hold if nonlinear processes (e.g., chemistry) are involved. The sequential EnSRF algorithm achieves further efficiency by operating only on vector and scalar quantities and not on matrices making its implementation particularly simple.

2.2. State and Covariance Forecast Model

[21] As mentioned, an important role in data assimilation is played by the so-called dynamical model (denoted $\mathcal{M}$). This model describes the evolution of the state vector in time and thus provides a first guess of the state vector before new observations are introduced to the system:

$$\mathbf{x}_{\mathrm{t+1}}^\mathrm{b} = \mathcal{M}\left(\mathbf{x}_\mathrm{t}^\mathrm{a}\right) \tag{14}$$

The same dynamical model can be used to forecast covariances in time:

$$\mathbf{P}_{\mathrm{t+1}}^\mathrm{b} = \mathbf{M}\mathbf{P}_\mathrm{t}^\mathrm{a}\mathbf{M}^\mathrm{T} + \mathbf{Q} \tag{15}$$

where $\mathbf{Q}$ represents an increase in state covariance introduced into the forecast by the imperfect dynamical model, and $\mathbf{M}$ is the linear(ized) matrix form of operator $\mathcal{M}$. In an ensemble framework equation (15) is not used. Instead the covariance $\mathbf{P}$ is forecasted through the individual ensemble members as:

$$\mathbf{x}_{\mathrm{i,t+1}}^\mathrm{b} = \mathcal{M}\left(\mathbf{x}_\mathrm{i,t}^\mathrm{a}\right) + \eta \tag{16}$$

Where η represents a random vector of forecast errors with the structure of $\mathbf{Q}$ to be added to the new background state. Through equations (14) and (16), current estimates depend on all previously optimized state vectors introducing a coupling between past and present state and covariance. Also, information on the state vector and its covariance structure derived from previous observations are propagated into the next estimate which gives the system a "learning" ability. Covariance structures are thus derived from the observations and the dynamical model, including its dynamical model errors. The model error serves an important purpose: it maintains spread in the ensemble and prevents it from converging to unrealistically small values after repeated exposure to observations. Dynamical model errors can be introduced explicitly as a vector of perturbations on the state as in (16), or stochastically by varying key parameters in the model $\mathcal{M}$. Nonlinear dynamical models such as a weather forecast model or ocean GCM usually have error growth intrinsic in the nonlinear physics causing the spread of the ensemble (covariances) to increase during a state forecast.

[22] An important realization for the CO_2 flux estimation problem as formulated here is that our state vector does not hold a dynamical variable in the sense that future CO_2 fluxes do not normally depend on our analysis of current CO_2 fluxes. Compare that to a 3D atmospheric temperature field, where tomorrow's temperature distribution will depend very strongly on the analyzed temperature field through atmospheric physics. CO_2 fluxes should thus be viewed as the system forcing (or boundary condition)

rather than system variables. The time evolution of this forcing is not readily captured in a state-dependent dynamical model. State dependence in the dynamical model is needed to couple the forecast to the analysis and thus propagate information. We refer to section 5 for further discussion of the lack of a state-dependent dynamical model for the fluxes.

[23] In absence of a suitable dynamical model we couple forecasted CO_2 fluxes to analyzed CO_2 fluxes through a simple form of persistence forecasting:

$$\mathcal{M} = \mathbf{I} \tag{17}$$

where $\mathbf{I}$ is the identity matrix. This means that we assume the background CO_2 fluxes for one time step to equal the once optimized fluxes of the previous time step. Obviously this is a poor model in the sense that it cannot propagate or add information to the system. We should therefore allow the analysis to deviate substantially from our first guess, and add a large error η to the new background state. Also, persistence forecasting does not contain any intrinsic or stochastic error growth and can therefore not be relied upon to balance decreasing covariances making the task of choosing suitable errors η even more complicated. As a result we find that it is better to only use (14) to propagate the mean of the state, and prescribe its covariance structure at each new step (we will explain the chosen covariance structure for this study in section 3). This has the advantage that we can still forecast our own background mean state and not depend on precalculated prior flux products, whereas we do not need to worry about the difficult task of creating a model $\mathcal{M}$ with associated errors $\mathbf{Q}$ to model uncertainty in the system. A disadvantage is that we lose the ''learning'' ability of the filter and need to start each new flux estimate with reasonably large uncertainties everywhere. In this mode of operation our filter resembles a 3d variational technique which also lacks the dynamic coupling between analyzed and background covariances.

2.3. Assimilation Cycle

[24] SEAT-A is a combination of observations with the above persistence forecast, the state analysis equations, and a tracer transport model. The first important aspect of this system is that the state vector contains flux estimates for multiple time steps, each corresponding to a one week mean. This is indicated by the system's ''lag'' which can span a long time window for CO_2 because the only process that slowly erases source signatures is atmospheric diffusion. In other words, the relationship between the state vector x and observations $\mathbf{y}^\circ$ (described by operator $\mathcal{H}$) spans several months. *Bruhwiler et al.* [2005] found that for monthly flux inversions, observations had to be linked to 6–9 months of past fluxes to reliably retrieve CO_2 fluxes. For our current problem we will estimate weekly fluxes, where tests (see section 4.2) indicate that 8–10 weeks of lag still captures most of the spatial information in the observations. We speculate that this shorter lag is possible because we use instantaneous CO_2 signals from weekly fluxes that hold more spatial information, and are more quickly replaced by new flux signals from subsequent weeks than the smoother monthly mean CO_2 signals that were used previously in monthly mean flux inversions.

[25] The time stepping in the assimilation scheme is illustrated in Figure 1. In this example, twelve weeks of lag in the state vector are indicated by $\mathbf{x}_i(0,\ldots,11)$, in which the number in parentheses denotes how many times a particular week of fluxes has been estimated previously on the basis of different observations from previous cycles. Each shaded box represents an ensemble [i = 1,..,N] of global surface fluxes [*s*]. Light shaded boxes denote background fluxes whereas dark shaded boxes denote posterior fluxes. A cycle of SEAT-A proceeds as follows: (1) We run the TM5 model forward from the background concentration fields in $CO_{2i}(x,y,z,t)$ to $CO_{2i}(x,y,z,t+12)$ forced by the fluxes in $x_i(0,\ldots,11)$, and extract CO_2 mixing ratios at the observation times and locations. This allows us to construct an ensemble of modeled CO_2 at each site. (2) Equations (2) and (10) are solved to give an analyzed ensemble of fluxes for each element of the state vector and each week; (3) the ensemble of final fluxes in $\mathbf{x}_i^a(12)$ will no longer be estimated in the next cycle and are therefore incorporated into $CO_{2i}(x,y,z,t+1)$ by running the TM5 model one week forward starting from $CO_{2i}(x,y,z,t)$ forced with the final ensemble fluxes $\mathbf{x}_i(12)$. (4) Each analyzed state vector becomes the background state vector for the next cycle. A new background mean flux is created to go into $\overline{\mathbf{x}}(0)$ by propagation with model $\mathcal{M}$ (equation (14)), (5) we draw a new ensemble of N flux deviations $\mathbf{x}'_i(0)$ from the specified background covariance structure to represent the Gaussian PDF around the new mean flux $\overline{\mathbf{x}}(0)$, and finally (6) new observations $\mathbf{y}^\circ$ are read and the next cycle starts.

[26] The TM5 chemistry transport model serves two purposes in SEAT-A: (1) It is used to sample the state vector and return predicted CO_2 concentrations (TM5 is operator $\mathcal{H}$), and (2) it carries the ''memory'' of all optimized fluxes occurring previous to those currently in our state vector in the form of the 3D field of CO_2 concentrations ($CO_{2i}(x,y,z,t)$). Note that one cycle of SEAT-A thus requires (nlag + 1) weeks of simulation with TM5 for each ensemble member. Detailed descriptions of the TM5 model are given by *Peters et al.* [2004] and *Krol et al.* [2005]. In Appendix A we describe two important aspects of the model: the two-way nested horizontal grid definition, and parallel operation of the model.

2.4. System Initialization

[27] At t = 0 the ensemble of background concentration fields $CO_{2i}(x,y,z,t)$ and the ensemble of fluxes $\mathbf{x}_i(0,..,11)$ need to be initialized. SEAT-A is initialized with a mean flux estimate in the initial state vector based on the CASA biosphere model [*Randerson et al.*, 1997] for land fluxes, and the *Takahashi et al.* [2002] ocean fluxes. The ensemble members are created by making N state vectors using unconditional realizations following the prescribed covariance structure $\mathbf{P}^b$ (see section 3 for a description of $\mathbf{P}^b$), and adding these to the central estimate $\overline{x}$. The initial concentration fields $CO_{2i}(x,y,z,0)$ for the first step are taken from a previous CO_2 simulation with TM5 and were not optimized in this work even though the influence of initial concentrations on flux estimates is well known [*Bousquet et al.*, 2000; *McKinley et al.*, 2004; *Peylin et al.*, 2005]. In real

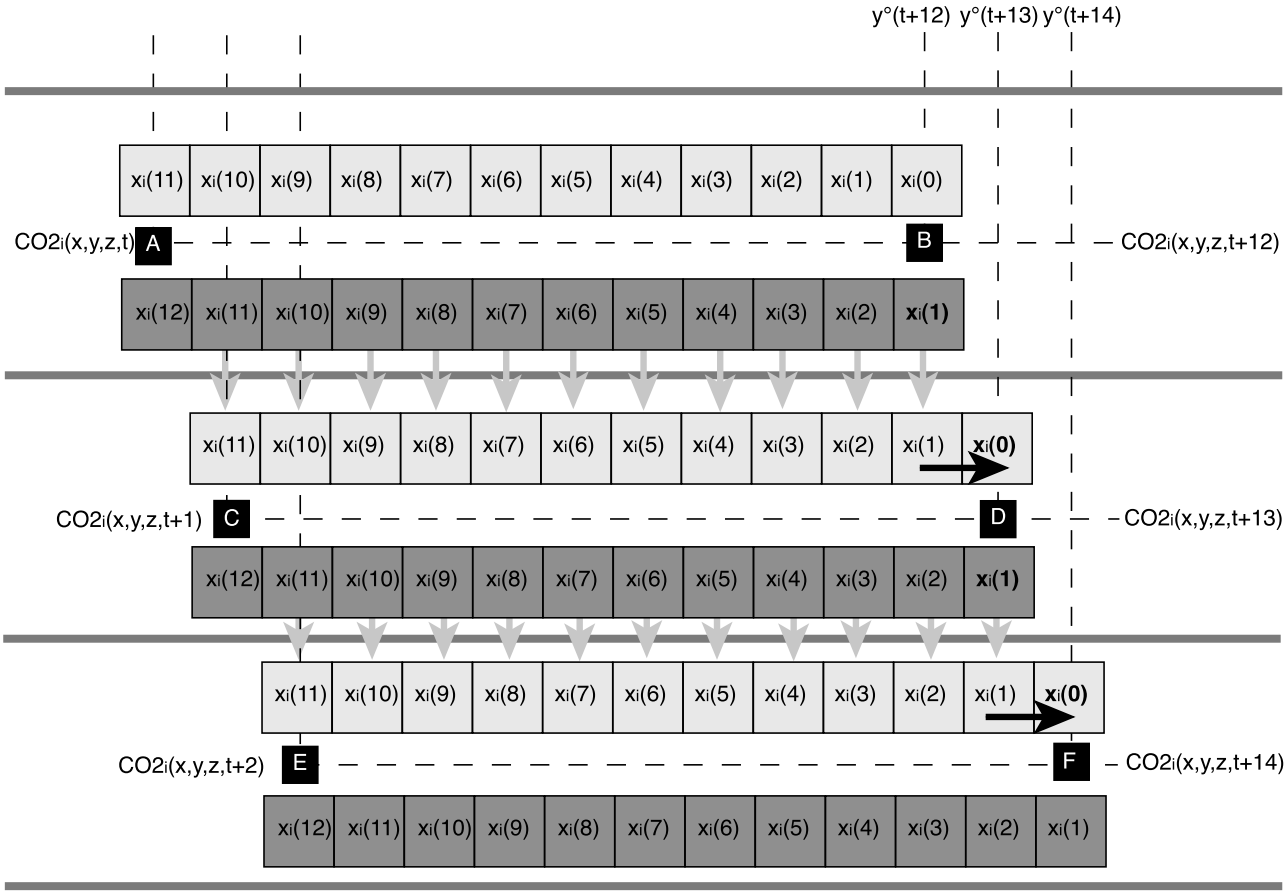

Figure 1. Illustration of three cycles in SEAT-A when 12 weeks of fluxes compose the state vector. Light shaded boxes denote the background fluxes, and dark shaded boxes denote posterior fluxes. Each box represents N ensemble members each with [s] surface fluxes around the globe. The number in parentheses indicates how many times a week of fluxes has been estimated previously on the basis of observations from past cycles, and the subscript i refers to an individual ensemble member. The scheme proceeds as follows: (1) N instances of TM5 are run from point A to B with $CO_{2i}(x,y,z,t)$ as starting concentration, forced by the fluxes $\mathbf{x}_i(11)$ to $\mathbf{x}_i(0)$. (2) The resulting concentrations at $CO_{2i}(x,y,z,t+12)$ are compared to observations $\mathbf{y}^\circ(t=12)$, and the state vector is optimized. (3) The final flux estimate in $\mathbf{x}_i(12)$ is incorporated into the background concentration by running TM5 from A to C starting from $CO_{2i}(x,y,z,t)$ forced with fluxes $\mathbf{x}_i(12)$. (4) Posterior fluxes from step t become background fluxes for step t + 1 indicated by shaded vertical arrows. The background mean flux for the new $\mathbf{x}(0)$ is forecasted from $\mathbf{x}(1)$ with the dynamical model $\mathcal{M}$ (solid horizontal arrow), and (5) new ensemble members are drawn from a specified background covariance to form the Gaussian PDF around $\mathbf{x}(0)$; the other fluxes $\mathbf{x}_i(11)$ to $\mathbf{x}_i(1)$ remain the same. (6) A new cycle starts with new observations.

applications of SEAT-A initial conditions will be accounted for as part of the "inversion" parameters.

[28] Once SEAT-A is initialized, all future background state vectors are created using the state propagation model $\mathcal{M}$, with no other information than CO_2 observations and the covariances $\mathbf{P}^b$ determining the analyzed solution. Because of the influence of the initial condition on the initial results of the filter we suggest to discard the first months of flux estimates if one wants a result independent of the background flux information. The number of time steps still influenced by the chosen initial condition is difficult to assess, but is on the order of two times the lag time of the filter as the first nlag weeks get an assigned flux, and their direct influence last another nlag weeks. It could be argued however that reasonable initial values will not influence the first months to such an extent that their results are useless.

2.5. Covariance Localization

[29] Covariance localization is used to keep the covariance structure of the ensemble system well behaved [*Houtekamer and Mitchell*, 1998]. Since the number of ensemble members N is finite, the representation of $\mathbf{P}$ in N-dimensional space is not perfect resulting in a varying number of off-diagonal covariance matrix values that do not truly describe coherent behavior of flux means, but rather statistical noise of the ensemble. It can be argued that the performance of the system improves if such noise in the covariances is suppressed in some way. This can be achieved by covariance localization, which is done by

calculating the Schur product (i.e., piecewise multiplication) $\mathbf{P} = \mathbf{P} \cdot \mathbf{L}$ where $\mathbf{L}$ is an $[s \times s]$ matrix. Usually, $\mathbf{L}$ describes the exponential decay of the covariance structure with distance between state variables [*Gaspari and Cohn*, 1999]:

$$\mathbf{L}_{ij} = e^{-d_{ij}/l} \tag{18}$$

Where d_{ij} is the great circle distance between state vector element i and j and l is a specified length scale. The optimal length scale increases as more ensemble members are used, and too strong localization can degrade the quality of the analysis [*Houtekamer and Mitchell*, 1998].

[30] We apply $\mathbf{L}$ only to spatial correlations within each week by multiplying each observation's Kalman gain vector $\mathbf{k}$ $[s]$ with a selected row i of matrix $\mathbf{L}$. This decreases the magnitude of state updates as a function of distance d_{ij} to the state vector element i that is maximally constrained by an individual observation. We have tested the performance of the filter for different decay lengths (section 4.1). The assumption that flux covariances are small for regions separated by large distances is physically credible, as processes controlling the spatial flux patterns (such as temperature, soil moisture, sunlight) covary on the local scale, but rarely globally. It was found for NWP [*Houtekamer and Mitchell*, 1998; *Whitaker and Hamill*, 2002] and also in this study that covariance localization strongly reduces the number of ensemble members needed to get satisfactory performance in the assimilation system.

2.6. Comparison to NWP

[31] Finally, we briefly discuss parallels between NWP ensemble assimilation systems and SEAT-A. First of all, in NWP the state vector $\mathbf{x}$ contains no lag, but simply represents one state at one time. Variables in the state vector are the prognostic variables of a weather forecast model such as temperature, humidity, and wind speed that will serve as initial conditions for a weather forecast. Because those are often the same quantities that are measured, observation operators $\mathcal{H}$ usually simply perform sampling in 3D state space, although $\mathcal{H}$ can be as complex as a radiative transport model to assimilate satellite radiances. State propagation ($\mathcal{M}$) is achieved with the full nonlinear weather forecast model that takes an analyzed state vector as the initial condition to make a weather forecast several days into the future (state at t = t + 1,2,3,). The physics of the atmosphere contained in the formulation of such models allows propagation and even addition of information from one time step to the next, a skill that renders covariance propagation feasible for NWP in contrast to the work we present here.

[32] In NWP, covariances are thus derived from the observations and serve mainly to maintain the structure of organized weather patterns (such as low-pressure areas or fronts) during assimilation, as well as to determine uncertainty in the initial conditions for the forecast. In the CO_2 problem, covariances have been used mostly as a formal quantitative estimate of flux uncertainty, a practice that could be questioned (see section 5). NWP assimilation systems estimate $\mathcal{O}(10^7)$ parameters using ensembles of less than 100 members. Part of this efficiency is possible because of the short memory of the atmosphere for weather patterns: nonlinear and chaotic behavior erase the imprint of organized weather systems within several days.

3. Assimilation System Tests

[33] We have performed several tests to assess the performance of our ensemble Kalman filter for the CO_2 problem. These tests all focused on the "engineering" parameters of the method such as the number of ensemble members needed, the optimal strength of localization, and the number of weeks of lag in the assimilation system. "Inversion" parameters more specific to the flux retrieval problem itself (such as model-data mismatch, flux uncertainty, number of regions, covariance length scale, initial conditions, and choice of observation network) are commonly also varied in real applications, but were not varied in this work. These latter parameters will strongly depend on the particular problem to be solved and do not necessarily relate to the technical performance of the assimilation method that we intend to demonstrate in this work. There is however a dependence of the "engineering" parameters on the "inversion" parameters and in addition to the customary variation of the inversion parameters, sensitivity tests of the engineering parameters will need to be performed for future applications of the filter.

[34] In the problem for which we chose to demonstrate this technique, SEAT-A retrieves CO_2 fluxes for each week of the year 2000 on a regular 9×6 degree global grid (1200 unknowns per week), constrained by pseudo-observations sampled from known fluxes, while mimicking the actual sampling from the year 2000 CMDL Cooperative Air sampling network. A list of sites is included in Table 2. Since not all CMDL sites successfully take a sample each week, the configuration of the network changes from week to week. Although this study does not yet include continuous observations, several sites (Ascension, Baltic, Guam, Station M, Zeppelin) already collect flask samples at a frequency exceeding once per week. Including aircraft measurements from several locations (12 flasks per flight, indicated by a variable height in Table 2) a total of 2460 observations are assimilated in this work. This stresses our point that even for a moderately sized problem, precalculating the observation operators is a daunting task. Pseudo observations were created with a set of known fluxes on the basis of *Takahashi et al.* [2002] for ocean regions, and Simple Biosphere model [*Denning et al.*, 2003] V3.0 fluxes over land regions. Fluxes were averaged to weekly means to ensure that no differences between true fluxes and retrieved fluxes would arise because of the estimation of weekly average parameters by our method. This also means that the system exhibits "perfect" transport further isolating the effect of the assimilation technique in our tests. The ability of SEAT-A to reproduce the true fluxes thus depends on (1) the capacity to "see" the true fluxes with the year 2000 NOAA-CMDL cooperative air sampling network, (2) the limited ability of the state propagation model to set proper background fluxes, and (3) limitations in our assimilation method related to imperfect covariances, localization, or assimilation window length.

[35] The inversions presented were performed with TM5 in a configuration with two nested grids over North

Table 2. Sites Used in This Study, Their Observation Frequency, and Assigned Model-Data Mismatch[a]

Code	Number of Observations	Longitude, °	Latitude, °	Altitude, m	Name	$\sqrt{R}$, ppm
ALT	53	82.45	−62.52	210	Alert, Nunavut, Canada	0.50
ASC	76	−7.92	−14.42	54	Ascension Island, United Kingdom	0.50
ASK	40	23.18	5.42	2728	Assekrem, Algeria	1.50
AZR	47	38.77	−27.38	40	Terceira Island, Azores, Portugal	0.50
BAL	95	55.42	17.07	28	Baltic Sea, Poland	7.50
BME	43	32.37	−64.65	30	St. Davids Head, Bermuda, United Kingdom	0.50
BMW	44	32.27	−64.88	30	Tudor Hill, Bermuda, United Kingdom	0.50
BRW	48	71.32	−156.60	11	Barrow, Alaska, United States	0.50
BSC	50	44.17	28.68	3	Black Sea, Constanta, Romania	7.50
CAR	129	40.90	−104.80	Variable	Carr, Colorado, United States	1.00
CBA	27	55.20	−162.72	25	Cold Bay, Alaska, United States	3.00
CGO	45	−40.68	144.68	94	Cape Grim, Tasmania, Australia	3.00
CHR	37	1.70	−157.17	3	Christmas Island, Republic of Kiribati	0.50
CRZ	34	−46.45	51.85	120	Crozet Island, France	0.125
EIC	35	−27.15	−109.45	50	Easter Island, Chile	0.125
GMI	82	13.43	144.78	6	Mariana Islands, Guam	0.50
HAA*	56	21.23	−158.95	Variable	Molokai Island, Hawaii, United States	1.0
HBA	48	−75.58	−26.50	33	Halley Station, Antarctica, United Kingdom	0.125
HFM*	54	42.54	−72.17	Variable	Harvard Forest, Massachusetts, United States	1.00
HUN	51	46.95	16.65	344	Hegyhatsal, Hungary	7.50
ICE	44	63.34	−20.29	127	Storhofdi, Vestmannaeyjar, Iceland	0.50
IZO	37	28.30	−16.48	2360	Tenerife, Canary Islands, Spain	1.50
KEY	42	25.67	−80.20	3	Key Biscayne, Florida, United States	7.50
KUM	53	19.52	−154.82	3	Cape Kumukahi, Hawaii, United States	0.50
KZD	53	44.45	75.57	412	Sary Taukum, Kazakhstan	3.00
KZM	49	43.25	77.88	2519	Plateau Assy, Kazakhstan	3.00
LEF	5	45.93	−90.27	868	Park Falls, Wisconsin, United States	5.00
LEF	30	45.93	−90.27	Variable	Park Falls, Wisconsin, United States	2.00
MHD	48	53.33	−9.90	25	Mace Head, County Galway, Ireland	3.00
MID	47	28.21	−177.38	7.7	Sand Island, Midway, United States	0.50
MLO	49	19.53	−155.58	3397	Mauna Loa, Hawaii, United States	0.50
NMB	12	−23.58	15.03	461	Gobabeb, Namibia	1.50
NWR	46	40.05	−105.58	3475	Niwot Ridge, Colorado, United States	1.50
PFA*	75	65.07	−147.29	Variable	Poker Flat, Alaska, United States	1.00
PSA	51	−64.92	−64.00	10	Palmer Station, Antarctica, United States	0.125
PTA*	22	38.95	−123.73	17	Point Arena, California, United States	3.00
RPB	42	13.17	−59.43	45	Ragged Point, Barbados	0.50
RTA*	40	−21.25	−159.83	Variable	Rarotonga, Cook Islands	1.00
SEY	37	−4.67	55.17	7	Mahe Island, Seychelles	0.50
SHM	45	52.72	174.10	40	Shemya Island, Alaska, United States	0.50
SMO	50	−14.24	−170.57	42	Tutuila, American Samoa	0.50
SPO	46	−89.98	−24.80	2810	South Pole, Antarctica, United States	0.125
STM	93	66.00	2.00	5	Ocean Station M, Norway	3.00
SUM	30	72.58	−38.48	3238	Summit, Greenland	1.50
SYO	24	−69.00	39.58	14	Syowa Station, Antarctica, Japan	0.125
TDF	27	−54.87	−68.48	20	Tierra Del Fuego, La Redonda Isla, Argentina	0.50
UTA	41	39.90	−113.72	1320	Wendover, Utah, United States	3.00
UUM	48	44.45	111.10	914	Ulaan Uul, Mongolia	3.00
WIS	54	31.13	34.88	400	Sede Boker, Negev Desert, Israel	7.50
WKT*	8	31.32	−97.33	251	Moody, Texas, United States	5.00
WLG	24	36.29	100.90	3810	Mt. Waliguan, Peoples Republic of China	1.50
ZEP	94	78.90	11.88	475	Ny-Alesund, Svalbard, Norway and Sweden	0.50

[a]Sampling at sites with an asterisk was started after 1999.

America. Sites in the NOAA CMDL air sampling network were divided into six categories, each with their own assigned model-data mismatch value. The categories and respective model-data mismatches [ppm] are: Antarctic sites (0.125), marine boundary layer (0.50), land sites (3.0), mountain sites (1.50), aircraft samples (1.0), tower sites (5.0), and difficult sites (7.5). These values represent subjective choices and are not based on an optimization or analysis of representation errors in our model. A list of sites and their assigned category is found in Table 2. Note that we did not add noise to the pseudo-observations before feeding them back into SEAT-A. We chose not to perturb the observations because we never use the true fluxes as background, and only see the true fluxes through a limited network. This will already prevent us from fully reproducing the truth even without perturbed observations.

[36] A priori information for SEAT-A comes in two forms (1) the background mean fluxes persisted from the previous cycle and (2) a prescribed background covariance structure. The background covariances $\mathbf{P}$ at each time step are prescribed to decay isotropically with distance as in this 4×4 example covariance matrix with 2 land (subscript l) and 2 ocean fluxes (subscript o):

$$\mathbf{P}^{\mathrm{b}} = \begin{pmatrix} \sigma_l & \sigma_l \cdot e^{-d/L_l} & 0 & 0 \\ \sigma_l \cdot e^{-d/L_l} & \sigma_l & 0 & 0 \\ 0 & 0 & \sigma_o & \sigma_o \cdot e^{-d/L_o} \\ 0 & 0 & \sigma_o \cdot e^{-d/L_o} & \sigma_o \end{pmatrix} \quad (19)$$

where σ_l is the variance on terrestrial fluxes equaling 1.0×10^{-16} [kgC/m^2/s]^2, and σ_o is the variance on ocean fluxes equal to 1.0×10^{-18} [kgC/m^2/s]^2 and d is the great circle distance between the center of two regions. The length scales for exponential decay of the covariations were chosen as L_l = 900 km and L_o = 2000 km. The χ^2 on the fluxes calculated from the ensemble of posterior fluxes has a mean of 0.89 indicating that the chosen length scales and $\sigma_{l,o}$ values were not optimal given the values in matrix **R**. We could probably have used slightly smaller length scales or lower variance parameters $\sigma_{l,o}$. Using a covariance structure that leads to the optimal value of $\chi^2 = 1$ [see, e.g., *Michalak et al.*, 2005] is not crucial for the work presented here and will be saved for a future application of SEAT-A. We confirmed that the distribution of χ^2 on the ensemble of posterior fluxes and observations followed a normal distribution indicating proper operation of SEAT-A.

[37] Fluxes used to produce pseudodata include fossil fuel CO_2 emissions taken from the CDIAC estimates for 1995 [*Brenkert*, 1998]. In the assimilation, the same fossil fuel flux was incorporated into the background field CO_{2i}(x,y,z,t) and thus presubtracted from the observations. In one experiment we perturbed the fossil fuel fluxes randomly by an arbitrary 10% in each grid box, separately for each ensemble member thus including fossil fuel uncertainty stochastically in the filter. This approach will also be used in real applications but is not discussed further here as it does not inform us on the assimilation system performance. Other uncertainties such as boundary layer mixing strength or convective overturning could be added stochastically in the same way, adding to the spread of the ensemble and thus to the posterior covariances in a more realistic manner than simply adding a constant transport model error term to the observation covariance matrix **R**.

[38] The quality of several runs is assessed through two statistics: (1) the root-mean-square of the difference between all true (superscript *t*) and analyzed fluxes:

$$\mathrm{RMS} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{x}^{t} - \mathbf{x}^{a}\right)^2} \tag{20}$$

and (2) the number of degrees of freedom (d.o.f.) in the final covariance estimate measured as [*Patil et al.*, 2001]:

$$\mathrm{d.o.f.} = \frac{\left(\sum_{i=1}^{N}\omega_i\right)^2}{\sum_{i=1}^{N}\omega_i^2} \tag{21}$$

where ω_i are the singular values obtained from a singular value decomposition of matrix **X**. The d.o.f will initially increase with increasing number of ensemble members N, until added columns of **X** yield singular vectors close to zero. At that point, the columns of matrix **X** span the complete covariance structure and adding more ensemble members does not add more information to the matrix. We present the d.o.f only for one randomly chosen week of our final results, as we found that it varied only slightly from week to week.

[39] Although we realize that many other tests could be performed with pseudodata experiments we chose this particular setup because it illustrates the power of the ensemble method in a realistic but controlled setting. We could have created tests with perfect data coverage, or very few degrees of freedom, or near infinite ensemble members, or nonlinear tracers. However, the extensive literature on ensemble Kalman filtering in the fields of NWP and ocean modeling already covers those situations in much more detail than we can address here. This includes comparisons of Kalman filter methods to batch methods [*Bruhwiler et al.*, 2005] and 4d-var [*Lorenc*, 2003]. Our pseudodata experiment is meant to show that the CO_2 problem falls within the wide range of applications for which ensemble methods have been shown to work well and offer clear benefits.

4. Results

[40] Starting our discussion of the results at the large scale, Figure 2 shows the annual mean for "true" fluxes and those recreated with our assimilation system after aggregation of the results to 22 "superregions" corresponding to the TransCom 3 regions given by *Gurney et al.* [2002]. We show these annual means for an ensemble size of 1500 members as well as for a 200 member ensemble where localization was applied. At the continental scales the method is obviously able to reproduce the true flux well with differences always within the posterior uncertainty and largest differences occurring in the poorly observed tropical regions. Total ocean and land fluxes are both overestimated slightly mostly because of misallocation of fluxes in Northern Africa, the Temperate Pacific Ocean, and Indian Ocean. The inability to properly separate tropical land and ocean fluxes on the basis of current observations is a well-known feature of inversions [*Bousquet et al.*, 2000]. Seasonal cycles for these superregions based on weekly flux estimates are shown in Figure 3 for three reasonably well constrained regions with large seasonal flux variations (Temperate North America, Europe, Boreal Eurasia), as well as for the Southern Ocean since it dominates uptake of CO_2 in the Southern Hemisphere. The ability of our assimilation system to represent seasonal peak-to-trough amplitudes as large as 15 PgC/yr without bottom-up fluxes to guide the solution is very promising and demonstrates its power to extract information from the observations. Without data assimilation, the fluxes would have remained equal to the initial values at t = 0 for the rest of the year. Further aggregating these weekly fluxes to monthly means (not shown) reduces some spurious temporal variations introduced by the inhomogeneous sampling network and further reduces the differences.

[41] Figure 4 shows that the assimilation contains finer details of the true flux field as well. Maps of monthly mean flux patterns were made for July and November with ocean and land fluxes on separate panels to bring out the details more clearly. In July, major features of the true land biosphere flux distribution are strong uptake in the boreal regions and net CO_2 emissions in South America, Northern Africa, and India. These features are also visible in the assimilation results with amplitudes matching the truth closely in most locations. Areas of uptake or emissions are more widespread in the retrieved fluxes with smaller peak uptake signals, reflecting our inability to observe such

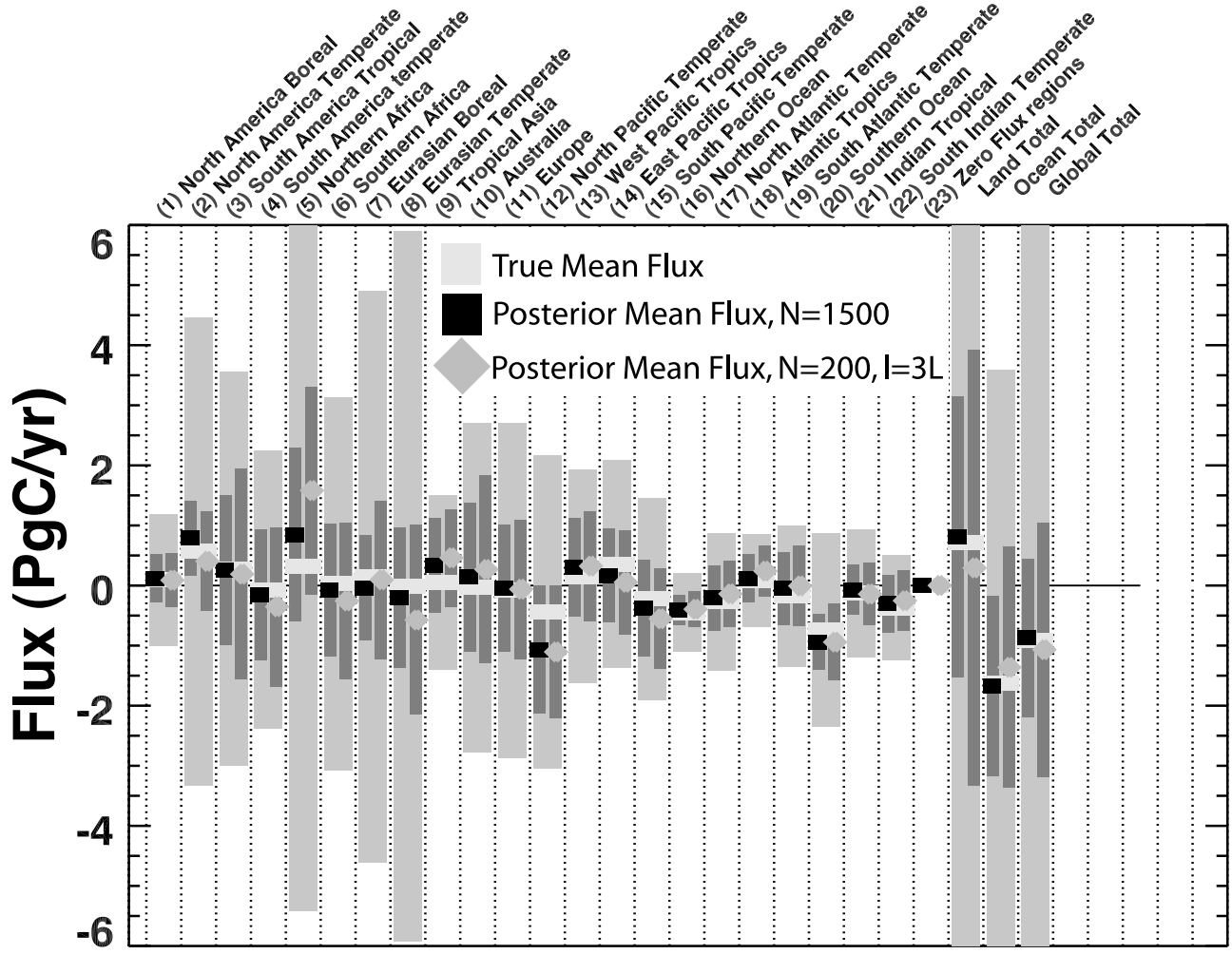

Figure 2. Annual mean fluxes aggregated to TransCom regions. Light blue symbols are the "true" values, green symbols represent analyzed fluxes from a run with 1500 ensemble members, and red symbols are results for 200 ensemble members and a localization length of three times the covariance scale (see text). The error bars have light shading for 1-sigma background flux uncertainties and dark shading for 1-sigma posterior uncertainties. The off-scale a priori land, ocean, and global uncertainties are 19.28, 5.2, and 19.9 PgC/yr, respectively. See color version of this figure at back of this issue.

detail with our current network. Also, some anticorrelated dipoles in the fluxes are visible that yield correct aggregated fluxes at continental scales, for instance in Australia. This is an example of our limited capability to see regional flux details from the CMDL network. In November, respiration dominates in the Northern Hemisphere with maximum CO_2 fluxes to the atmosphere in the Eurasian Boreal regions and large parts of North America. Tropical fluxes are strongly positive in the Amazon region and Southern Africa, balanced by uptake in Northern Africa. All these features are also seen in our retrieved fluxes including local maxima in Western Canada, the eastern US, and eastern Siberia. The tropical fluxes, although reasonable in magnitude, are again not always located correctly. This is specifically visible in Northern Africa where fluxes appear in the Sahara. Without more detailed a priori information our method is not prohibited to create this unphysical solution as long as it is mathematically correct.

[42] The true ocean flux in Figure 4c shows outgassing throughout the tropics peaking in the central Pacific, balanced by uptake in the Southern and Northern Ocean as well as the Southern Pacific and Indian Ocean. The retrieved ocean fluxes generally show similar features but are somewhat more noisy than the true fluxes. This is likely a result of the imposed covariance length scale that can locally force the retrieved solution away from the true solution.

4.1. Ensemble Size and Localization

[43] The results discussed above were produced with an ensemble size of 1500 members, a number that is generally too large to allow efficient calculation of the ensemble statistics. An important challenge therefore is to obtain similarly good answers with fewer ensemble members. Assimilation runs were done with fewer ensemble members, introducing various degrees of localization to ensure spurious covariances are suppressed. Figure 5 shows the results of these tests. The minimum RMS error is obtained at N = 1500, no localization which also has the most d.o.f. indicating that this solution is closest to the "truth" and

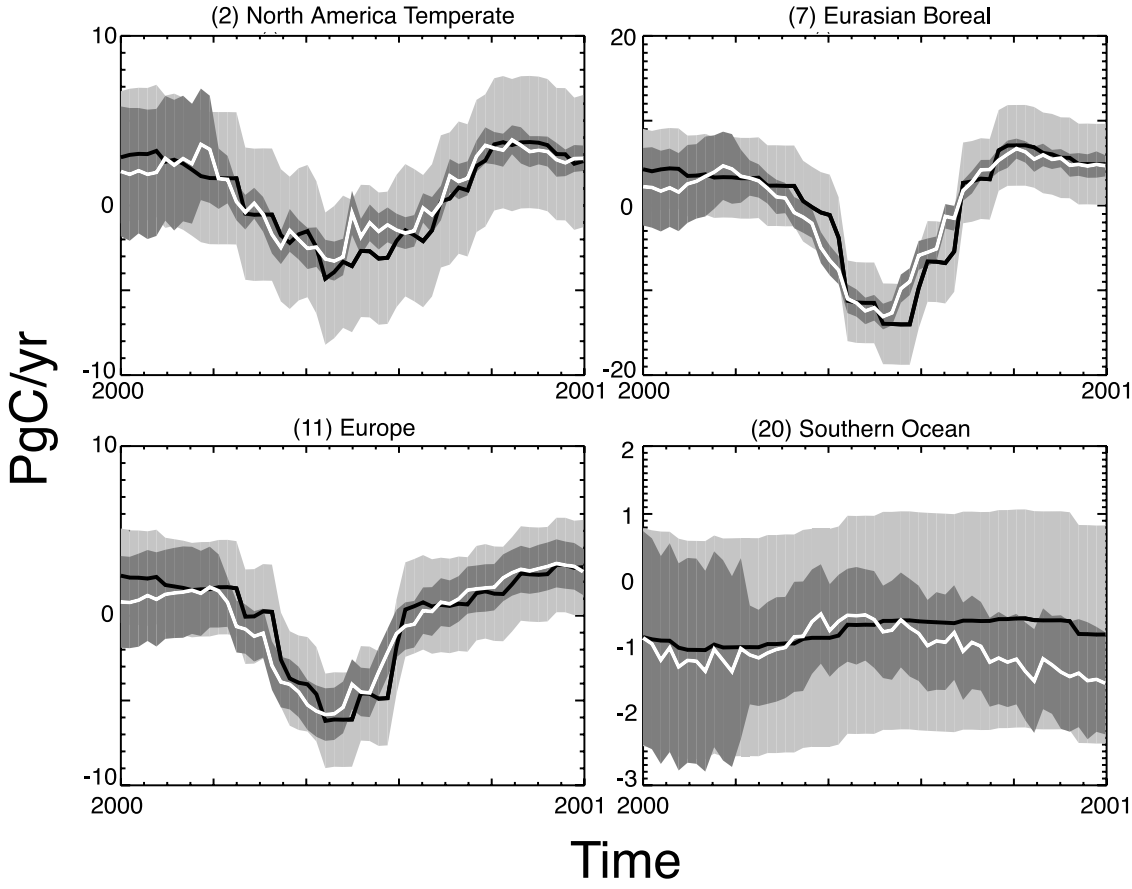

Figure 3. Seasonal cycles of the fluxes for aggregated TransCom regions North America, Boreal Eurasia, Europe, and the Southern Ocean. Solid lines are the "true" fluxes, and open lines are the assimilated ones. Light shaded bars are the a priori uncertainty, and dark shaded bars are posterior. Units are in PgC/yr. Note the different y scale on the panels and the reduction of uncertainty that starts after the first few months because of the filter spin-up.

contains the most information on the covariances. When fewer than 500 ensemble members are used, the results without localization deteriorate quickly as the covariance matrix becomes more inaccurate causing spurious adjustments to the state vector. When some of these spurious covariances are suppressed through localization (where the localization length scale l in equation (18) is chosen as 3× the covariance length scale L from equation (19)), the RMS continues to be low even for a 200 member ensemble. All simulations with fewer ensemble members show decreasing d.o.f.. Although we still retrieve the mean of the fluxes quite well with fewer members, we do not have as much detailed information on its covariance structure anymore. At some point this will affect the estimated variances as well as the covariances leading to biases in the uncertainty assigned to individual regions. Figure 6 shows an example of this for the variance on the fluxes. The pattern of uncertainty reduction on the annual mean fluxes deteriorates quickly when we go from N = 1500 to N = 200 members without localization as the representation of distant covariances becomes increasingly inaccurate because of statistical noise. However, when we suppress that noise through localization with l = 3L, the N = 200 simulation is nearly identical to the N = 1500 case. When we further reduce the number of ensemble members to N = 50, the result clearly misrepresents the posterior variances again. Note that these results are likely dependent on other parameters in the problem and should be determined for each inversion separately.

4.2. Assimilation Window Size

[44] Important considerations when choosing the assimilation window size are computational efficiency and estimation accuracy. A larger assimilation window means that each week of fluxes is constrained by more observations, but also requires longer integrations of the transport model, and more parameters to be estimated per cycle. The representation of a larger covariance matrix, moreover, requires more ensemble members.

[45] To quantitatively estimate how many weeks of lag our filter needs, we look at the spatial information contained in each consecutive estimate of a particular week of fluxes. Spatial information on the flux distribution is contained in two quantities: $\mathcal{H}$ and $\mathbf{P}^b$. The operator $\mathcal{H}$ determines which regions are constrained directly by observations because they are under the "footprint" of a sampling site. $\mathbf{P}^b$ determines which regions are constrained indirectly by inferring information from a neighboring region. We discuss the trade-off between these two methods of inferring information in section 5. Operator $\mathcal{H}$ and matrix

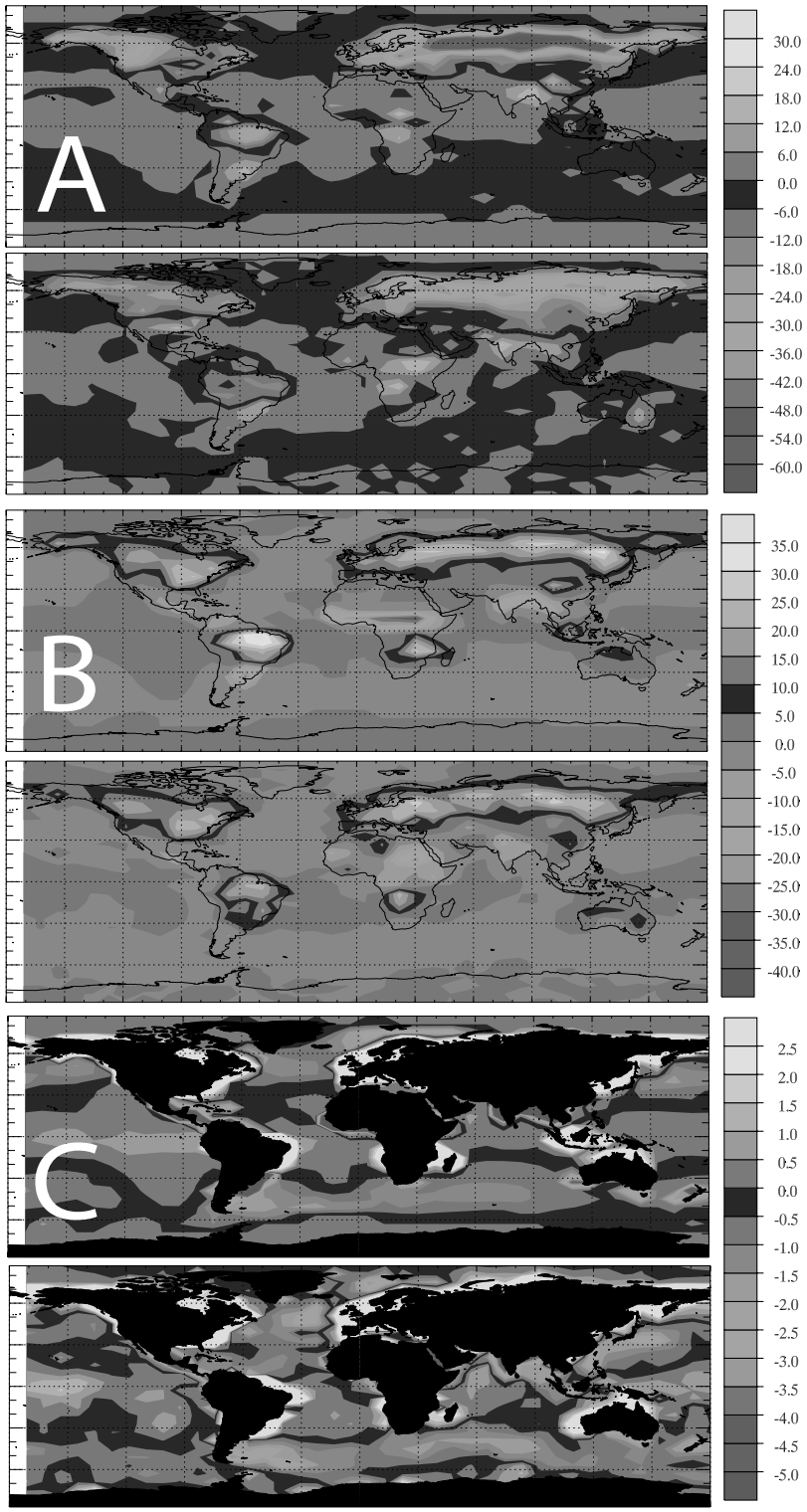

Figure 4. (a) Land fluxes in July, (b) land fluxes in November, and (c) ocean fluxes in November. Top plots show true fluxes, and bottom plots show assimilated fluxes. Note that some large fluxes near coastal regions are due to the coarse grid and do not represent true ocean fluxes. Original flux estimates were done at weekly timescales of which five are averaged in this plot. Units are 10^{-9} kgC/m^2/s. See color version of this figure at back of this issue.

$\mathbf{P}^b$ are convoluted in the Kalman gain matrix $\mathbf{K}$ through the numerator $\mathbf{PH}^T$ (equations (4) and (9)). If vector $\mathbf{k}$ has limited structure, the spatial information added to the state vector will be low.

[46] The question of how many weeks of lag to use can thus be rephrased: How much does each consecutive state estimate add to explaining the spatial structure of the final solution? This is investigated by calculating the amount of

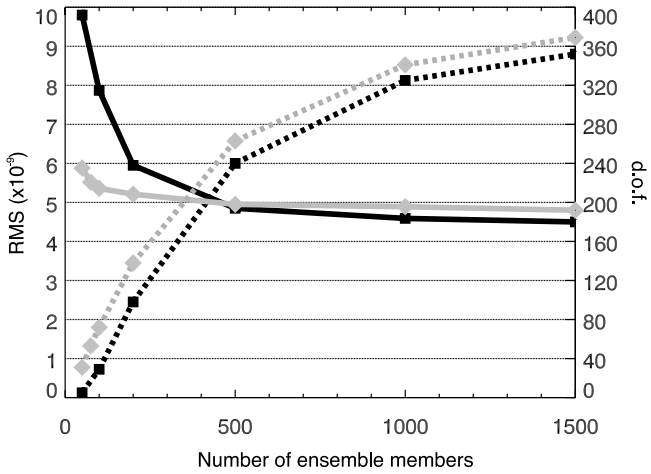

Figure 5. Root-mean-square flux differences (solid, units of 1×10^{-9} kgC/m^2/s) and number of degrees of freedom (d.o.f.) in the posterior ensemble (dashed) as a function of number of ensemble members. Black lines with squares indicate results without localization, while grey lines with diamonds indicate results with a localization strength of $l = 3 \times L_{l,o}$.

spatial structure in the final flux vector (assumed to be attained after twelve weeks) that is explained after each update of the state. This is calculated as the squared linear correlation coefficient (r) between the final vector **x**(12) and its partial solutions from previous steps **x**(0,..,11). The variance (r^2) will initially increase quickly because the first few state estimates generally use large and local transport signals carrying the bulk of the information. Later updates of the state occur with more diffuse transport signals causing the added variance to level off. Figure 7 illustrates this for 40 different state vectors from different weeks. It can be seen that on average, 90% of the variance of the final solution is captured after estimating the state eight times and all 40 estimates attained this percentage after ten weeks. Although the added variance from updates 9–12 is not zero, it is less than 3% per week. If we had assumed the final solution to be attained after 24 estimates instead of twelve, the tail of the curve would have been longer and contained a larger fraction of the total variance. Each estimate beyond eight weeks would again carry only a small fraction (<3%) of the variance. Although this analysis does not show beyond doubt that 8–10 time steps is enough to accurately retrieve the fluxes, it does show the quick decrease in information content as observations and fluxes get further

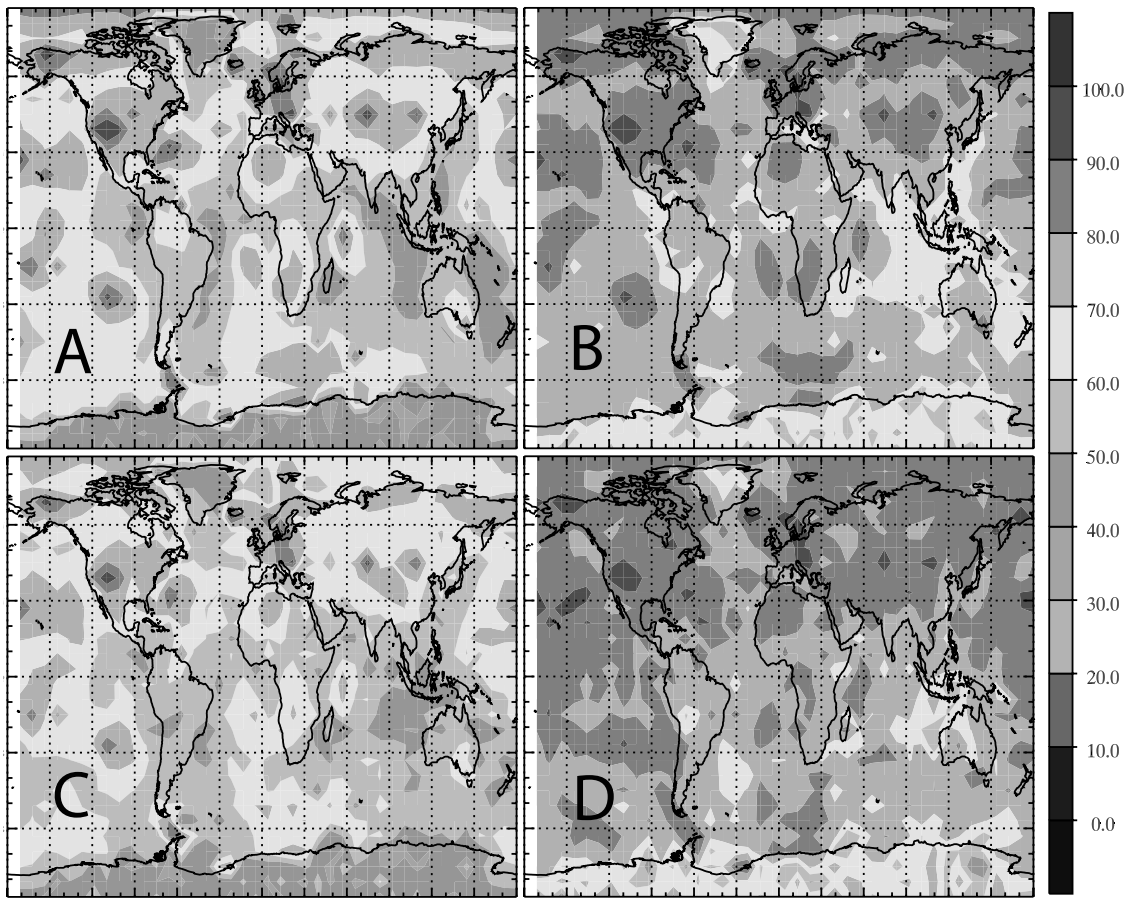

Figure 6. Annual mean uncertainty reduction (%) for a case with (a) 1500 members, no localization; (b) 200 members, no localization; (c) 200 members, localization with l = 3L; and (d) 50 members, localization with l = 3L. Inaccurate representation of the covariances leads to large and incorrect decreases of uncertainty. Comparing Figures 6b and 6c shows that localization can improve results with fewer ensemble members. See color version of this figure at back of this issue.

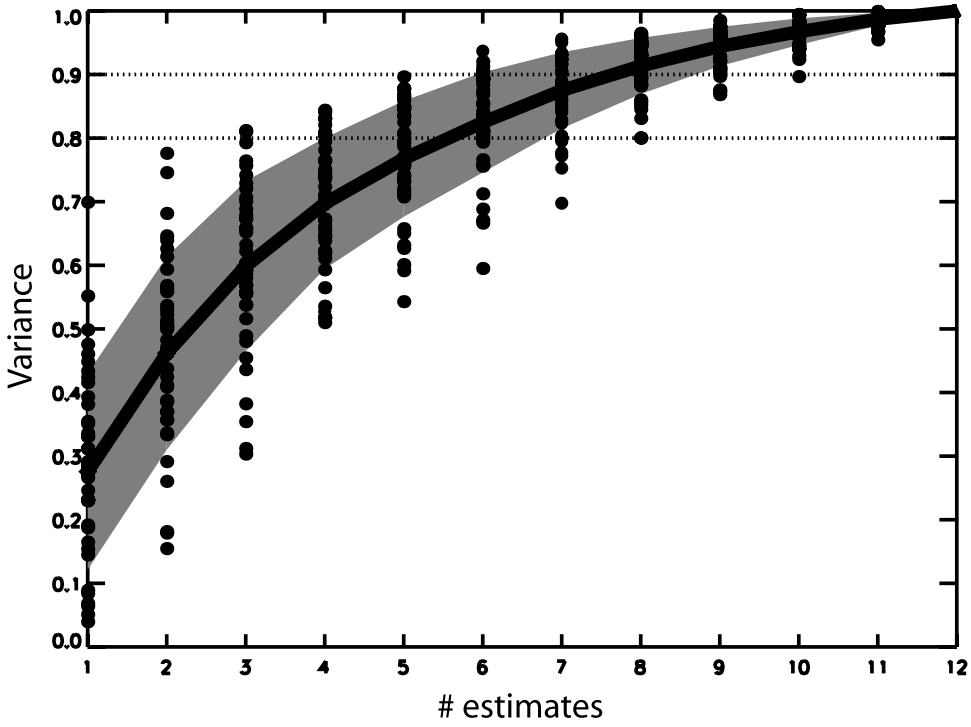

Figure 7. Amount of variance in a 12-week lag state estimate explained after each consecutive estimate. Individual points show the fractional variance (defined as r^2 with r the correlation coefficient between final and partial state update) of 40 different states, each estimated twelve times before being fixed. The thick line is the mean with 1-sigma standard deviation as shaded area. Note that the variance equals 1.0 after 12 estimates as this is assumed to be the final flux value. After 10 weeks more than 90% of the spatial structure of the final solution is captured.

separated in time. Interestingly, our results given by *Bruhwiler et al.* [2005] also suggested that 8–10 (one month) time steps in the assimilation window is sufficient to extract the spatial information of the flux patterns, whereas *Law* [2004] suggested six (one month) time steps.

[47] It is important to understand that signals beyond the lag time of our filter are not lost in the inversion. The background fields CO_{2i}(x,y,z,t) will carry information far beyond the lag time of the filter to ensure that all emissions will affect all sites eventually. The subtle difference is that we decide not to extract information about the spatial flux patterns anymore after 8–10 weeks. The influence of fluxes beyond the lag time shows up at each site through CO_{2i}(x,y,z,t + 12) instead of through $\mathcal{H}(\mathbf{x}_i)$, and are thus presubtracted to ensure consistency of all past fluxes with all future observations. The existence of an ensemble of flux histories in the form of N background CO_2 mixing ratio distributions causes also the covariance of past fluxes to be incorporated contrary to a regular Kalman smoother where only the mean background CO_2 mixing ratio distribution is retained.

5. Discussion

[48] The choice of the number of ensemble members to use directly influences the quality as well as the cost of the solution and should therefore be done with care. Although there are no formal rules to determine this parameter, the required ensemble size is related to the number of degrees of freedom in the covariance matrix. More ensemble members are needed to accurately represent a covariance matrix with more degrees of freedom. When background covariances are prescribed like in our method, the number of degrees of freedom can be controlled by prescribing covariances between parameters in the state. For example, it might be beneficial to set tighter covariances between regions in the poorly observed tropics, as this reduces the number of degrees of freedom and will thus improve our statistical representation of the covariance matrix, to benefit other regions where an abundance of observations allows us to retrieve fluxes in more detail such as North America or Europe. Note that if in the future satellite observations or a vastly improved surface network exists, the need to couple regions a priori through the covariance matrix will disappear. Our current experiences with 100–200 ensemble members required to represent a state vector of 14,400 elements is moderately promising for future applications of SEAT-A.

[49] As stated earlier, the state propagation model $\mathcal{M}$ plays an important role in an assimilation system. The skill of this model determines the quality of the first guess of the fluxes, and allows information to propagate through the filter from one time step to the next reducing the uncertainty, and thereby the number of degrees of freedom in the system to be solved. The poor skill of our current "persistence" model forces us to prescribe the covariances for each new estimate, and each estimate therefore starts with a relatively large and homogeneous uncertainty. Although the

problem of a poor state propagation model $\mathcal{M}$ could be partly solved by starting each new assimilation cycle from a ''climatological'' first guess (e.g., fluxes produced by process-based biosphere or ocean models), we prefer not to use such bottom-up products to prevent mixing the different information streams that we want to compare independently later on. Improvements to our state propagation model will likely improve our assimilation skills and allow us to retrieve fluxes more reliably. It could perhaps even allow us to couple the background covariances to the analyzed ensemble properties again. One approach could be to optimize for parameters in the covariance matrix at each step [see *Michalak et al.*, 2005; *Krakauer et al.*, 2004] on the basis of the fit of the ensemble CO_2 mixing ratios to the observed ones, causing the analyzed ensemble fluxes to influence the background covariances. In the work presented here we have not pursued this yet.

[50] An important way to improve the current method is through better specification of the background flux covariances $\mathbf{P}^b$. Given the relatively sparse observation network there are two ways to constrain the unknown fluxes: (1) by including the full coupling (through transport) between all fluxes to be estimated and all observations available in the inversion and (2) by specifying correlation structures that allow flux regions to be constrained indirectly through inference from flux regions that are directly constrained by observations. The flux estimates of *Gurney et al.* [2002], *Bousquet et al.* [2000], *Rayner et al.* [2002], *Houweling et al.* [1999], and *Law et al.* [2003] all employ method (1) and fully couple all observations to all fluxes. In addition, the studies of *Rödenbeck et al.* [2003] and *Michalak et al.* [2004] use a covariance structure as in (2) to further constrain the problem. This should not be confused with the much more rigorous constraint of imposing a fixed pattern of fluxes (essentially a covariance of 1.0) within a limited number of large regions to be estimated, as is commonly done in lower-resolution inversions. The covariance specification method (2) has the disadvantage that the true correlation structure of the fluxes is largely unknown and specification of background covariances affects the solutions to a degree that is difficult to assess. Using the transport model (1) has the disadvantage of making the inverse problem larger and thus more expensive to solve, and requires running the transport model for a long period of time just to constrain a few hard-to-observe regions while the majority of the regions are already well constrained in the early stages of assimilation. It also relies on the ability of coarse transport models to track the sampled air masses through the atmosphere accurately. Given the crude grids, time step, and parameterizations of vertical exchange in these models one should strongly question their ability to track the air masses even after just a few weeks of atmospheric transport. Moreover, atmospheric tracer transport calculations rely on reanalyzed meteorological products which only represent the best guess of the meteorological fields while not doing any justice to the stochastic nature of atmospheric transport, which could only be captured using an ensemble of transport fields.

[51] Several sources of information could be used to produce background covariance structures for our flux inversions. A first obvious choice could be to use an ecosystem database to introduce covariations only between similar ecosystems that are close together geographically. This would, for instance, allow us to infer information on the fluxes from a large forest by only observing parts of it, or separate the contribution from adjacent crops and grasslands even though we observe their combined signals in the atmosphere. Satellite products delivering land-surface characteristics such as greenness (NDVI), soil moisture, or even fire disturbances could be exploited as well. Such products have the advantage of high spatial and temporal coverage while not explicitly informing on the mean carbon fluxes. We want to stress that this covariance information is only needed while observations are limited. If available, observations of CO_2 and related trace gases will always be a preferred source of information over background covariance structures in SEAT-A.

[52] In all the discussions of covariances in this work, this quantity is used to inform on the relationship between individual a priori or posterior parameters in the state vector. The absolute magnitude of the posterior uncertainty is of lesser importance because it has a strong dependence on the choice of ''inversion'' parameters such as model-data mismatch and background covariance magnitude and length scales. This strong dependence implies that discussions of posterior uncertainty can only be useful in the context of the specific choices for these parameters. Much more than setting absolute uncertainties on each regions contribution to the atmospheric carbon pool, posterior covariances should therefore be used to (1) check the independence of the retrieved fluxes and aggregate regions where needed and (2) check the statistical correctness of the assimilation by analyzing the posterior PDF (Gaussian?) and its relationship to the inversion parameters (uncertainty reduction, innovation statistics, χ^2). The true uncertainty of top-down derived flux estimates should include many more aspects of the error structure such as possible biases in the sampling of observations, biases in model transport, covariances in sampling errors, and aggregation errors. The total uncertainty of regional or continental carbon fluxes should thus be constructed from a number of different methodologies with different models, different data sets, and different assumptions and should not be confused with the posterior covariances from an assimilation system such as SEAT-A.

[53] Finally, we want to mention the current weak points and drawbacks of our approach. The most important drawback of the current scheme is the previously mentioned lack of a dynamical model. This prevents us from exploiting the ''learning'' ability of the EnKF and introduces a reliance on prescribed background covariances. A second drawback is the statistical representation of some key properties of the system. Not only can the statistical representation itself be poor when insufficient ensemble members are used, the underlying statistical assumptions themselves (Gaussian PDFs, uncorrelated errors) can be wrong. Careful selection of the ''engineering'' parameters can be demanding and adds yet another dimension for sensitivity analysis of the results. From an implementation point of view the computer resources needed to run an ensemble data assimilation system like SEAT-A (transport model that scales over many parallel processors, large amounts of disk space and memory, 50–100 CPUs per simulation) can also be a limitation. The incomplete representation of the posterior

covariances is the most important concession to make in a system like SEAT-A. Therefore we recommend to only use an ensemble data assimilation system if the dimensions of the state vector, or of the observation vector necessitate this approach.

[54] In the near future, SEAT-A will be used to study the North American carbon cycle in more detail. For such an endeavor, more regions will be added over North America while fluxes at larger distances will be coupled either as larger regions or through a stronger covariance. Our ability to retrieve North American carbon fluxes at higher spatial resolution will furthermore depend on the number of observations available, and on TM5s ability to simulate CO_2 concentrations at continental sites. Introducing continuous CO_2 data measured from tall towers as well as constraints from the observed $^{13}C/^{12}C$ isotopic ratios will be an important step in the application of SEAT-A. Further developments of this method should focus on creating proper background covariance structures, creation of a dynamical model to forecast fluxes, stochastic treatment of uncertain transport model parameters, extension to other trace gases, and possibly transport model error estimates in the assimilation technique.

6. Conclusions

[55] We have demonstrated the use of an ensemble data assimilation method to estimate CO_2 surface fluxes from atmospheric observations. The new system overcomes some limitations of previously used inversion methods and shares some of their strengths and should be viewed as another possible Bayesian approach for CO_2 flux estimates. It can ingest large amounts of observations without the need to precalculate observation operators, estimate surface fluxes at the model grid scale without storing and inverting large matrices (similar to 4d-var), and provide a top-down view of the surface fluxes without reliance on bottom-up flux estimates (similar to regular Kalman filters and geostatistical methods). The representation of covariances between fluxes by an ensemble of states is necessarily limited by the ensemble size. The details of the system were demonstrated through a realistic pseudodata experiment in which fluxes were retrieved satisfactorily. For this problem, nine weeks of lag and 200 ensemble members with a localization over three times the covariance correlation length scale was found to work well. We stress that these values will likely depend on the particular problem one tries to solve. Extension of this method to other trace gases that include nonlinear chemical interactions will be part of future work.

Appendix A: TM5 Model

[56] The TM5 chemistry transport model is simplified to a tracer transport model for the CO_2 problem, and is a fully linear operator on CO_2 fluxes. Tracer transport (advection, vertical diffusion, cloud convection) is done by offline meteorological fields taken from the European Centre for Medium Range Weather Forecast (ECMFW) model, run in either forecast (T512L60) or reanalysis (ERA40) mode. All physical parameterizations in the TM5 model are kept as close as possible to the ECMWF formulation to achieve similarity between the two. TM5 offers the possibility to use online two-way nested grids in the model, giving it regional-scale capabilities in a global framework. For the NACP, we have defined a global $6^\circ \times 4^\circ$ TM5 grid with two nested grids that focus on North America at $3^\circ \times 2^\circ$, and the US plus parts of Canada on $1^\circ \times 1^\circ$. A tracer transport assessment in this configuration using SF_6 was recently published [*Peters et al.*, 2004]. Note that the model offers the flexibility to easily change this grid, allowing test runs without nested grids or at coarser global resolutions.

[57] The TM5 model is fully parallel implemented with MPI, with each processor carrying model variables distributed either over the number of tracers, or the number of vertical levels. The first option is very beneficial for SEAT-A, as it allows us to run each member of our ensemble as a separate tracer on a separate processor, greatly reducing the time required to sample N ensembles. Furthermore, an adjoint of TM5 is available that was used to calculate and store linearized observation operators (matrix form of **H**) for all observations as part of previous research. This allowed us to perform tests with large numbers of ensembles N by sampling model concentrations from a coarse grid (and thus quick) TM5 model calculation of $CO_{2i}(x,y,z,t + 12)$ from $CO_{2i}(x,y,z,t)$ with zero fluxes, augmented with many ensemble flux influences $\mathcal{H}(x_i)$ created with a simple matrix multiplication of the stored **H** with the ensemble of vectors x_i. Note that this approach will not be used in real applications but we decided to take advantage of the existing observation operators for the tests shown here.

[58] **Acknowledgments.** We are grateful to NOAA's High Performance Computing Division of the Forecast Systems Lab for providing an excellent supercomputing platform and support. We would like to thank two anonymous reviewers for their helpful comments. S.D. was supported by NASA cooperative agreement NCC5-621, and D.Z. was supported by NASA contract NNG05GD15G.

References

Andres, R. J., G. Marland, I. Fung, and E. Matthews (1996), A 1 degrees × 1 degrees distribution of carbon dioxide emissions from fossil fuel consumption and cement manufacture, 1950–1990, *Global Biogeochem. Cycles*, *10*(3), 419–429.

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*(5495), 1342–1346.

Brenkert, A. L. (1998), Carbon dioxide emission estimates from fossil-fuel burning, hydraulic cement production, and gas flaring for 1995 on a one degree grid cell basis, http://cdiac.esd.ornl.gov/epubs/ndp/ndp058a/ndp058a.html, Carbon Dioxide Inf. Anal. Cent., Oak Ridge Natl. Lab., Oak Ridge, Tenn.

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

Ciais, P., P. P. Tans, J. W. C. White, M. Trolier, R. J. Francey, J. A. Berry, D. R. Randall, P. J. Sellers, J. G. Collatz, and D. S. Schimel (1995), Partitioning of ocean and land uptake of CO_2 as inferred by delta-C-13 measurements from the NOAA Climate Monitoring and Diagnostics Laboratory global air sampling network, *J. Geophys. Res.*, *100*(D3), 5051–5070.

Cohn, S. E., N. S. Sivakumaran, and R. Todling (1994), A fixed-lag Kalman smoother for retrospective data assimilation, *Mon. Weather Rev.*, *122*(12), 2838–2867.

Conway, T. J., P. P. Tans, L. S. Waterman, and K. W. Thoning (1994), Evidence for interannual variability of the carbon-cycle from the National Oceanic and Atmospheric Administration/Climate Monitoring and Diagnostics Laboratory Global Air Sampling Network, *J. Geophys. Res.*, *99*(D11), 22,831–22,855.

Denning, A. S., I. Y. Fung, and D. Randall (1995), Latitudinal gradient of atmospheric CO_2 due to seasonal exchange with land biota, *Nature*, *376*(6537), 240–243.

Denning, A. S., M. Nicholls, L. Prihodko, I. Baker, P. L. Vidale, K. Davis, and P. Bakwin (2003), Simulated variations in atmospheric CO_2 over a Wisconsin forest using a coupled ecosystem-atmosphere model, *Global Change Biol.*, *9*(9), 1241–1250.

Enting, I. (2002), *Inverse Problems in Atmospheric Constituent Transport*, Cambridge Univ. Press, New York.

Eskes, H. J., P. F. J. Van Velthoven, P. J. M. Valks, and H. M. Kelder (2003), Assimilation of GOME total-ozone satellite observations in a three-dimensional tracer-transport model, *Q. J. R. Meteorol. Soc.*, *129*(590), 1663–1681.

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

Fan, S., M. Gloor, J. Mahlman, S. Pacala, J. Sarmiento, T. Takahashi, and P. Tans (1998), A large terrestrial carbon sink in North America implied by atmospheric and oceanic carbon dioxide data and models, *Science*, *282*(5388), 442–446.

Francey, R. J., P. P. Tans, C. E. Allison, I. G. Enting, J. W. C. White, and M. Trolier (1995), Changes in oceanic and terrestrial carbon uptake since 1982, *Nature*, *373*(6512), 326–330.

Gaspari, G., and S. E. Cohn (1999), Construction of correlation functions in two and three dimensions, *Q. J. R. Meteorol. Soc.*, *125*(554), 723–757.

Gurney, K. R., et al. (2002), Towards robust regional estimates of CO_2 sources and sinks using atmospheric transport models, *Nature*, *415*(6872), 626–630.

Hartley, D., and R. Prinn (1993), Feasibility of determining surface emissions of trace gases using an inverse method in a three-dimensional chemical transport model, *J. Geophys. Res.*, *98*(D3), 5183–5198.

Houtekamer, P. L., and H. L. Mitchell (1998), Data assimilation using an ensemble Kalman filter technique, *Mon. Weather Rev.*, *126*(3), 796–811.

Houweling, S., T. Kaminski, F. Dentener, J. Lelieveld, and M. Heimann (1999), Inverse modeling of methane sources and sinks using the adjoint of a global transport model, *J. Geophys. Res.*, *104*(D21), 26,137–26,160.

Ide, K., P. Courtier, M. Ghil, and A. C. Lorenc (1997), Unified notation for data assimilation: Operational, sequential and variational, *J. Meteorol. Soc. Jpn.*, *75*(1B), 181–189.

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.

Keeling, R. F., S. C. Piper, and M. Heimann (1996), Global and hemispheric CO_2 sinks deduced from changes in atmospheric O_2 concentration, *Nature*, *381*(6579), 218–221.

Khattatov, B. V., J. F. Lamarque, L. V. Lyjak, R. Menard, P. Levelt, X. X. Tie, G. P. Brasseur, and J. C. Gille (2000), Assimilation of satellite observations of long-lived chemical species in global chemistry transport models, *J. Geophys. Res.*, *105*(D23), 29,135–29,144.

Kleiman, G., and R. G. Prinn (2000), Measurement and deduction of emissions of trichloroethene, tetrachloroethene, and trichloromethane (chloroform) in the northeastern United States and southeastern Canada, *J. Geophys. Res.*, *105*(D23), 28,875–28,893.

Krakauer, N. Y., T. Schneider, J. T. Randerson, and S. C. Olsen (2004), Using generalized cross-validation to select parameters in inversions for regional carbon fluxes, *Geophys. Res. Lett.*, *31*, L19108, doi:10.1029/2004GL020323.

Krol, M. C., S. Houweling, B. Bregman, M. van den Broek, A. Segers, P. van Velthoven, W. Peters, F. J. Dentener, and P. Bergamaschi (2005), The two-way nested global chemistry-transport zoom model TM5: Algorithm and applications, *Atmos. Chem. Phys.*, *5*, 417–432.

Law, R. M. (2004), Technical note: An interannual inversion method for continuous CO_2 data, *Atmos. Chem. Phys.*, *4*, 477–484.

Law, R. M., Y. H. Chen, and K. R. Gurney (2003), TransCom 3 CO_2 inversion intercomparison: 2. Sensitivity of annual mean results to data choices, *Tellus, Ser. B*, *55*(2), 580–595.

Lorenc, A. C. (2003), The potential of the ensemble Kalman filter for NWP—A comparison with 4D-Var, *Q. J. R. Meteorol. Soc.*, *129*, 3183–3203.

Lyster, P. M., S. E. Cohn, R. Menard, L. P. Chang, S. J. Lin, and R. G. Olsen (1997), Parallel implementation of a Kalman filter for constituent data assimilation, *Mon. Weather Rev.*, *125*(7), 1674–1686.

Maksyutov, S., T. Machida, H. Mukai, P. K. Patra, T. Nakazawa, and G. Inoue (2003), Effect of recent observations on Asian CO_2 flux estimates by transport model inversions, *Tellus, Ser. B*, *55*(2), 522–529.

Masarie, K. A., and P. P. Tans (1995), Extension and integration of atmospheric carbon-dioxide data into a globally consistent measurement record, *J. Geophys. Res.*, *100*(D6), 11,593–11,610.

McKinley, G. A., M. J. Follows, and J. Marshall (2004), Mechanisms of air-sea CO_2 flux variability in the equatorial Pacific and the North Atlantic, *Global Biogeochem. Cycles*, *18*, GB2011, doi:10.1029/2003GB002179.

Menard, R., S. E. Cohn, L. P. Chang, and P. M. Lyster (2000), Assimilation of stratospheric chemical tracer observations using a Kalman filter. part I: Formulation, *Mon. Weather Rev.*, *128*(8), 2654–2671.

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.

Michalak, A. M., A. Hirsch, L. Bruhwiler, K. R. Gurney, W. Peters, J. B. Miller, and P. P. Tans (2005), Maximum likelihood estimation of covariance parameters for Bayesian atmospheric trace gas surface flux inversions, *J. Geophys. Res.*, doi:10.1029/2005JD005970, in press.

Miller, S. M., H. E. Snell, and J. L. Moncet (1999), Simultaneous retrieval of middle atmospheric temperature and trace gas species volume mixing ratios from Cryogenic Infrared Radiance Instrumentation for Shuttle (CIRRIS 1A), *J. Geophys. Res.*, *104*(D15), 18,697–18,714.

Patil, D. J., B. R. Hunt, E. Kalnay, J. A. Yorke, and E. Ott (2001), Local low dimensionality of atmospheric dynamics, *Phys. Rev. Lett.*, *86*(26), 5878–5881.

Peters, W., M. C. Krol, E. Dlugokencky, F. J. Dentener, P. Bergamaschi, G. Dutton, P. van Velthoven, J. B. Miller, L. Bruhwiler, and P. P. Tans (2004), Toward regional-scale modeling using the two-way nested global model TM5: Characterization of transport using SF_6, *J. Geophys. Res.*, *109*, D19314, doi:10.1029/2004JD005020.

Pétron, G., C. Granier, B. Khattatov, V. Yudin, J. F. Lamarque, L. Emmons, J. Gille, and D. P. Edwards (2004), Monthly CO surface sources inventory based on the 2000–2001 MOPITT satellite data, *Geophys. Res. Lett.*, *31*, L21107, doi:10.1029/2004GL020560.

Peylin, P., D. Baker, J. Sarmiento, P. Ciais, and P. Bousquet (2002), Influence of transport uncertainty on annual mean and seasonal inversions of atmospheric CO_2 data, *J. Geophys. Res.*, *107*(D19), 4385, doi:10.1029/2001JD000857.

Peylin, P., P. Rayner, P. Bousquet, C. Carouge, F. Hourdin, P. Heinrich, and P. Ciais (2005), Daily CO_2 flux estimates over Europe from continuous atmospheric measurements: 1: Inverse methodology, *Atmos. Chem. Phys. Disc.*, *5*, 1647–1678.

Randerson, J. T., M. V. Thompson, T. J. Conway, I. Y. Fung, and C. B. Field (1997), The contribution of terrestrial sources and sinks to trends in the seasonal cycle of atmospheric carbon dioxide, *Global Biogeochem. Cycles*, *11*(4), 535–560.

Rayner, P. J., and R. M. Law (1999), The interannual variability of the global carbon cycle, *Tellus, Ser. B*, *51*(2), 210–212.

Rayner, P. J., R. M. Law, D. M. O'Brien, T. M. Butler, and A. C. Dilley (2002), Global observations of the carbon budget: 3. Initial assessment of the impact of satellite orbit, scan geometry, and cloud on measuring CO_2 from space, *J. Geophys. Res.*, *107*(D21), 4557, doi:10.1029/2001JD000618.

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.

Stajner, I., and K. Wargan (2004), Antarctic stratospheric ozone from the assimilation of occultation data, *Geophys. Res. Lett.*, *31*, L18108, doi:10.1029/2004GL020846.

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.

Tans, P. P., I. Y. Fung, and T. Takahashi (1990), Observational Constraints on the global atmospheric CO_2 budget, *Science*, *247*(4949), 1431–1438.

Tarantola, A. (2004), *Inverse Problem Theory and Methods for Model Parameter Estimation*, Soc. for Ind. and Appl. Math., Philadelphia, Pa.

Whitaker, J. S., and T. M. Hamill (2002), Ensemble data assimilation without perturbed observations, *Mon. Weather Rev.*, *130*(7), 1913–1924.

Wofsy, S. C., and R. C. Harriss (2002), The North American Carbon Program (NACP), report, 75 pp., U.S. Global Change Research Program, Washington, D. C.

Yudin, V. A., G. Pétron, J. F. Lamarque, B. V. Khattatov, P. G. Hess, L. V. Lyjak, J. C. Gille, D. P. Edwards, M. N. Deeter, and L. K. Emmons (2004), Assimilation of the 2000–2001 CO MOPITT retrievals with optimized surface emissions, *Geophys. Res. Lett.*, *31*, L20105, doi:10.1029/2004GL021037.

L. Bruhwiler, A. Hirsch, J. B. Miller, W. Peters, P. P. Tans, and J. Whitaker, Global Monitoring Division, NOAA Earth Systems Research Lab, 325 Broadway R/CMDL-1, Boulder, CO 80302, USA. (wouter.peters@noaa.gov)

A. S. Denning, Department of Atmospheric Science, Colorado State University, Fort Collins, CO 80523, USA.

M. C. Krol, Netherlands Institute for Space Research, NL-3584 CA Utrecht, Netherlands.

D. Zupanski, Cooperative Institute for Research in the Atmosphere, Fort Collins, CO 80523, USA.

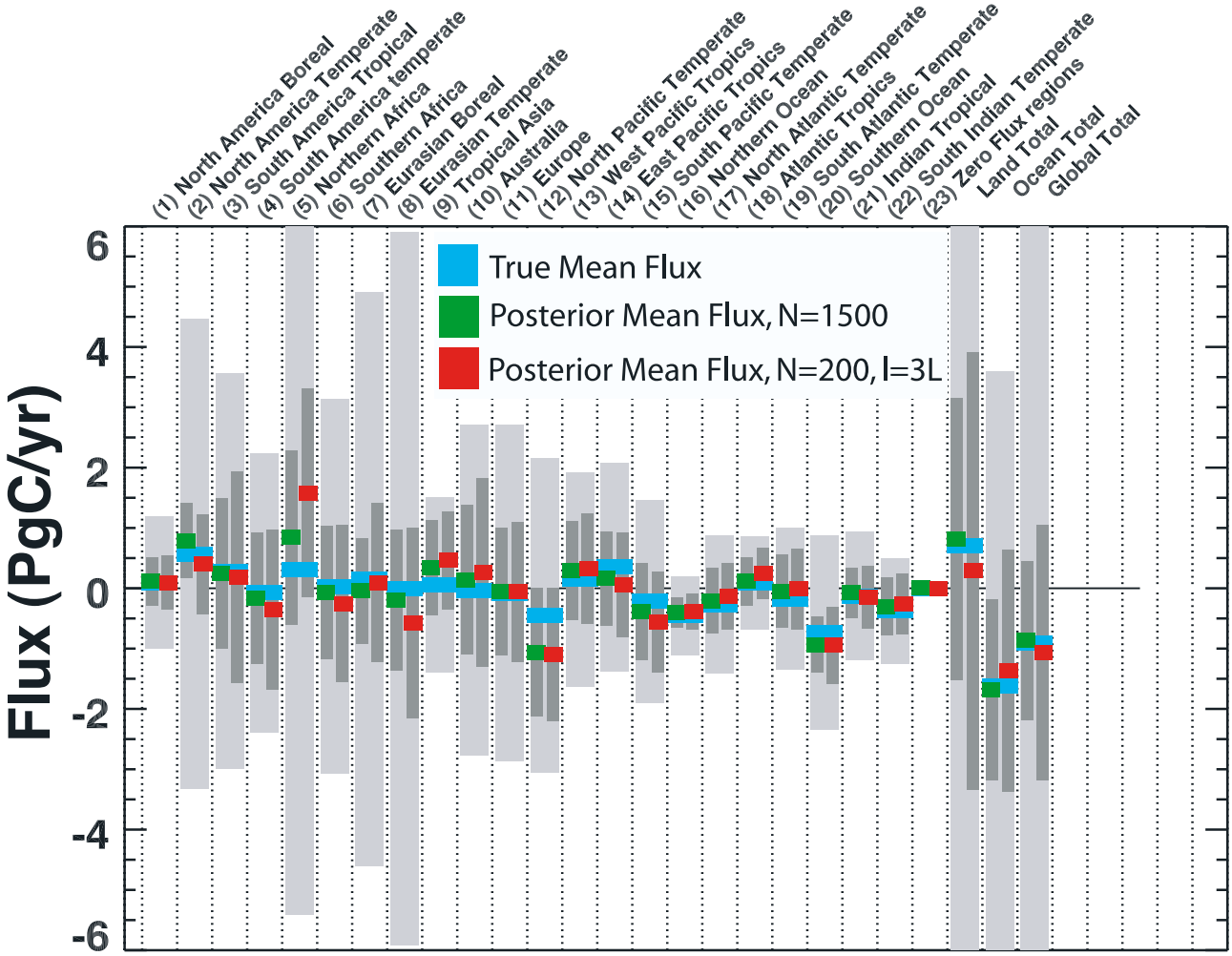

Figure 2. Annual mean fluxes aggregated to TransCom regions. Light blue symbols are the "true" values, green symbols represent analyzed fluxes from a run with 1500 ensemble members, and red symbols are results for 200 ensemble members and a localization length of three times the covariance scale (see text). The error bars have light shading for 1-sigma background flux uncertainties and dark shading for 1-sigma posterior uncertainties. The off-scale a priori land, ocean, and global uncertainties are 19.28, 5.2, and 19.9 PgC/yr, respectively.

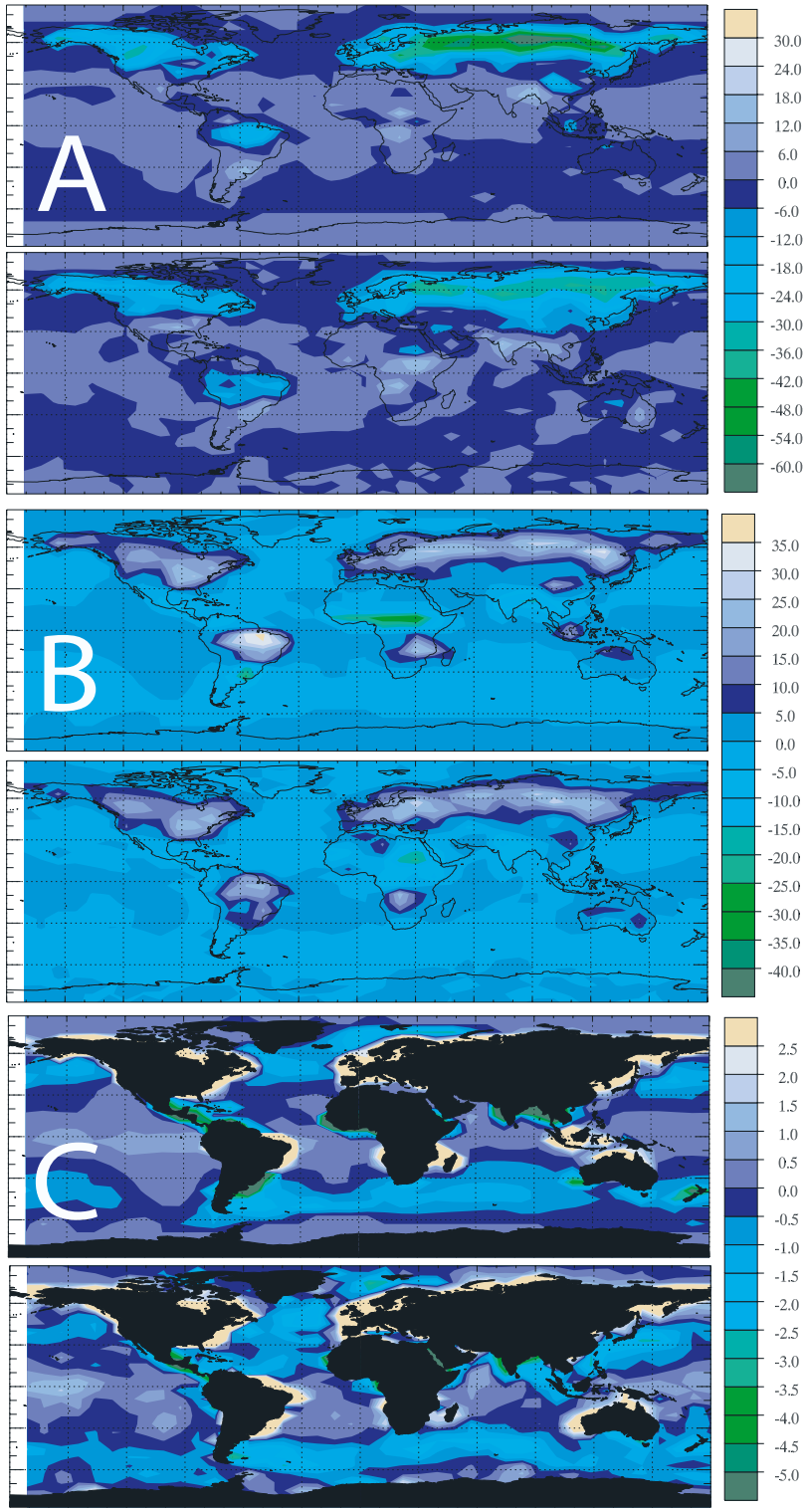

Figure 4. (a) Land fluxes in July, (b) land fluxes in November, and (c) ocean fluxes in November. Top plots show true fluxes, and bottom plots show assimilated fluxes. Note that some large fluxes near coastal regions are due to the coarse grid and do not represent true ocean fluxes. Original flux estimates were done at weekly timescales of which five are averaged in this plot. Units are 10^{-9} kgC/m^2/s.

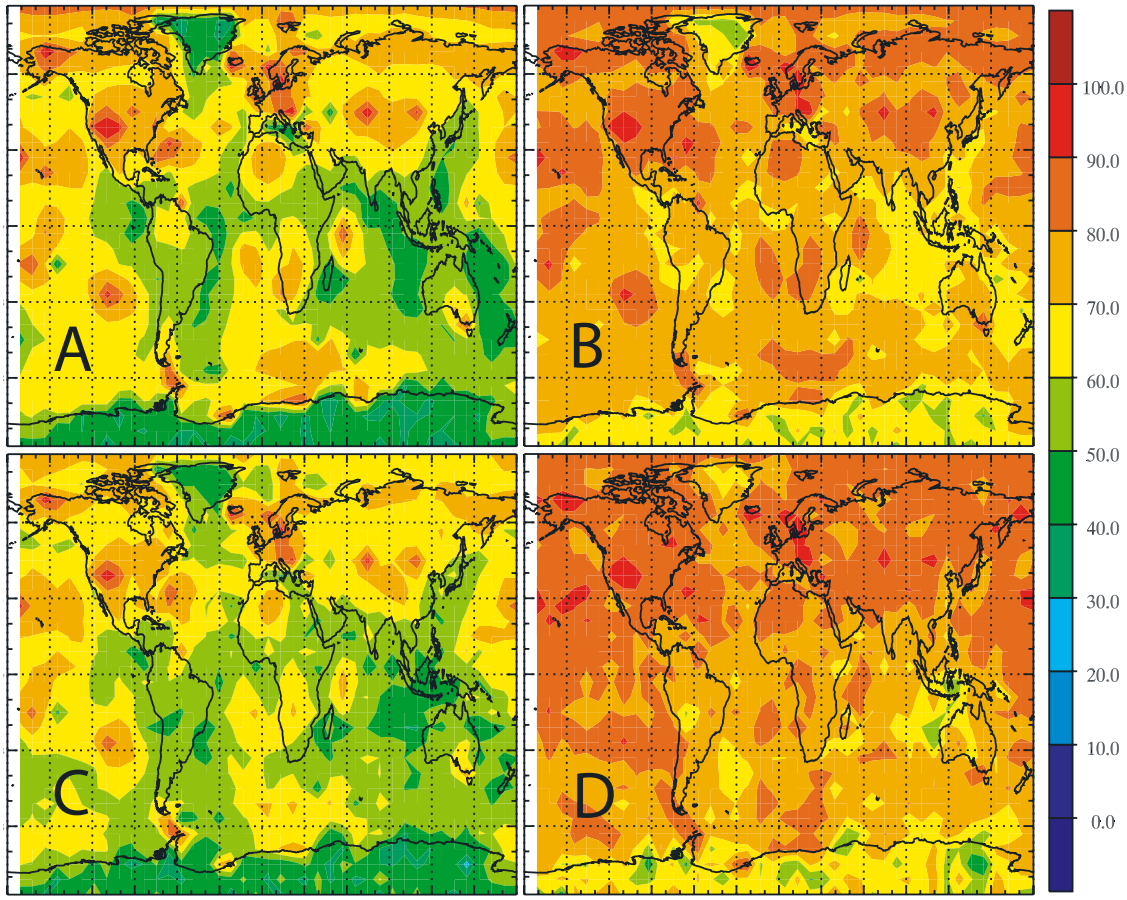

Figure 6. Annual mean uncertainty reduction (%) for a case with (a) 1500 members, no localization; (b) 200 members, no localization; (c) 200 members, localization with l = 3L; and (d) 50 members, localization with l = 3L. Inaccurate representation of the covariances leads to large and incorrect decreases of uncertainty. Comparing Figures 6b and 6c shows that localization can improve results with fewer ensemble members.