UMD-UCB carbon cycle data assimilation system

Ji-Sun Kang, Eugenia Kalnay, Inez Fung, Junjie Liu, and Takemasa Miyoshi

1. Introduction

Our project focused on quantifying surface carbon fluxes varying in space and in time, using observations of atmospheric CO2 concentrations. Although it is a "top-down" approach, our methodology, denoted UMD-UCB carbon cycle data assimilation system (UMD-UCB CDA) and described in Kang et al. 2011, 2012), introduced several new techniques compared to previous top-down methods (e.g. Peters et al. 2005, 2007; Baker et al. 2006, 2010; Rayner et al. 2008; Michalak, 2008; Chevallier et al. 2009; Feng et al., 2009). The UMD-UCB CDA analyzes meteorological and carbon variables simultaneously, whereas other studies only focus on the estimation of surface carbon fluxes. It deals with uncertainties of atmospheric CO2 concentrations caused by surface carbon fluxes as well as by the wind. It introduces a new approach, known as "localization of variables" that optimizes the configuration of background error covariance among the variables by including only error covariances between physically relevant variables and zeroing out sampling errors in the covariance between unrelated variables (Kang et al. 2011).

In addition to other advanced ensemble Kalman filter data assimilation techniques such as adaptive inflation and localization methods, we use a short analysis window of just six hours, similar to that used in atmospheric data assimilation. Previous studies usually use long windows, from several weeks to months (Kang et al. 2012). Short windows should allow updating surface carbon fluxes using observed atmospheric CO2 with minimal attenuation of information between atmospheric CO2 and surface carbon fluxes. Although "top-down" approaches assume that trajectory or evolution of emitted CO2 from the surface is reasonably well known, it is impossible to avoid errors in the transport and turbulent mixing of CO2, especially with a long time window, which led Enting (2002) to describe the inversion problem as being intrinsically ill-posed. In the past, with only about hundred observations of CO2 per week over the entire globe, the lack of information was a more serious problem than the transport errors for each observation, so that it made complete sense to use each observation in such a way as to constrain a very broad area on the globe by using long assimilation windows. However, with the current and near future availability of satellite data measuring atmospheric CO2, together with the more abundant ground-based observations, it may be possible to provide a global constraint for surface carbon fluxes at finer resolution even within a short window. Also, a short window allows incorporating uncertainties of the weather variables into a carbon cycle data assimilation system. With a long assimilation window, we cannot include wind uncertainty in UMD-UCB CDA, because a long window leads to nonlinear growth of ensemble perturbations and the weather variables lose predictability.

The main purpose of this paper is to test whether it is possible to retrieve the evolving surface carbon fluxes accurately at grid point resolution by simultaneously assimilating weather and carbon variables using short windows with the advanced Local Ensemble Transform Kalman Filter (LETKF, Hunt et al., 2007) with additional features. For simplicity, we test this hypothesis using an Observing

System Simulation Experiment (OSSE) under the assumption of a perfect model. For this purpose, we use the SPEEDY model (Molteni, 2003), a realistic but fast spectral atmospheric model, to which we added a carbon prognostic variable as a tracer, and surface carbon fluxes assumed to be constant except when modified by the data assimilation. In the "nature" run from which observations are simulated, terrestrial carbon fluxes are obtained from CASA (Randerson et al. 1997), and oceanic flux from Takahashi et al. (2002).

In section 2, we describe the standard LETKF, section 3 discusses the analysis of surface CO2 fluxes and atmospheric CO2 concentrations in addition to weather variables within UMD-UCB CDA. Results are shown in section 4 and we summarize conclusions and future work in section 5.

2. Standard Local Ensemble Transform Kalman Filter

The analysis cycle for numerical weather prediction (NWP) starts by computing an ensemble of K 6-hour forecasts with the previous ensemble of K analyses as initial conditions. Next, we compute the background ensemble perturbation matrix $\mathbf{X}^b$, whose columns contain a departure of each ensemble forecast ($\mathbf{x}^{b(k)}$) from the ensemble mean ($\overline{\mathbf{x}}^b$): the k-th column of $\mathbf{X}^b$ is $\mathbf{x}^{b(k)} - \overline{\mathbf{x}}^b$, $\{k = 1,2,\ldots,K\}$, and $\mathbf{x}$ is a state vector of dynamic variables at the model grids. Then, the observation operator h is applied to the ensemble forecast $\mathbf{x}^b$ to transform the background from the model grid space to the observation space, $\mathbf{y}^{b(k)} = h(\mathbf{x}^{b(k)})$. Let $\mathbf{Y}^b = \mathbf{y}^{b(k)} - \overline{\mathbf{y}}^b$, $\{k = 1,2,\ldots,K\}$, be the ensemble background perturbations in observation space.

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

$$\overline{\mathbf{x}}^a_{(l)} = \overline{\mathbf{x}}^b_{(l)} + \mathbf{X}^b_{(l)}\overline{\mathbf{w}}_{(l)}, \tag{1}$$

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

$$\overline{\mathbf{w}}_{(l)} = \tilde{\mathbf{P}}^a_{(l)}(\mathbf{Y}^b_{(l)})^T \mathbf{R}^{-1}_{(l)}(\mathbf{y}^o_{(l)} - \overline{\mathbf{y}}^b_{(l)}). \tag{2}$$

Here, $\tilde{\mathbf{P}}^a_{(l)} = [(\mathbf{Y}^b_{(l)})^T \mathbf{R}^{-1}_{(l)}(\mathbf{Y}^b_{(l)}) + (K-1)\mathbf{I}/\rho]^{-1}$ is the analysis error covariance computed in ensemble space where $\mathbf{B}^{-1}_{(l)} = (K-1)\mathbf{I}$, $\mathbf{R}$ is the observation error covariance matrix, $\mathbf{y}^o$ is the observation vector, and ρ is the inflation factor, and the subscript (l) indicates a quantity defined on a local region centered at the analysis grid point l. Within a local region, space localization is carried out by multiplying the inverse observation error covariance matrix $\mathbf{R}^{-1}_{(l)}$ by a factor that decays from one to zero as the distance of the observations from the analysis grid point increases (Greybush et al, 2011).

From (1) and (2) the analysis increment, $\overline{\mathbf{x}}^a_{(l)} - \overline{\mathbf{x}}^b_{(l)}$ is given by the background perturbation matrix multiplied by the weight vector, which is a function of the innovation, $\mathbf{y}^o_{(l)} - \overline{\mathbf{y}}^b_{(l)}$, and error statistics of both background and observations. Thus, the analysis reflects observational information more than background information if the background error is greater than the observation errors, and vice versa. In addition, the ensemble perturbations of the analysis are determined by

$$\mathbf{X}^a_{(l)} = \mathbf{X}^b_{(l)}[(K-1)\widetilde{\mathbf{P}}^a_{(l)}]^{\frac{1}{2}} \tag{3}$$

With (3), we obtain the estimation of analysis uncertainty in addition to the analysis mean. The global analysis ensemble $\mathbf{x}^{a(i)}$, $\{i = 1,2,\ldots,K\}$, is formed by gathering the values obtained for $\overline{\mathbf{x}}^a_{(l)}$ and $\mathbf{X}^a_{(l)}$ at all the analysis grid points (Hunt et al. [2007]).

3. Modifications of the LETKF made in the UMD-UCB CDA

For standard NWP applications, the state vector **x** includes wind (U, V), temperature (T), humidity (q), and surface pressure (Ps). In the UMD-UCB CDA, we augment **x** with the atmospheric CO2 concentration (C), a prognostic variable treated in the SPEEDY-C model as a passive tracer, as well as with surface carbon flux (CF), which we define as an "evolving parameter" (i.e., a variable for which there are no observations). The forecast model that provides the background state to the analysis in the UMD-UCB CDA has six prognostic variables (U, V, T, q, Ps, C), and reads CF as a surface forcing every six hours. Thus, we use an augmented state vector **x** consisting of (U, V, T, q, Ps, C, CF) at all model grid points, where the "parameter" CF, like Ps, is defined at the model surface grid points. In the current version of the system, there is no process to update CF during the forecast step. On the other hand, we assimilate observations of (U, V, T, q, Ps) as well as atmospheric CO2 concentration (C), but no direct observations of CF, which is only updated by the analysis step through the multivariate background error covariance. With this experimental setting, UMD-UCB CDA is based on a flow-dependent multivariate background error covariance of (U, V, T, q, Ps, C, CF) when assimilating (U, V, T, q, Ps, C) observations every six hours at the analysis step.

3.1. Localization of variables

Kang et al. 2011 (hereafter K11) pointed out that error covariances between unrelated variables inevitably introduce sampling errors in the standard EnKF. Thus, K11 zeroed out unphysical error covariances between unrelated variables (denoted Localization of variables or LOCvar) and found that this strongly reduced sampling errors. Winds transport atmospheric CO2, so that the best strategy of the

LOCvar for carbon cycle data assimilation obtained by K11 was to keep the error covariance computed between wind fields (U, V) and atmospheric CO2 (C) but to zero out the error covariances between (U, V) and surface carbon fluxes (CF), as well as between (T, q, Ps) and (C, CF) (referred to as "localized 1-way multivariate data assimilation" in Figure 1). In this LOCvar system, the analysis of meteorological variables was not influenced by CO2 variables because the wind uncertainty was estimated from the analysis of meteorological variables and provided to the atmospheric CO2 analysis in a "one-way" mode (no feedbacks allowed from the CO2 analysis to the wind). This was done because it was found that modifying the winds with the information provided by the wind-atmospheric C error covariance, as would be done in the standard LETKF increased the wind errors.

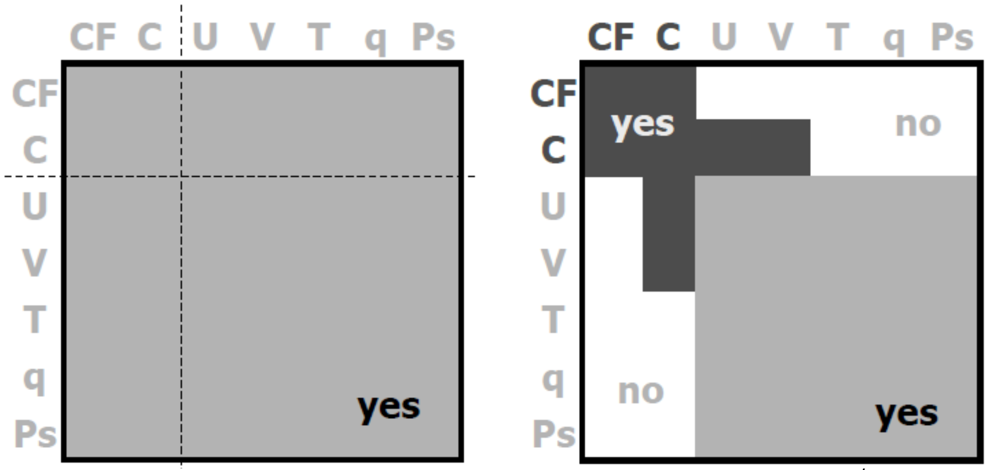


Figure 1. Schematic plots of background error covariance matrix $\mathbf{P}^b = \mathbf{X}^b \mathbf{X}^{b^t} / (K-1)$ for a fully multivariate (left) and a localized-1way (right) analysis system. The shading of the variable names are matched with the system used for their updates. White areas with "no" indicate the error correlation between variables is assumed to be zero during the analysis, while areas with "yes" indicate that the errors are allowed to be correlated (adapted from Figure 1 of Kang et al. 2011)

3.2. Inflation

In ensemble Kalman filter data assimilation (EnKF), the analysis is the optimal interpolation between the ensemble of short-range model forecasts and the observations. If the background (forecast) error variance is underestimated, the analysis system deemphasizes the observations because the analysis algorithm places too much confidence in the background. This process feedbacks, and may eventually lead to filter divergence, with the ensemble analysis no longer reflecting the observations. Conversely, an overestimation of background error variance leads to over-fitting the observations. In EnKF, the background error variance estimated from the ensemble forecasts tends to be underestimated because the ensemble size is limited and the forecast model is imperfect. This underestimation of background uncertainty would diminish the impact of the observations, especially in data-rich regions (Whitaker et al., 2008; Miyoshi et al., 2010). Thus, there is a need to inflate the background/analysis error covariances in such a way that the ensemble spread does not collapse in data-rich regions and is not excessively large in data-poor regions.

In order to better represent the background uncertainty in the EnKF analysis, we have implemented the adaptive multiplicative inflation of Miyoshi (2011). Since CF is not measured and its forecast is persistence, it was found necessary to further inflate the ensemble analyses of (C, CF) with an additive inflation in addition to the standard multiplicative inflation. The added random fields for each ensemble member are selected from the nature run: pairs of atmospheric CO2 and surface carbon flux fields are chosen randomly within 30 days centered at the analysis time and then scaled to a magnitude corresponding to a 6-hour difference. The additive fields for atmospheric CO2 are scaled again by a factor of 0.70, chosen by trial and error, before being added to each analysis ensemble. We scale the amplitude of the additive fields for CF by a factor of 0.1 and adjust this factor by requiring that the global mean spread of CF ensemble be at least 5.0×10^{-9} kg/m^2/s.

3.3. Vertical localization of column mixing CO2 data

For GOSAT column CO2 data, the averaging kernel is nearly uniform from the surface to the upper troposphere (Figure 2). However, the forcing of atmospheric CO2 that we would like to estimate takes place near the surface. Within a 6-hour time step, we do not expect a significant impact of CF on the CO2 mixing ratio throughout the entire troposphere, as we would observe if we used the averaging kernel as observation operator. Instead, we confine the changes in the CO2 mixing ratio due to the observed changes in the CO2 to the lower atmosphere and let the forecast model mix the CO2 vertically. This is equivalent to "localizing" the column CO_2 information, and updating only the lower atmospheric CO_2 rather than the full column of CO_2. In other words, the vertical localization function is large in the lower troposphere but zero in the upper layers, because atmospheric CO_2 is forced by the surface carbon fluxes and is well mixed in the upper layers. Therefore, after comparing column CO_2 background (forecast) with the GOSAT observation in the observation space, the innovations (observations minus model forecast) are used only in the lowest three vertical levels (up to sigma level of 0.685) for the analysis of CO_2. Here, the column CO_2 background is computed by multiplying the averaging kernel by the CO_2 concentration at each level: $\mathbf{y}_k^b = \mathbf{H}(\mathbf{x}_{k,l}^b) = \sum_{l=1}^{nlev} a_l S(\mathbf{x}_{k,l}^b)$ where the subscript k indicates the k-th ensemble member, the subscript l indicates the l-th vertical level of total *nlev* levels, $\mathbf{y}_k^b$ is the column CO_2 background at the observation space, $\mathbf{H}$ is an observation operator, $\mathbf{x}_{k,l}^b$ is the k-th background ensemble forecast of CO_2 concentration for the l-th vertical level at the model space, a_l is the value of averaging kernel at the l-th vertical level which is normalized in a way that its vertical sum is equal to unity, and S is a spatial interpolation operator.

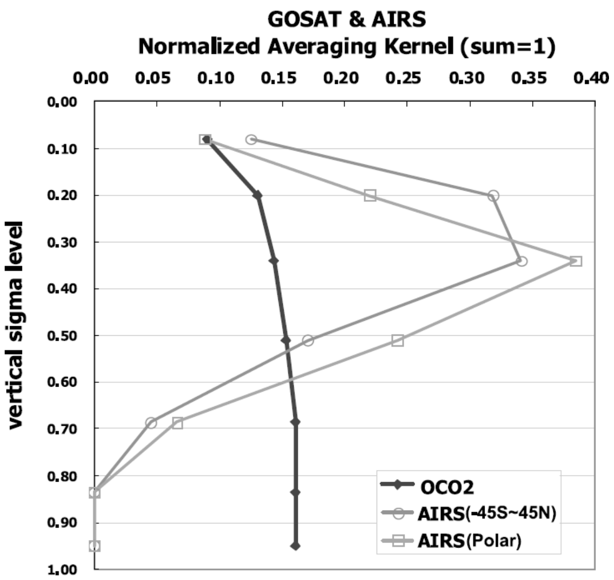


Figure 2. Averaging kernels of GOSAT and AIRS assumed in our simulation experiments, which are normalized to make each of a vertical sum the unity. (Figure 2 of Kang et al. 2012)

4. Results

4.1. Localization of variables

We tested the impact of the localization of variables (LOCvar) under a simple emission scenario with a simulation experiment where the only carbon emission is anthropogenic fossil fuel emission constant in time (Figure 3a). When assimilating synthetic data of ground based observations and GOSAT data for atmospheric CO2 and radiosonde data for weather variables, the experiment with LOCvar (Figure 3b) greatly outperforms the one including a fully multivariate error covariance (Figure 3c). This is because the system with LOCvar includes the important transport errors in the atmospheric CO2 analysis but eliminates the sampling errors introduced by computing the essentially zero error covariance between (C, CF) and (T, q, Ps), as well as the covariance between CF and (U, V).

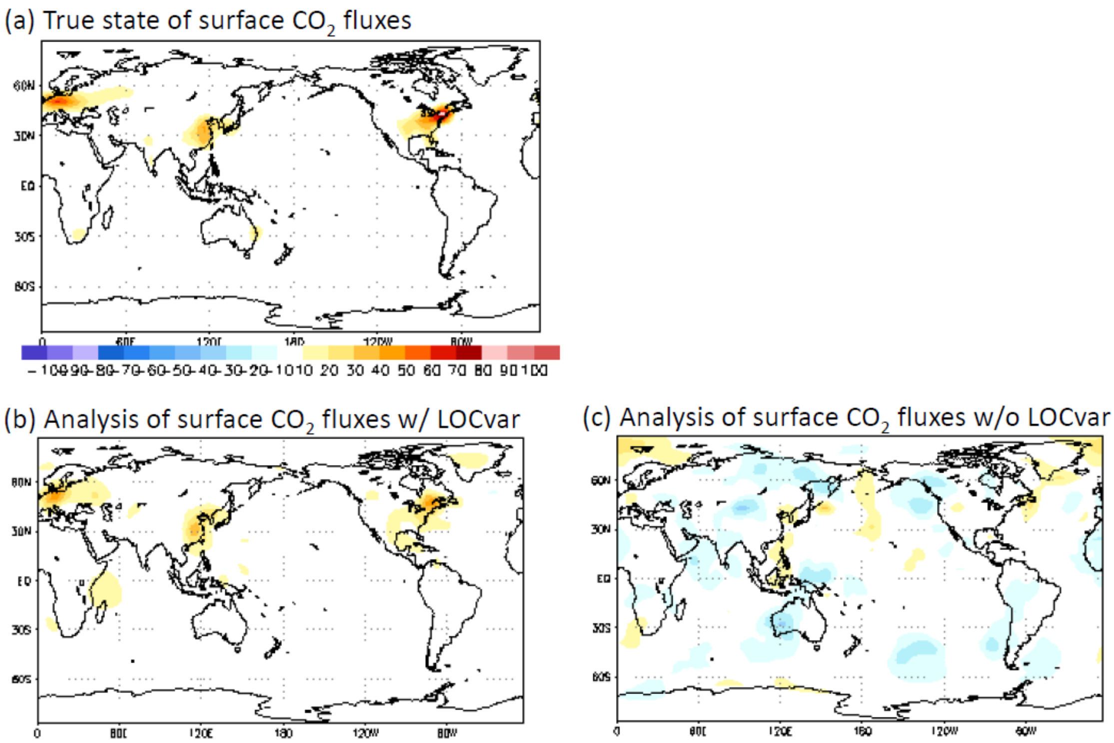


Figure 3. (a) True state of surface CO_2 fluxes and the analysis (b) with LOCvar and (c) without LOCvar (which means using the analysis with fully multivariate background error covariance as in Fig. 1a). (Units are 10^{-9} kg/m^2/s)

4.2. Inflation and vertical localization of column mixing CO2 data

We upgraded the UMD-UCB CDA system with more advanced EnKF data assimilation techniques such as advanced inflation methods and spatial localization for column mixing data (Kang et al. 2012). As a result, we have obtained very promising CF estimation with seasonal variations, indicating success in estimating time-evolving two-dimensional parameter (Figure 4). The results of the simulation experiments (with a perfect model) assume that the true CF includes both constant fossil fuel emission as well as time-evolving, biosphere driven terrestrial carbon fluxes from CASA (Randerson et al. 1997), and oceanic flux from Takahashi et al. (2002). Table 1 shows a quantitative comparison among the experiments with and without advanced inflation methods and vertical localization technique. FixedM experiment indicates a standard EnKF inflation setting which has constant multiplicative inflation in time and space, AdaptM indicates adaptive multiplicative inflation that updates multiplicative inflation factor in time and space every analysis step, and Addi is additive inflation as described before. LowLevel means the result from vertical localization of GOSAT column mixing CO2 data and FullColumn means the result without the vertical localization. More detailed descriptions of each method and the corresponding results can be found in Kang et al. 2012.

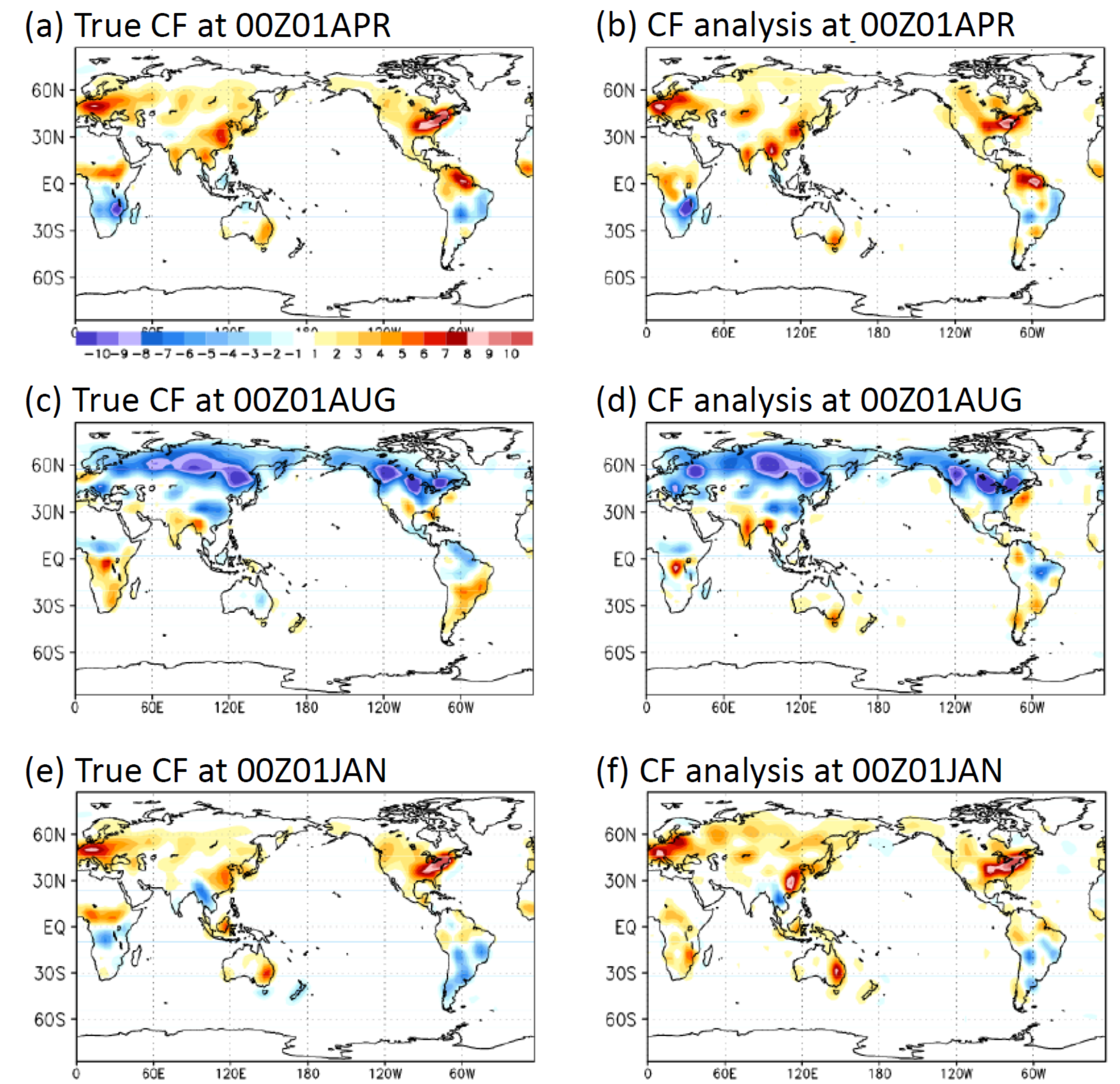


Figure 4. True state of surface CO2 fluxes (left) and its analysis (right) after three month of data assimilation (top), after seven months (middle), and one-year (bottom) of data assimilation. (Units are 10^{-8} kg/m^2/s)

Table 1. RMS errors of analyses in surface CO_2 fluxes from the sensitivity experiments to the inflation method and vertical localization of column CO_2 data: column ONE YEAR includes the RMSEs averaged over the one-year analysis period except for the first three months, and column SUMMER includes those only for July and August. (,) indicates the RMS errors over land and ocean respectively. Here, the RMS errors are computed over every grid point at the surface with respect to the true CO2 fluxes. (Table from Kang et al. 2012)

Inflation	Vertical localization	RMSE of surface CO_2 fluxes (gC/m^2/yr): Global mean (land, ocean)	
		ONE YEAR	SUMMER
FixedM	LowLevel	**185.76** (294, 118)	**218.44** (370,114)

FixedM + Addi	LowLevel	**159.10** (249, 106)	**155.66** (244,101)
AdaptM + Addi	LowLevel	**114.38** (179, 75)	**115.24** (182,73)
AdaptM + Addi	FullColumn	**124.70** (213, 63)	**122.12** (208, 62)

4.3. Short window vs. Long windows

In order to test the impact of shorter vs. longer windows, we performed the CO_2 analysis with an extended (three-week) analysis window. An Ensemble Kalman Smoother has been applied to constrain CF, assimilating future observations within a three-week analysis window. Note that LETKF-C (with a short window) incorporates meteorological variables in analyzing CO_2 variables in order to include wind uncertainties in CO_2 analysis. However, with an extended three-week window, we cannot include wind uncertainty in CO_2 analysis because such a long assimilation window leads to nonlinear growth of ensemble perturbations and three weeks are long enough for the weather variables to lose predictability. Indeed, the results from the simultaneous analysis with atmospheric variables in the long assimilation window experiment are greatly degraded compared to LETKF-C results (not shown). Thus, we have performed carbon-univariate data assimilation system with a three-week analysis window (LongWindow, hereafter), which excludes ensemble of meteorological variables and assimilates only atmospheric CO_2 observations for analyzing atmospheric CO_2 and surface CO_2 fluxes. LongWindow uses the six-hour analysis mean of wind fields from the LETKF-C in order to transport atmospheric CO_2. This LongWindow system is similar to many previous studies in several ways: 1) wind information is given by an independent analysis, 2) there is no explicit treatment of transport errors during CO_2 analysis, and 3) it uses a long assimilation window.

Figure 5 shows that, initially, the (3-week) LongWindow has larger analysis increments because it assimilates observations for three weeks from the initial time to constrain CF, while the (6 hours) LETKF-C assimilates the observations available only at the analysis time so that the initial increment is very small. Broad negative errors appear over the Southern Hemisphere in LongWindow because the random initial state of lower tropospheric CO_2 has positive errors over the region. The LongWindow analysis tries to reduce those errors using three weeks observations, and thus broad negative CF is estimated while LETKF-C has small CF analysis increments using six-hour observations.

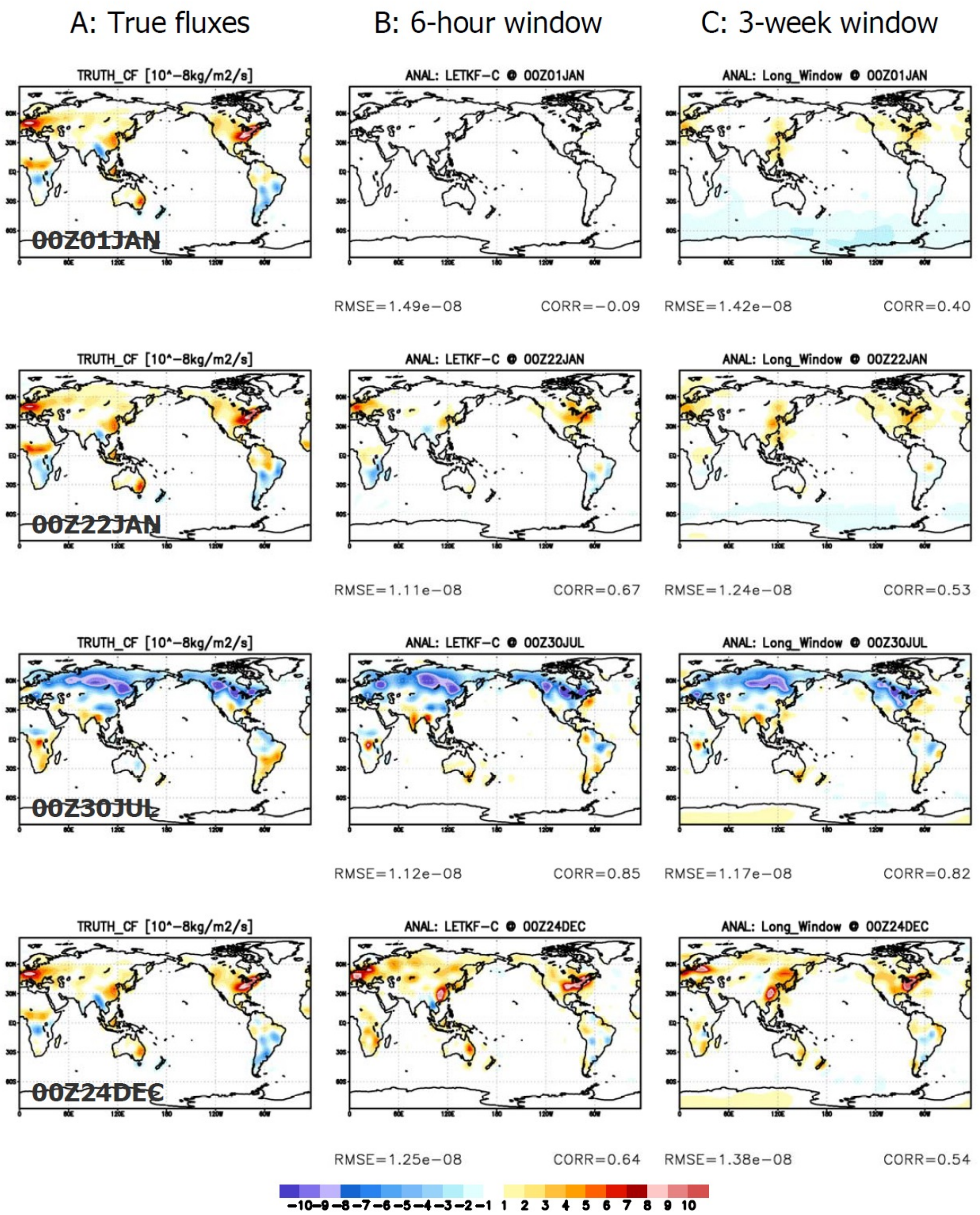


Figure 5. True surface CO_2 fluxes (left) and their analyses from LETKF-C (middle) and LongWindow (right) on January 1st (top), January 22nd (the second row), July 30th (the third row), December 24th (bottom). Unit for the color figures is 10^{-8}kg/m^2/s. Global RMS error and spatial correlation coefficient are included below each analysis plot.

In terms of the amount of assimilated observations, it would be fair to compare the result of LongWindow at 00Z01JAN with that of LETKF-C at 00Z22JAN, showing how different the CF analyses are according to the length of analysis window. Overall, both windows succeed in estimating the evolving

CF, but LETKF-C has more detailed and localized CF estimation than LongWindow. Including CO_2 observations far from the analysis time may not necessarily improve the CF analysis compared to an instantaneous analysis due to the attenuation of detailed information as discussed in *Enting* (2002), Figure 1.3. A shorter assimilation window reduces the attenuation of observed CO_2 information because the analysis system can use near-surface CO_2 observations before the transport of CO_2 blurs out the essential information of near-surface CO_2 forcing (schematic Figure 6). Note that the short assimilation window is also allowed to use real-time wind uncertainty information within LETKF-C system, whereas LongWindow could not benefit from it.

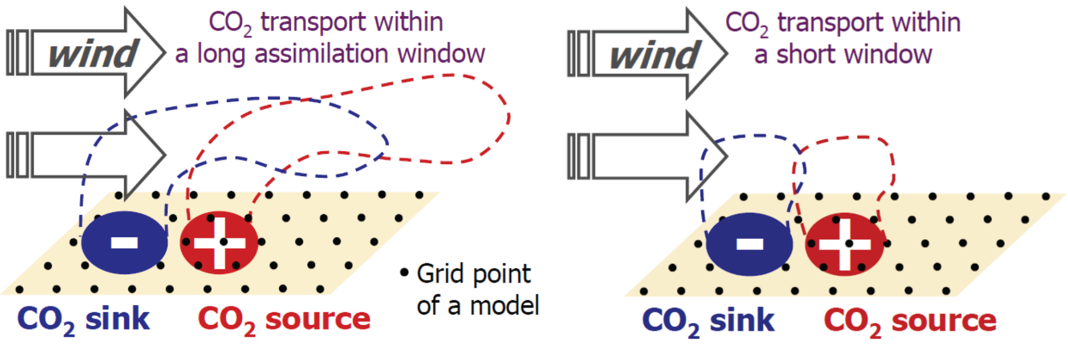


Figure 6. Schematic plot of carbon cycle data assimilation system with long assimilation window (left), and with short window (right): When we attempt to estimate surface carbon fluxes by assimilating atmospheric CO2 observations, a short window reduces the attenuation of observed CO2 information because the analysis system can use the strong correlation between C and CF before the transport of atmospheric CO2 blurs out the essential information of surface CO2 forcing. Thus, we cannot reflect the optimal correlation between C and CF within a long assimilation window, which can introduce sampling errors into the EnKF analysis.

5. Summary and discussion

We developed a simultaneous analysis system of CO_2 and meteorological variables using the LETKF data assimilation method. Through the use of OSSEs, we found that including advanced assimilation techniques makes it possible to estimate time-evolving surface CO_2 fluxes, even without direct observations or land surface models providing prior information.

The use of a simultaneous atmospheric and CO_2 analysis with short windows links errors in surface CO_2 fluxes to the information about near surface atmospheric CO_2 concentrations, accounting for uncertainties in the wind fields used for driving the transport model in the flux inversion, and before the signal gets blurred by nonlinear transport and turbulence effects. It estimates the background error covariance among the variables of atmospheric CO_2, surface CO_2 fluxes and wind, without having to run a separate transport model used as an observation operator in most previous studies. The "localization of variables" introduced in K11 accounts for wind uncertainties during the analysis cycle of carbon variables while zeroing out the covariance between variables that are not physically coupled. This substantially reduces sampling errors and improves the estimation of CO_2 fluxes. We also find that the accurate representation of background uncertainties with advanced inflation methods is essential in order to obtain good estimations of the carbon variables. Advanced inflation methods and vertical localization methods are important for the success of the UMD-UCB CDA in estimating time-evolving surface carbon fluxes.

LETKF-C uses a much shorter analysis window (6 hours) than previous studies in order to account for time-evolving error covariance between wind and CO_2 as well as to avoid the attenuation due to turbulent transport of the observed CO_2 information. While a longer assimilation window (3 weeks) also succeeds in estimating the evolving surface carbon fluxes, we find that it has less spatial structure than the short 6-hour window. This is because atmospheric CO_2 observations for several weeks ahead can contain surface CO_2 forcing information far from the analysis point, so that the analysis system loses information on the CO_2 transport due to errors in the transport model and the wind analysis. The results show that the flux inversion with a long-time assimilation window is not as accurate as the one obtained with a six-hour assimilation window, particularly the smaller-scale structures. In other words, a longer window allows more observations to be used for constraining surface CO_2 fluxes but the loss of information (Figure 1.3 of Enting, 2002) makes surface carbon flux estimates with long windows somewhat worse than analyses using 6-hour windows.

In a perfect model scenario, our results indicate that carbon cycle data assimilation system does not require using *a priori* information, including information on initial conditions. However, we have not accounted for error sources such as model errors of meteorological variables, the diurnal cycle of carbon fluxes, and observation biases, so that we will face additional difficulties when using real observations. These difficulties may be overcome by applying additional advanced methodologies, such as the use of information about the carbon fluxes climatology, more accurate models of surface fluxes, estimation and correction of model/observation errors [*Danforth et al. 2007; Li et al., 2009a and 2009b*]. We are currently working on a more realistic system based on the NCAR Community Atmosphere Model (CAM, version 5) and real observations following the work of *Liu et al.* [2011, 2012]. Liu et al. (2011) quantified the non-negligible impact of meteorology uncertainty on CO_2 forecast with the coupled LETKF-CAM. Liu et al. (2012) further demonstrated the system performance in estimating CO_2 concentrations by simultaneously assimilating meteorology observations and the mid-troposphere Atmospheric InfraRed Sounder (AIRS) CO2 observations with the LETKF-CAM. It is also possible to incorporate *a priori* information and models into our analysis system to get a physical update on the background state of carbon fluxes, such as done in CarbonTracker (Peters et al., 2007). This should improve the results further and reduce the difficulties associated with the assimilation of real observations with an imperfect model.

Finally, we note that the methodology that we have developed in this study can be also applied to any type of surface fluxes, not just carbon fluxes. In particular, we have obtained similarly good results in estimating surface fluxes of sensible and latent heat assuming a realistic coverage of AIRS temperature and moisture retrievals. Therefore, this approach could have a broader impact on various applications estimating time-evolving parameters within EnKF data assimilation system.

References

Baker, D. F., S. C. Doney, and D. S. Schimel (2006), Variational data assimilation for atmospheric CO2, Tellus, Ser. B, 58, 359–365, doi:10.1111/ j.1600-0889.2006.00218.x.

Baker, D. F., H. Bösch, S. C. Doney, D. O'Brien, and D. S. Schimel (2010), Carbon source/sink information provided by column CO2 measurements from the Orbiting Carbon Observatory, Atmos. Chem. Phys., 10, 4145–4165, doi:10.5194/acp-10-4145-2010.

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

Danforth, C. M., E. Kalnay, and T. Miyoshi (2007), Estimating and correcting global weather model error, Mon. Weather Rev., 135, 281–299, doi:10.1175/MWR3289.1.

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

Feng, L., P. I. Palmer, H. Bösch, and S. Dance (2009), Estimating surface CO2 fluxes from space-borne CO2 dry air mole fraction observations using an ensemble Kalman filter, Atmos. Chem. Phys., 9, 2619–2633, doi:10.5194/acp-9-2619-2009.

Greybush, Steven J., E. Kalnay, T. Miyoshi, K. Ide, and B. Hunt, 2011: Balance and Ensemble Kalman Filter Localization Techniques. *Mon. Wea. Rev.*, 139, 511-522.

Hunt, B. R., E. Kostelich, and I. Szunyogh (2007), Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter, Physica D, 230, 112–126, doi:10.1016/j.physd.2006.11.008.

Kang, J.-S., E. Kalnay, J. Liu, I. Fung, T. Miyoshi, and K. Ide (2011), "Variable localization" in an ensemble Kalman filter: Application to the carbon cycle data assimilation, J. Geophys. Res., 116, D09110, doi:10.1029/2010JD014673.

Kang, J.-S., E. Kalnay, T. Miyoshi, J. Liu, and I. Fung (2012), Estimation of surface carbon fluxes with an advanced data assimilation methodology, J. Geophys. Res., 117, D24101, doi:10.1029/2012JD018259.

Li, H., E. Kalnay, and T. Miyoshi (2009a), Simultaneous estimation of covariance inflation and observation errors within an ensemble Kalman filter, Q. J. R. Meteorol. Soc., 135, 523–533, doi:10.1002/qj.371.

Li, H., E. Kalnay, T. Miyoshi, and C. M. Danforth (2009b), Accounting for model errors in ensemble data assimilation, Mon. Weather Rev., 137, 3407–3419, doi:10.1175/2009MWR2766.1.

Liu, J., I. Fung, E. Kalnay, and J.-S. Kang (2011), CO2 transport uncertainties from the uncertainties in meteorological fields, Geophys. Res. Lett., 38, L12808, doi:10.1029/2011GL047213.

Liu, J., I. Fung, E. Kalnay, J.-S. Kang, E. T. Olsen, and L. Chen (2012), Simultaneous assimilation of AIRS XCO2 and meteorological observations in a carbon climate model with an ensemble Kalman filter, J. Geophys. Res., 117, D05309, doi:10.1029/2011JD016642.

Michalak, A. M. (2008), Adapting a fixed-lag Kalman smoother to a geostatistical atmospheric inversion framework, Atmos. Chem. Phys., 8, 6789–6799.

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 CO2 surface fluxes from atmospheric trace gas observations, J. Geophys. Res., 110, D24304, doi:10.1029/2005JD006157.

Peters, W, A R. Jacobson, C. Sweeney, A. E. Andrews, T. J. Conway, K. Masarie, J. B. Miller, L. M. P. Bruhwiler, G. Pétron, A. I. Hirsch, D. Worthy, G. van der Werf, J. Randerson, P. Wennberg, M. Krol, and P. Tans (2007), An atmospheric perspective on North American carbon dioxide exchange: CarbonTracker, Proc. Natl. Acad. Sci. U. S. A., 104, 18,925–18,930, doi:10.1073/pnas.0708986104.

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, doi:10.1029/97GB02268.

Rayner, P. J., R. M. Law, C. E. Allison, R. J. Francey, C. M. Trudinger, and C. Pickett-Heaps (2008), Interannual variability of the global carbon cycle (1992–2005) inferred by inversion of atmospheric CO2 and d13CO2 measurements, Global Biogeochem. Cycles, 22, GB3008, doi:10.1029/2007GB003068.

Takahashi, T., et al. (2002), Global sea-air CO2 flux based on climatological surface ocean pCO2, and seasonal biological and temperature effects, Deep Sea Res., Part II, 49, 1601–1622, doi:10.1016/S0967-0645(02) 00003-6.

Parameter Estimation in EnKF: Surface Fluxes of Carbon, Heat, Moisture and Momentum

Ji-Sun Kang, **Eugenia Kalnay**, Takemasa Miyoshi, Junjie Liu, Inez Fung, Kayo Ide, Brian Hunt, and many collaborators at the

University of Maryland **Weather-Chaos** group

Outline

- **Motivation:** Why did we decide to estimate carbon fluxes as parameters, not using traditional inversion?
- **LETKF:** brief introduction.
- **Simultaneous assimilation** of carbon and meteorological observations
- **Parameter estimation** with LETKF allows us to estimate surface fluxes.
- **Advanced methods:** "variable localization", vertical *physical* localization, one-way covariance coupling, covariance inflation.
- **Are short or long assimilation windows better?** We use 6hr windows, inversion methods use months, not hours.
- **Wouter Peters** developed the EnKF-based Carbon Tracker. He commented on our approach.

Results

Motivation

Unlike traditional carbon flux inversions,

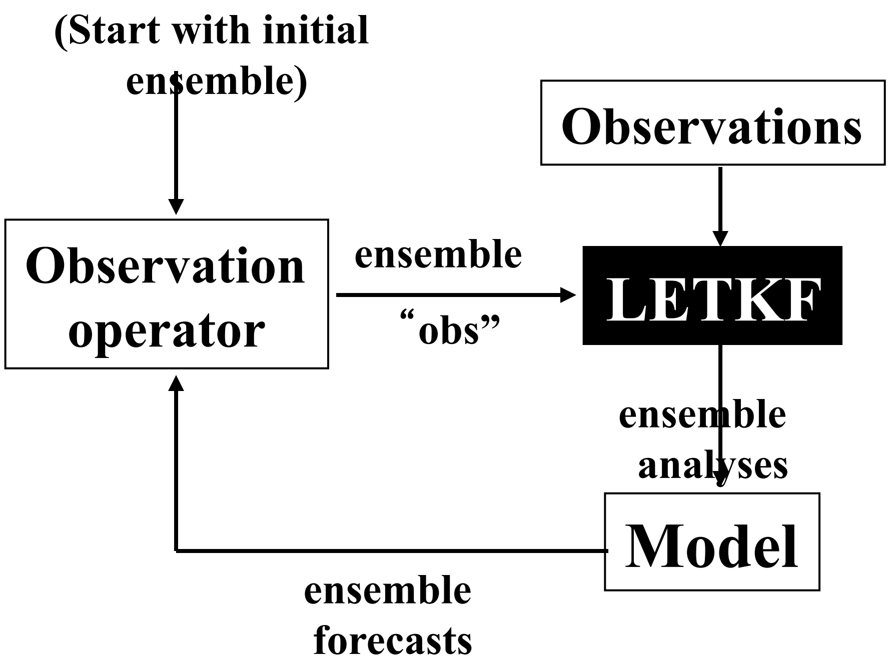
- We are estimating carbon fluxes as parameters
- Assimilating simultaneously carbon and atmospheric u,v,T,q,p_s
- Using short windows (6hr, not many months)

Inez Fung had the vision (~1990's) that data assimilation was the best approach to estimate carbon fluxes. Inez and I got a DOE grant to test this methodology (~2008)

It took us several years to make it work (Ji-Sun Kang's thesis 2009, JGR 2011, 2012). Also Junjie Liu, now at JPL.

For "perfect model" simulations we recover the true fluxes accurately at model grid resolution, without any *a priori* information.

4D-Local Ensemble Transform Kalman Filter (Ott et al, 2004, Hunt et al, 2004, 2007)

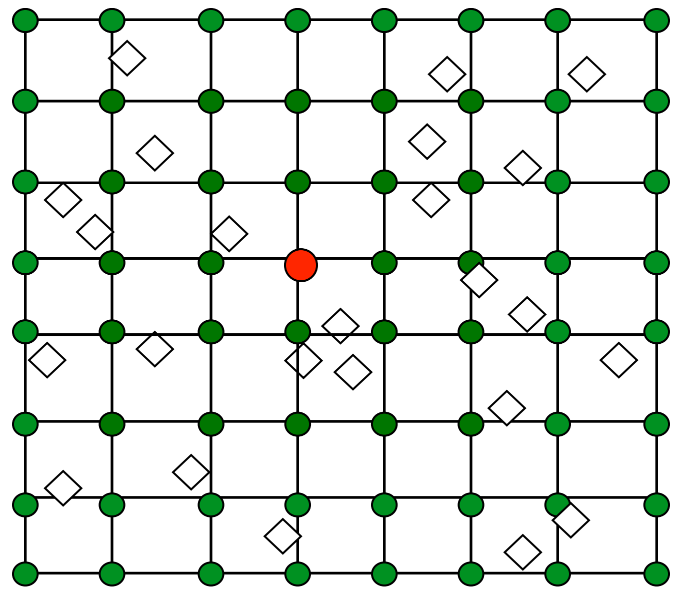


- **Model independent (black box)**
- **No adjoint needed**
- **4D LETKF extension**
- **Obs. assimilated simultaneously at each grid point**
- **LETKF computes the weights for the ensemble forecasts explicitly**

Localization based on observations

Perform data assimilation in a local volume, choosing observations

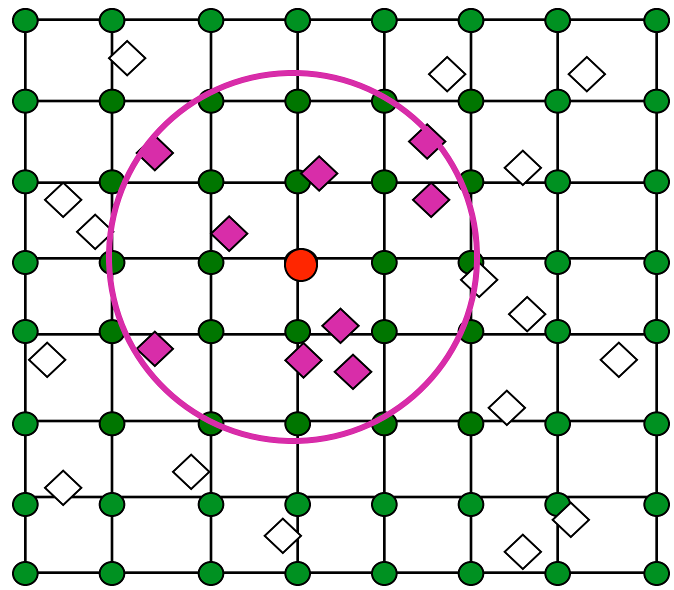
The state estimate is updated at the central grid red dot



Localization based on observations

Perform data assimilation in a local volume, choosing observations

The state estimate is updated at the central grid red dot



All observations (purple diamonds) within the local region are assimilated

The LETKF algorithm can be described in a single slide!

Local Ensemble Transform Kalman Filter (LETKF)

Globally:

Forecast step: $\mathbf{x}^b_{n,k} = M_n\left(\mathbf{x}^a_{n-1,k}\right)$

Analysis step: construct $\mathbf{X}^b = \left[\mathbf{x}^b_1 - \overline{\mathbf{x}}^b \mid \ldots \mid \mathbf{x}^b_K - \overline{\mathbf{x}}^b\right];$

$$\mathbf{y}^b_i = H(\mathbf{x}^b_i);\ \mathbf{Y}^b_n = \left[\mathbf{y}^b_1 - \overline{\mathbf{y}}^b \mid \ldots \mid \mathbf{y}^b_K - \overline{\mathbf{y}}^b\right]$$

Locally: Choose for each grid point the observations to be used, and compute the local analysis error covariance and perturbations in ensemble space:

$$\tilde{\mathbf{P}}^a = \left[(K-1)\mathbf{I} + \mathbf{Y}^T\mathbf{R}^{-1}\mathbf{Y}\right]^{-1};\ \mathbf{W}^a = [(K-1)\tilde{\mathbf{P}}^a]^{1/2}$$

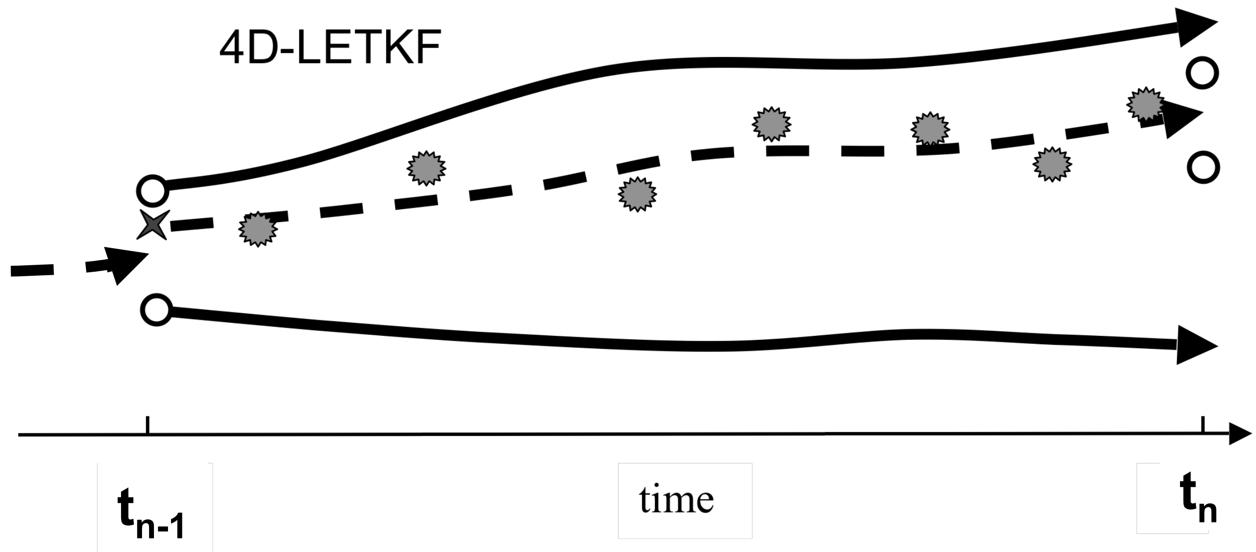
Analysis mean in ensemble space: $\overline{\mathbf{w}}^a = \tilde{\mathbf{P}}^a\mathbf{Y}^{bT}\mathbf{R}^{-1}(\mathbf{y}^o - \overline{\mathbf{y}}^b)$

and add to $\mathbf{W}^a$ **to get the analysis ensemble in ensemble space.**

The new ensemble analyses in model space are the columns of $\mathbf{X}^a_n = \mathbf{X}^b_n\mathbf{W}^a + \overline{\mathbf{x}}^b$ **. Gathering the grid point analyses forms the new global analyses. Note that the the output of the LETKF are analysis weights** $\overline{\mathbf{w}}^a$ **and perturbation analysis weight matrices** $\mathbf{W}^a$

These weights multiply the ensemble forecasts.

No-cost LETKF smoother (×): apply at t_{n-1} the same weights found optimal at t_n. It works for 3D- or 4D-LETKF



The no-cost smoother makes possible:

- ✓ Quasi Outer Loop (QOL)
- ✓ "Running in place" (RIP) for faster spin-up
- ✓ Use of future data in reanalysis
- ✓ Ability to use longer windows and nonlinear perturbations

Kalnay & Yang, 2010, Yang et al, 2012, 2013, Penny et al 2013

Estimation of surface fluxes as evolving parameters

Work of Ji-Sun Kang (now at KIAPS), with Kalnay, Liu and Fung.

(Kang et al., 2011, JGR, Kang et al., 2012, JGR)

- important for carbon cycle
- surface fluxes of heat, moisture, and momentum
- eventually for coupled data assimilation

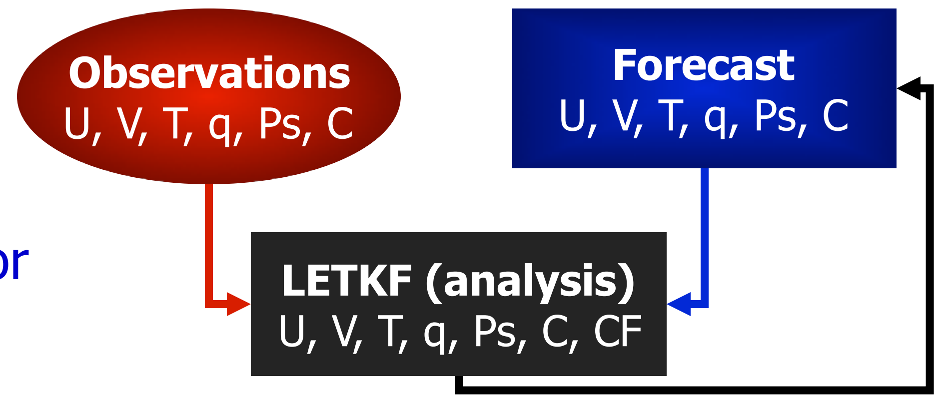
UMD-UCB LETKF-C System

Parameter estimation:
state vector augmentation

$$\mathbf{X}^b = \begin{bmatrix} \mathbf{X} \\ \mathbf{CF} \end{bmatrix}$$

: model state vector (U, V, T, q, Ps, C)

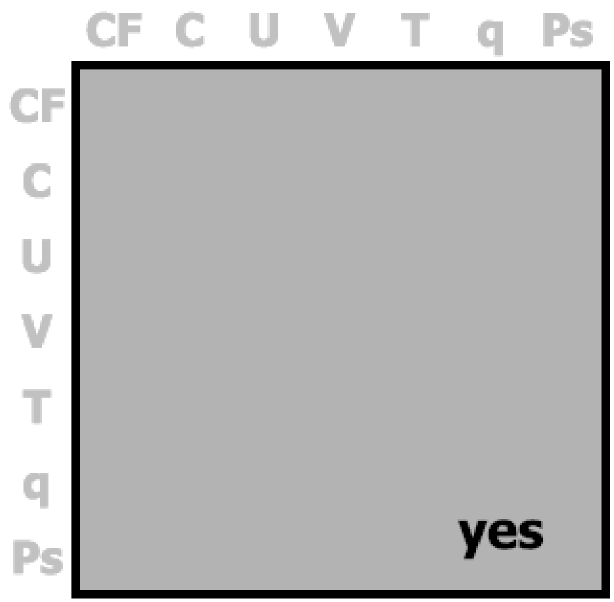
: surface CO_2 flux

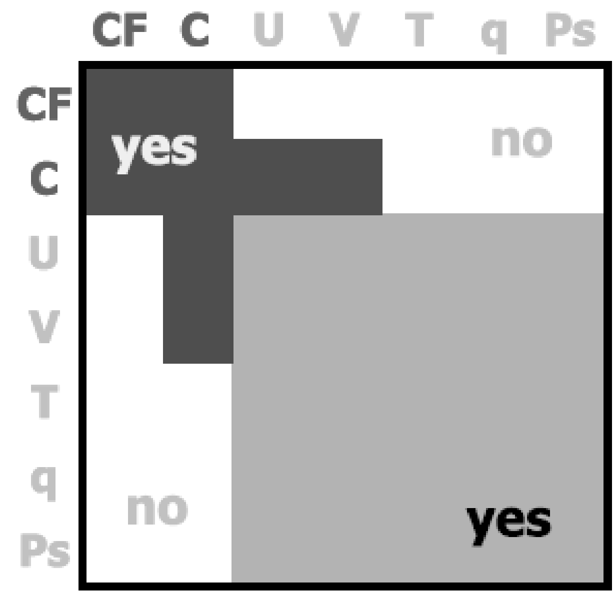


- Append **CF** (surface CO_2 fluxes with no observations)
- Update **CF** as part of the data assimilation process

- **Simultaneous** assimilation of carbon and meteorological variables
  - **Multivariate** analysis with **a localization of the variables** (Kang et al., 2011)
  - Update all variables (including **CF**) every **6** hours

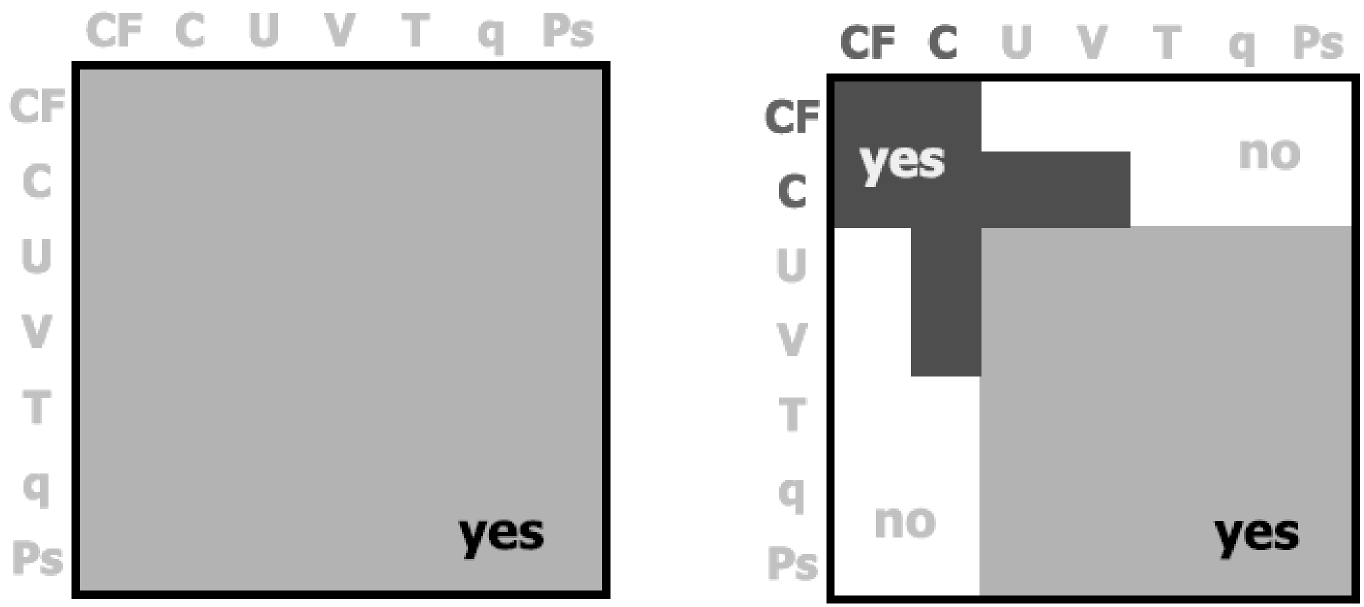
“One-way coupling” (Kang et al, JGR 2011)





The winds improve the CO2, but the CO2 makes the winds worse. So we keep a **one-way coupling** rather than dropping the coupling completely.

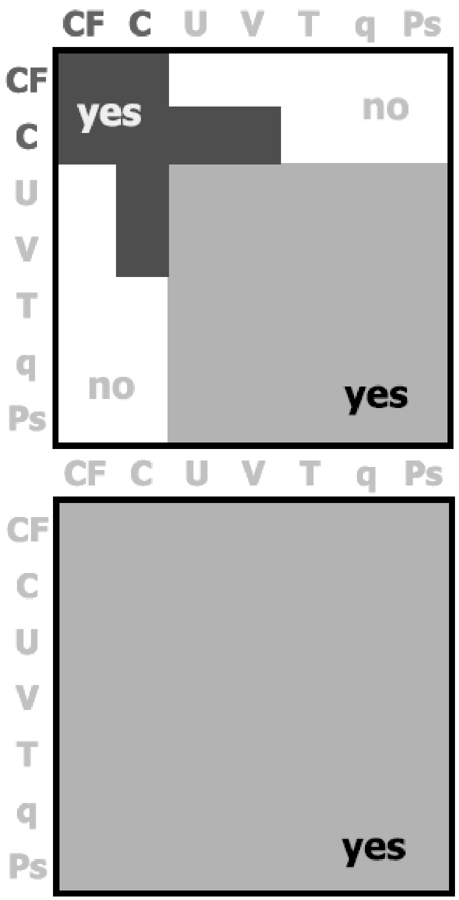
“Localization of variables” (Kang et al, JGR 2011)



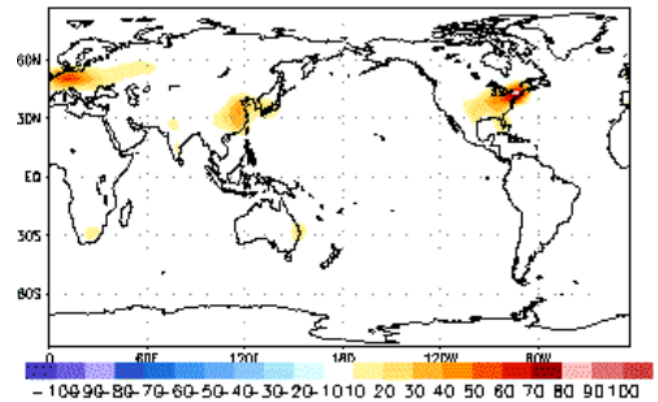
Schematic background error covariance matrix P^b.

➔ **Zeroing out the background error covariance between unrelated variables improves the result of the analysis by reducing sampling errors.**

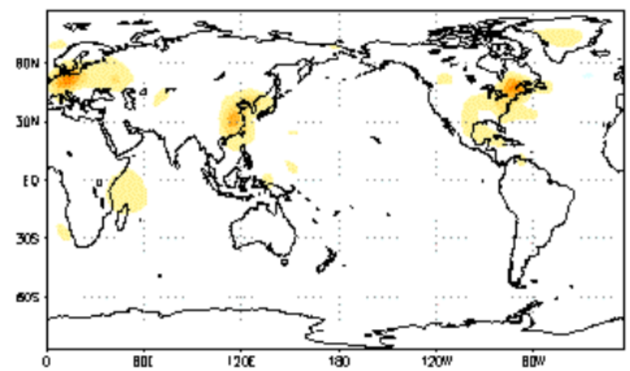
Results: “Variable localization” reduces sampling errors



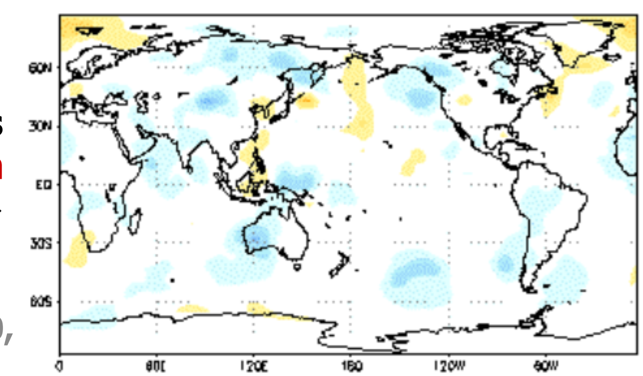
True CO_2 fluxes ➔
(anthropogenic)



Analysis of CO_2 fluxes
with variable localization
➔



Analysis of CO_2 fluxes
without variable localization
➔

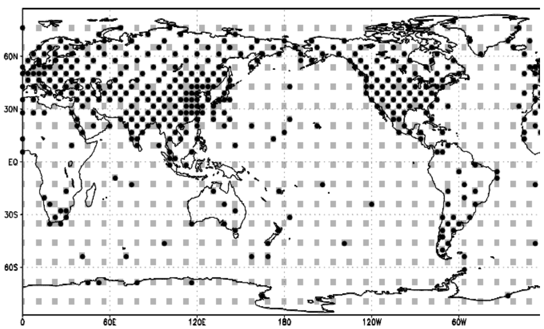


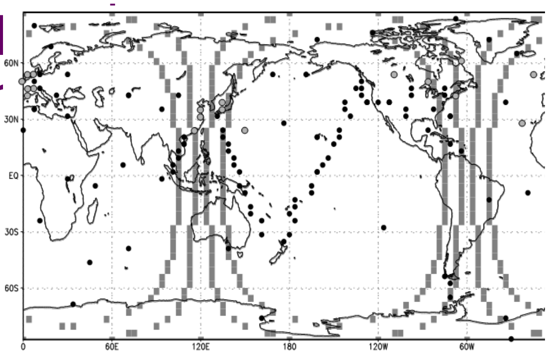
LETKF-C with SPEEDY-C

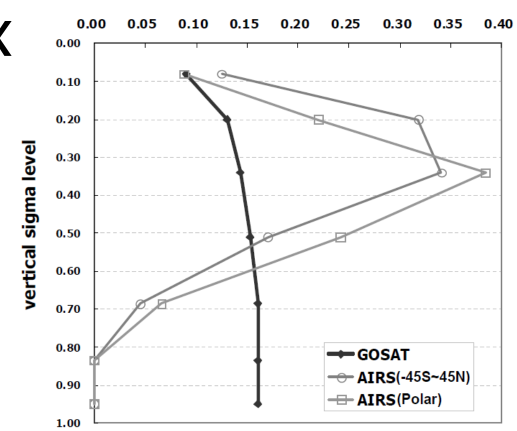
- Model: **SPEEDY-C** (Molteni, 2003; Kang, 2009)
  - Spectral AGCM model with T30L7
  - Prognostic variables: U, V, T, q, Ps, C
    - C (atmospheric CO_2): an inert tracer
  - **Persistence** forecast of Carbon Fluxes (CF), no observations
- **"True" CO2 fluxes: From CASA (Gurney et al, 2004)**
- Simulated observations
  - Rawinsonde observations of U, V, T, q, Ps
  - Ground-based observations of atmospheric CO_2
    - 18 hourly and 107 weekly data on the globe
  - Remote sensing data of column mixing CO_2
    - **AIRS** whose averaging kernel peaks at mid-troposphere
    - **GOSAT** whose averaging kernel is nearly uniform throughout the column
- Initial condition: random (**no *a-priori* information**)
- 20 ensembles

LETKF-C with SPEEDY-C

- Simulated observations
  - Rawinsonde observations of U, V, T, q, Ps
  - Ground-based observations of atmospheric C(
    - 18 hourly and 107 weekly data on the globe
  - Remote sensing data of column mixing CO_2
    - **AIRS** whose averaging kernel peaks at mid-tro|
    - **GOSAT** whose averaging kernel is nearly unifor column
- Initial conditions: random (**no *a-priori* info**)
- 20 ensemble members
- No direct measurement of surface Carbon Flux
- CF only changes through the LETKF:
  persistence forecast.

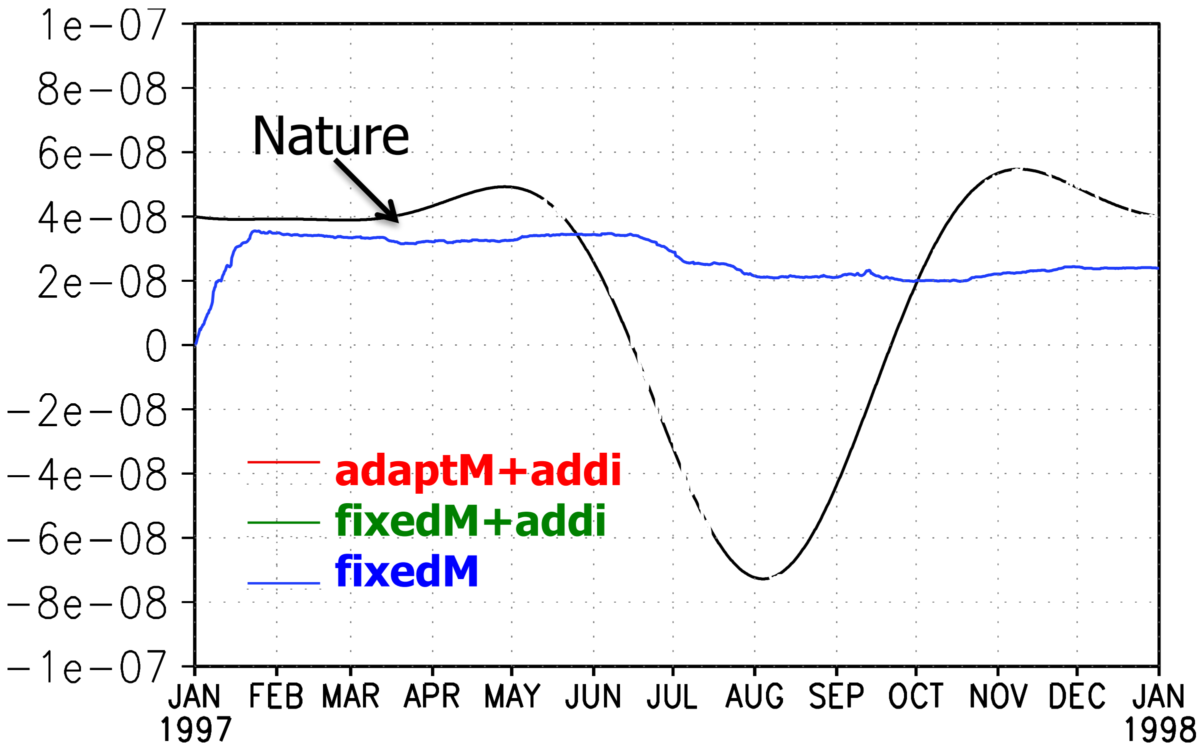




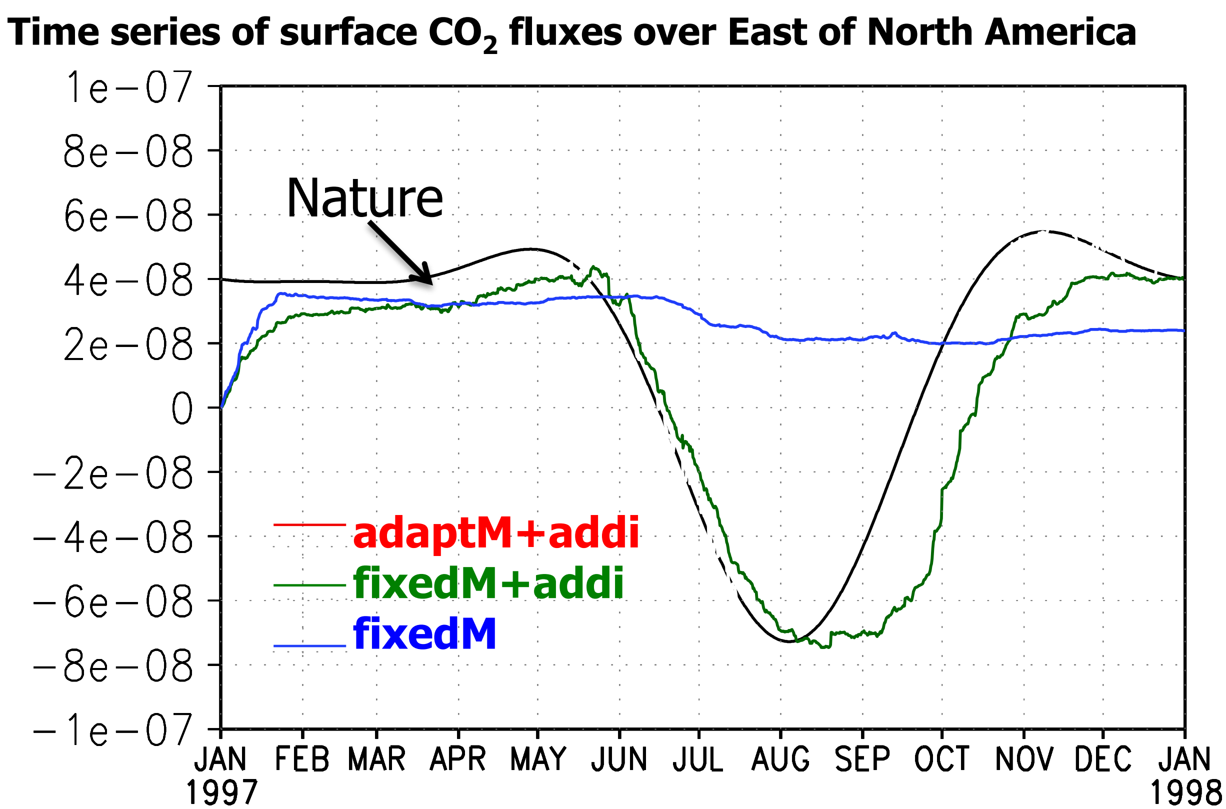


Impact of inflation: fixed multiplicative

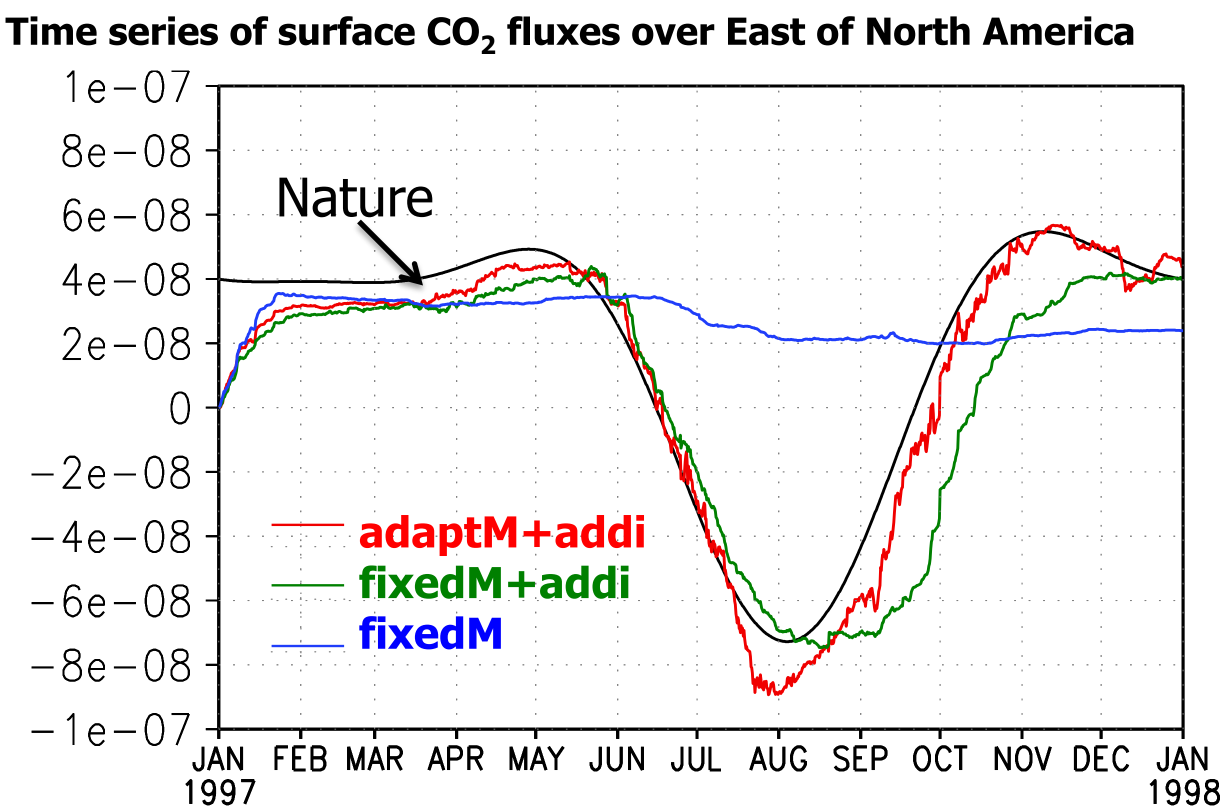
Time series of surface CO_2 fluxes over East of North America



Impact of inflation: fixed multiplicative+additive



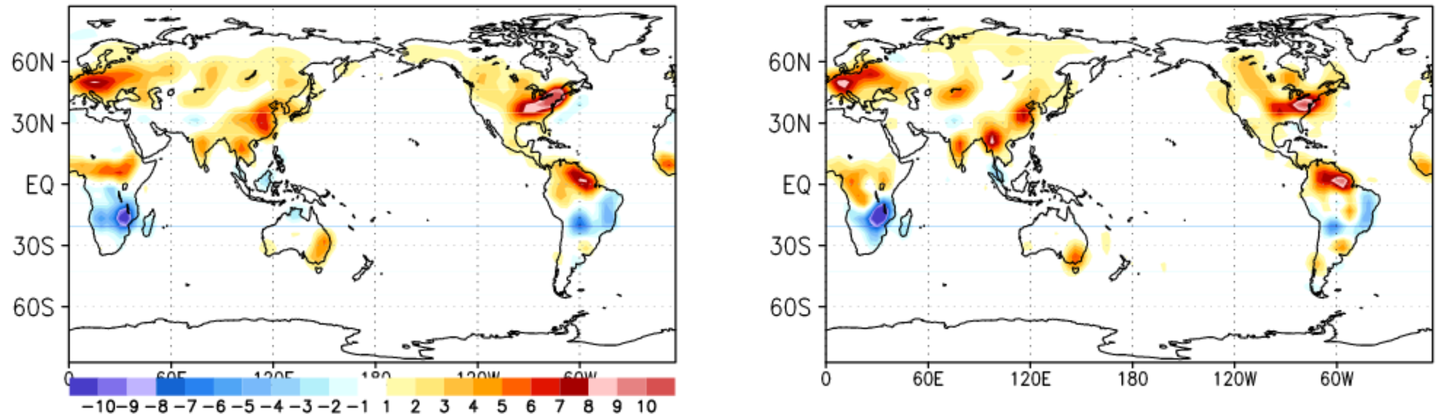
Impact of inflation: Adaptive multiplicative+additive



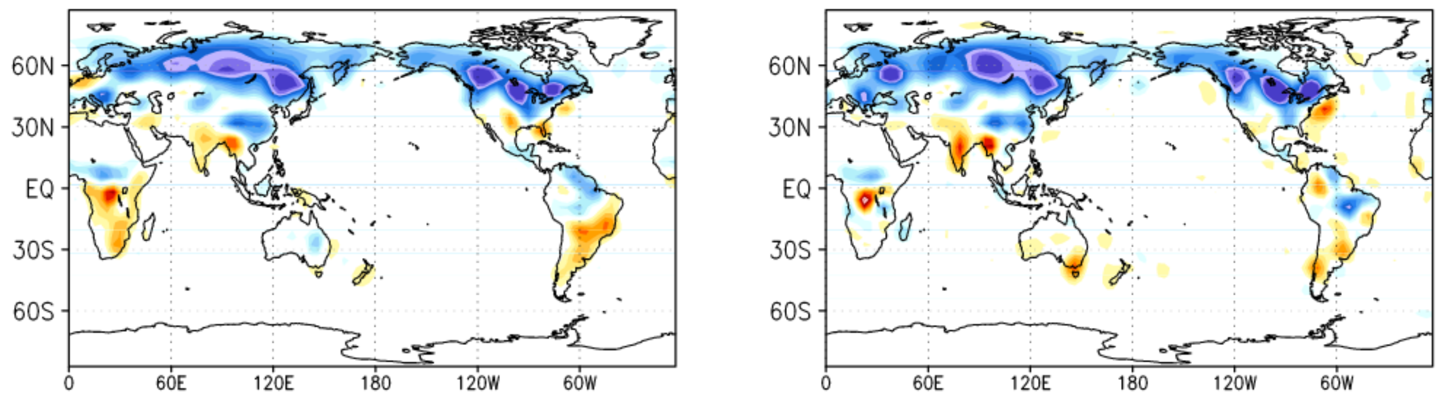
Results

▼ True CF ▼ ▼ Analysis of CF ▼

00Z01APR ▶
After three months of DA

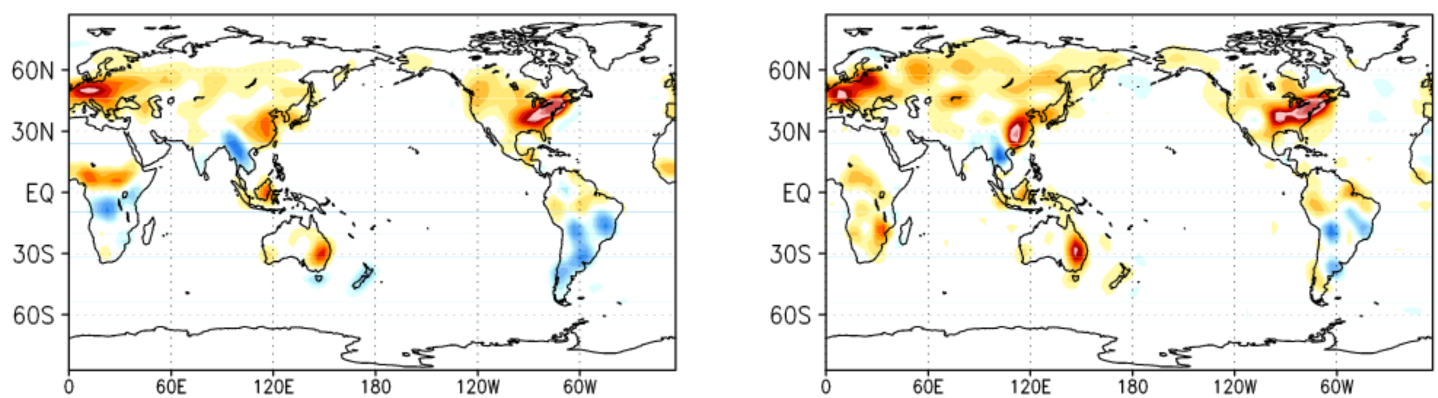


00Z01AUG ▶
After seven months of DA

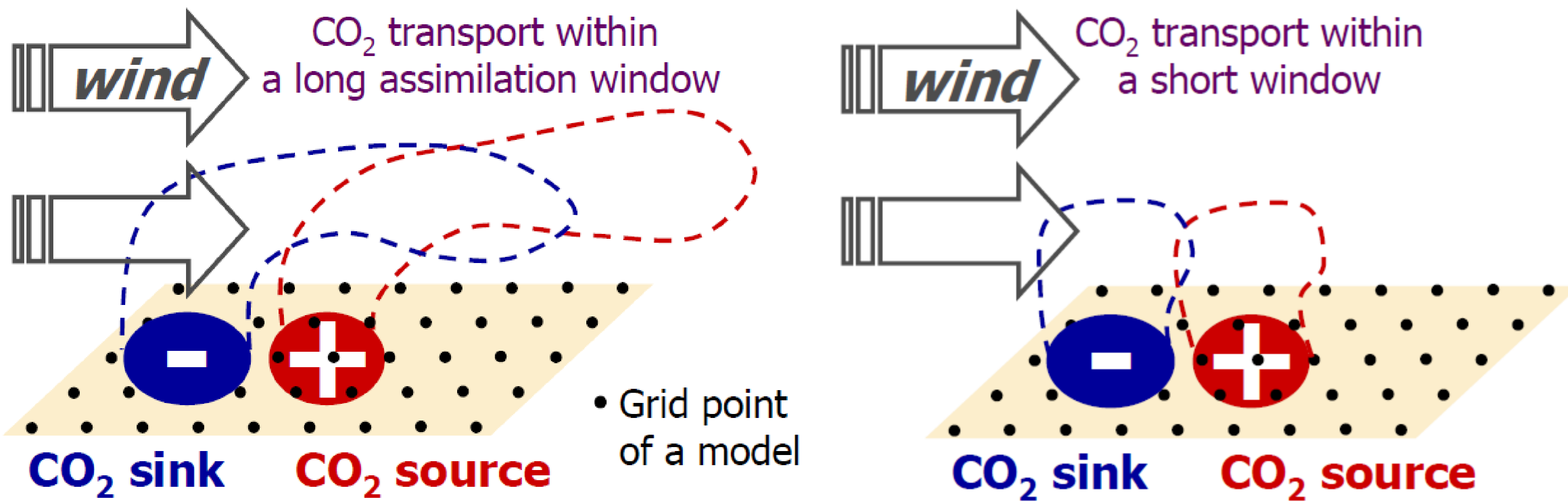


We succeeded in estimating time-evolving CF at model-grid scale!

00Z01JAN ▶
After one year of DA



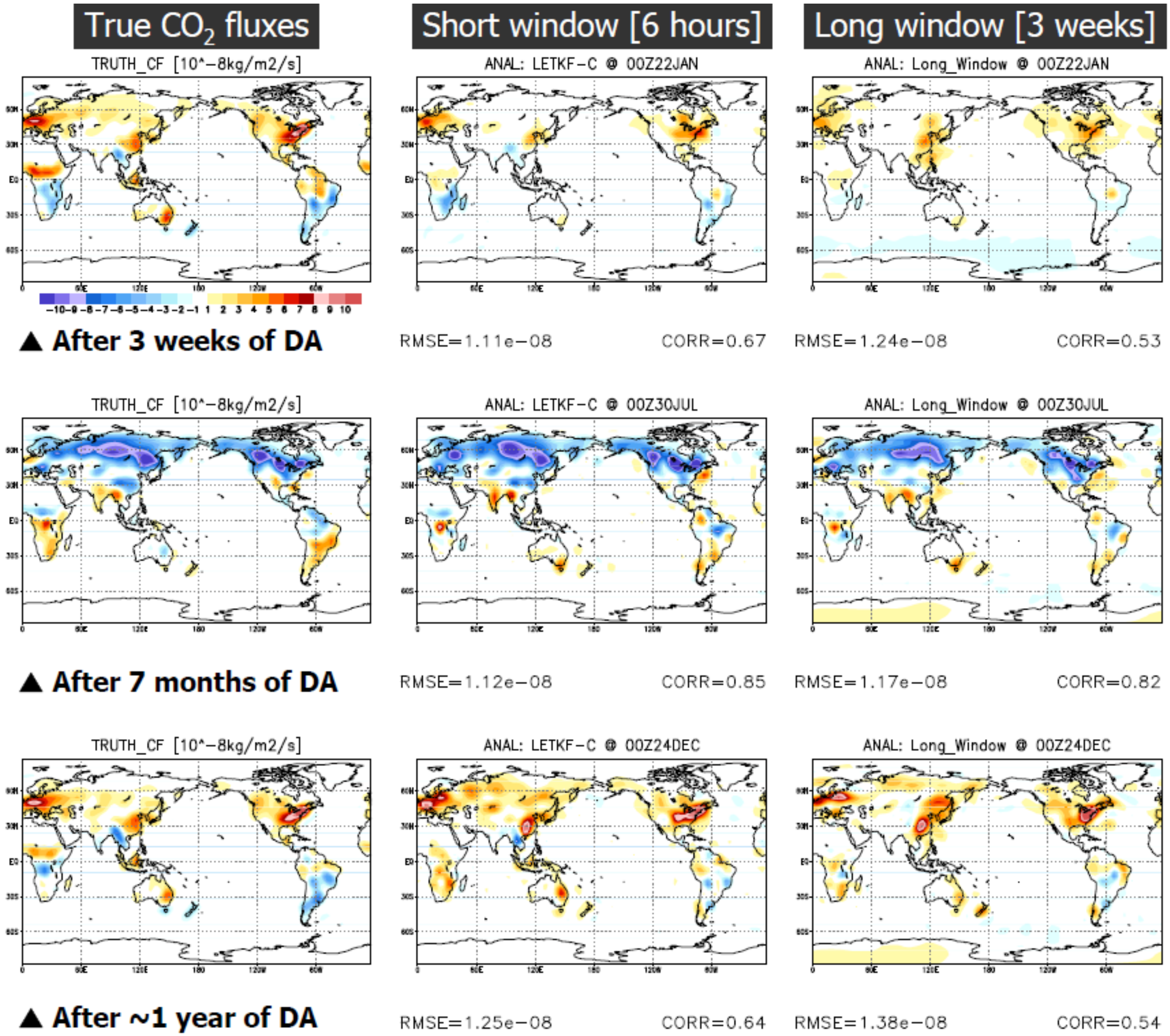
Assimilation windows for Carbon fluxes inversion: current systems use a very long window



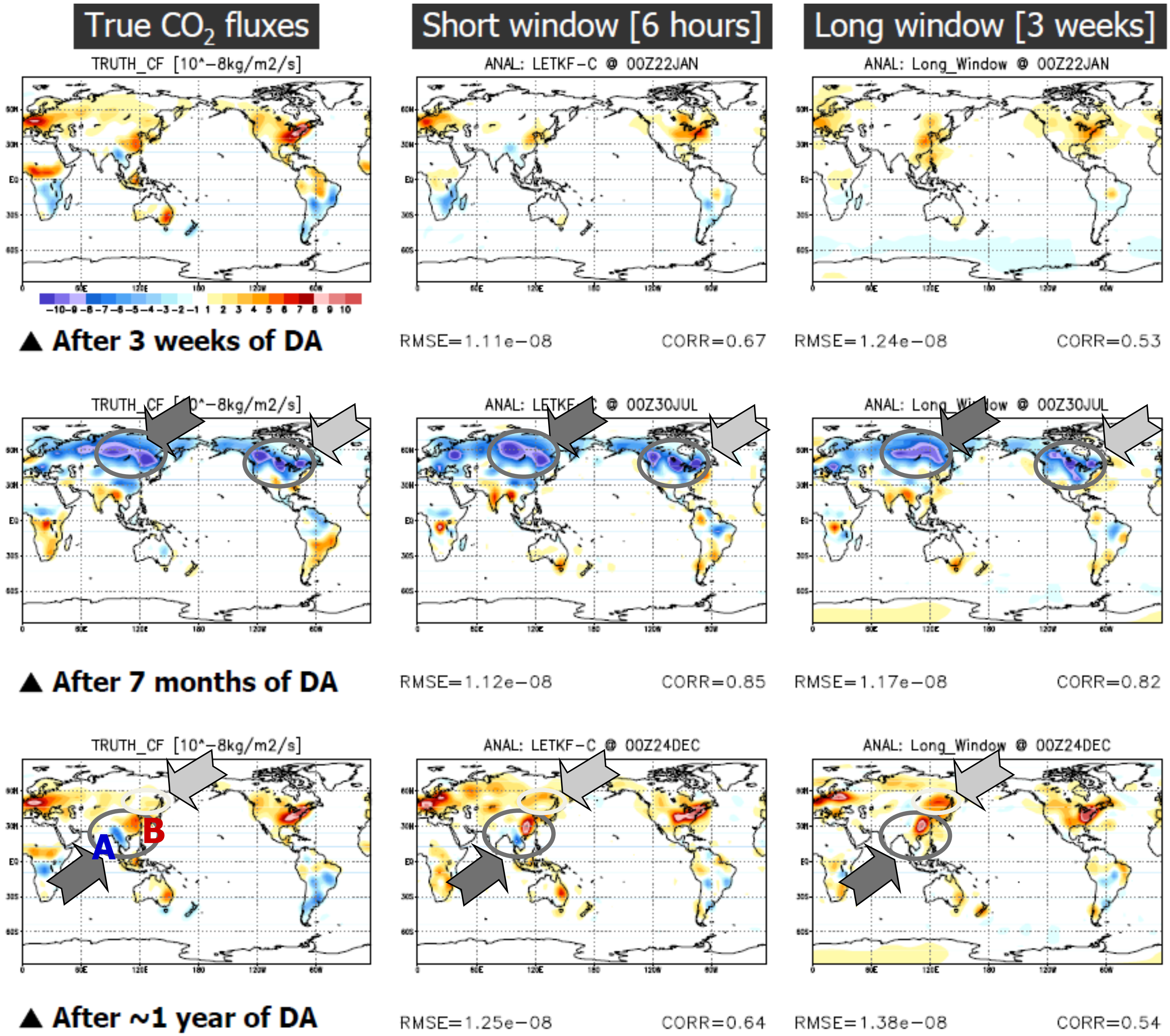
- CO_2 data assimilation system
  - A short assimilation window reduces the attenuation of observed CO_2 information because the analysis system can use the strong correlation between C and CF before the transport of atmospheric CO_2 blurs out the essential information of surface CO_2 forcing
  - We may not be able to reflect the optimal correlation between C and CF within a long assimilation window, which can introduce sampling errors into the EnKF analysis

Long vs. short windows in LETKF-C

- **OSSEs with SPEEDY-C**
  - Realistic observation distributions for meteorological variables and CO_2
    - Rawinsonde observation for (U, V, T, q, Ps)
    - Ground-based observations, AIRS and GOSAT CO_2 mixing ratio for C
- **Experiment 1: Analysis from LETKF-C**
  - Simultaneous analysis with a 6-hour assimilation window
- **Experiment 2: Analysis from a long (3-week) assimilation window**
  - With this long assimilation window, **ensemble perturbations** of meteorological variables become **non-linear** so that we do not include wind uncertainty for CO_2 data assimilation (**Carbon-Univariate DA**)



True CO2 fluxes
Short window [6 hours]
Long window [3 weeks]
TRUTH_CF [10^-8kg/m2/s]
ANAL: LETKF-C @ 00Z22JAN
ANAL: Long_Window @ 00Z22JAN
-10 -9 -8 -7 -6 -5 -4 -3 -2 -1 1 2 3 4 5 6 7 8 9 10
▲ After 3 weeks of DA
RMSE=1.11e-08
CORR=0.67
RMSE=1.24e-08
CORR=0.53
TRUTH_CF [10^-8kg/m2/s]
ANAL: LETKF-C @ 00Z30JUL
ANAL: Long_Window @ 00Z30JUL
▲ After 7 months of DA
RMSE=1.12e-08
CORR=0.85
RMSE=1.17e-08
CORR=0.82
TRUTH_CF [10^-8kg/m2/s]
ANAL: LETKF-C @ 00Z24DEC
ANAL: Long_Window @ 00Z24DEC
▲ After ~1 year of DA
RMSE=1.25e-08
CORR=0.64
RMSE=1.38e-08
CORR=0.54



True CO_2 fluxes
Short window [6 hours]
Long window [3 weeks]
TRUTH_CF [10^-8kg/m2/s]
ANAL: LETKF-C @ 00Z22JAN
ANAL: Long_Window @ 00Z22JAN
▲ After 3 weeks of DA
RMSE=1.11e-08
CORR=0.67
RMSE=1.24e-08
CORR=0.53
ANAL: LETKF-C @ 00Z30JUL
ANAL: Long_Window @ 00Z30JUL
▲ After 7 months of DA
RMSE=1.12e-08
CORR=0.85
RMSE=1.17e-08
CORR=0.82
ANAL: LETKF-C @ 00Z24DEC
ANAL: Long_Window @ 00Z24DEC
A
B
▲ After ~1 year of DA
RMSE=1.25e-08
CORR=0.64
RMSE=1.38e-08
CORR=0.54

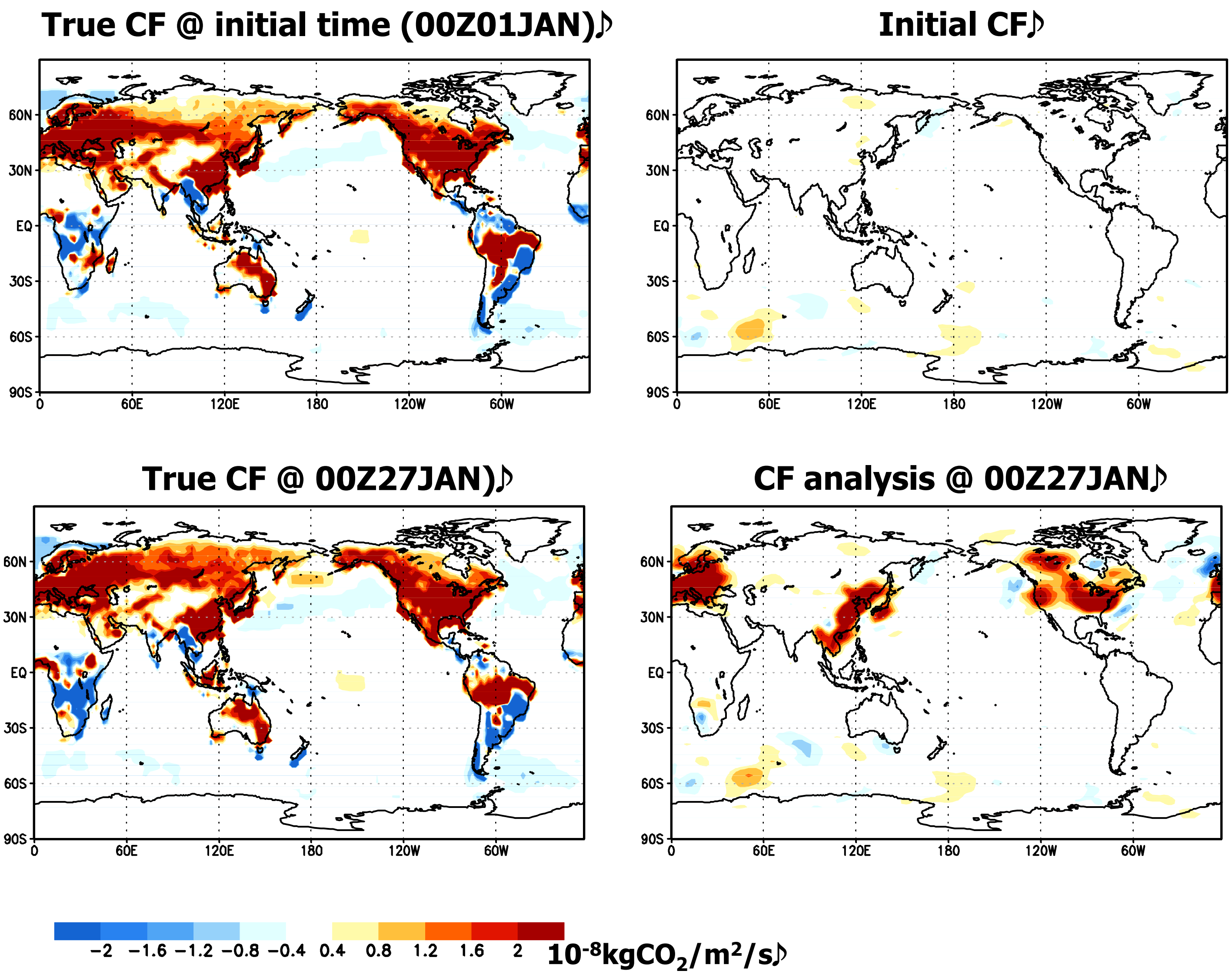
Summary of LETKF-C carbon fluxes

- **Assimilation window**
  - EnKF has better performance with a short window
  - CO_2 observations may be able to provide some information to distant CF, but it becomes blurred (an ill-posed problem).
- Implement LETKF-C on the NCAR CAM model
  - OSSE with realistic observations
  - Very slow (only 26 days)
  - Preliminary results are encouraging

LETKF-C with NCAR CAM3.5

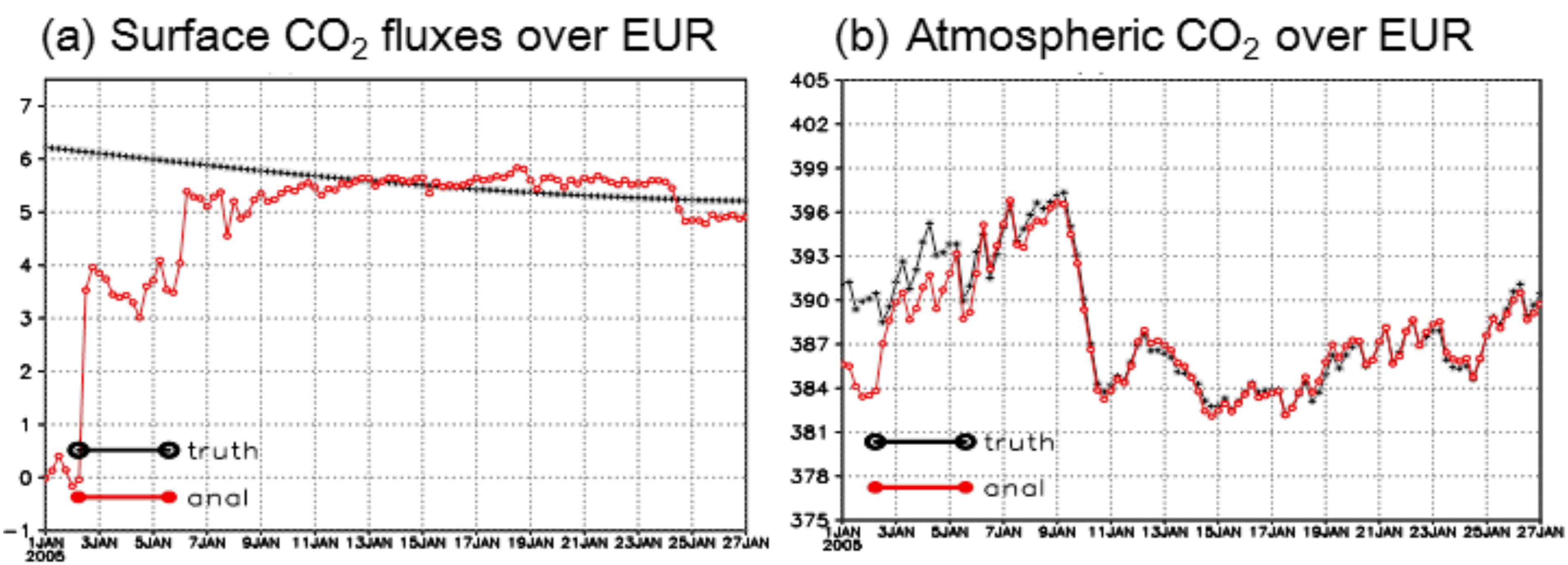
- Model: **CAM 3.5**
  - Finite Volume dynamical core
  - 2.5°×1.9° of horizontal resolution with 26 layers in the vertical
  - C (atmospheric CO_2) is an inert tracer
  - Persistence forecast of CF
- Simulated observations with **real observation coverage**
  - Conventional data for U, V, T, q, Ps
  - Ground-based observations of atmospheric CO_2
    - ~10 hourly and ~100 weekly records on the globe
  - Remote sensing data of column mixing CO_2
    - **AIRS** whose averaging kernel peaks at mid-troposphere
- Initial conditions: random (no *a-priori* information)
- 64 ensembles

LETKF-CAM 3.5 analysis after 26 days



Over a data rich region (Europe)

Time series of surface CO_2 fluxes and atmospheric CO_2 concentrations over Europe

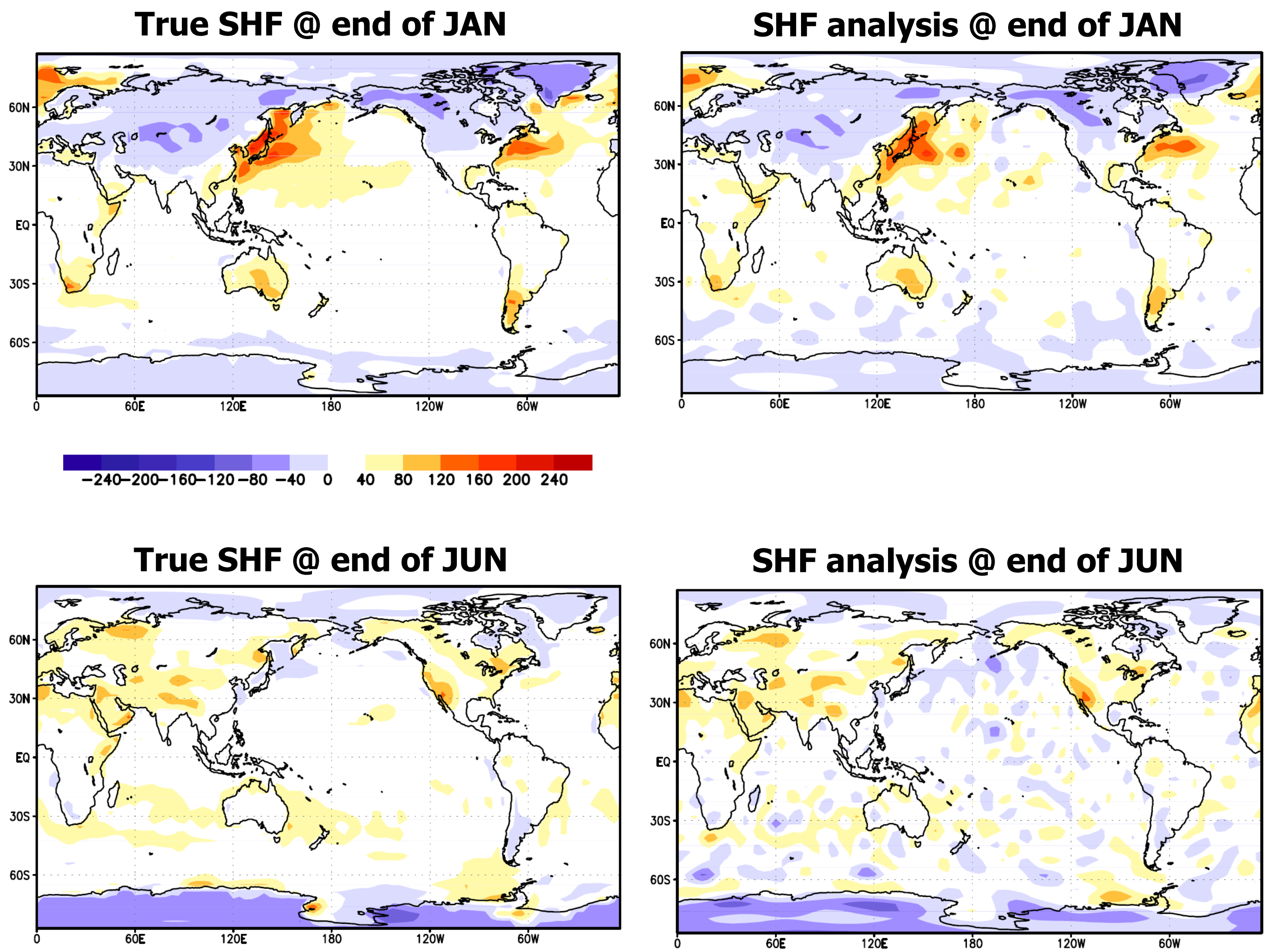


Surface Heat and Moisture Fluxes

- Can we estimate **surface moisture/heat fluxes** by assimilating atmospheric moisture/temperature observations? *We can use the same methodology…*
- OSSEs
  - Nature: SPEEDY (perfect model)
  - Forecast model: SPEEDY with **persistence forecast of Sensible/Latent heat fluxes (SHF/LHF)**
  - Observations: conventional observations of (U, V, T, q, Ps) and AIRS retrievals of (T, q)
  - Analysis: U, V, T, q, Ps + SHF & LHF
- Fully multivariate data assimilation
- Adaptive multiplicative inflation + additive inflation
- Initial conditions: random (no *a-priori* information)

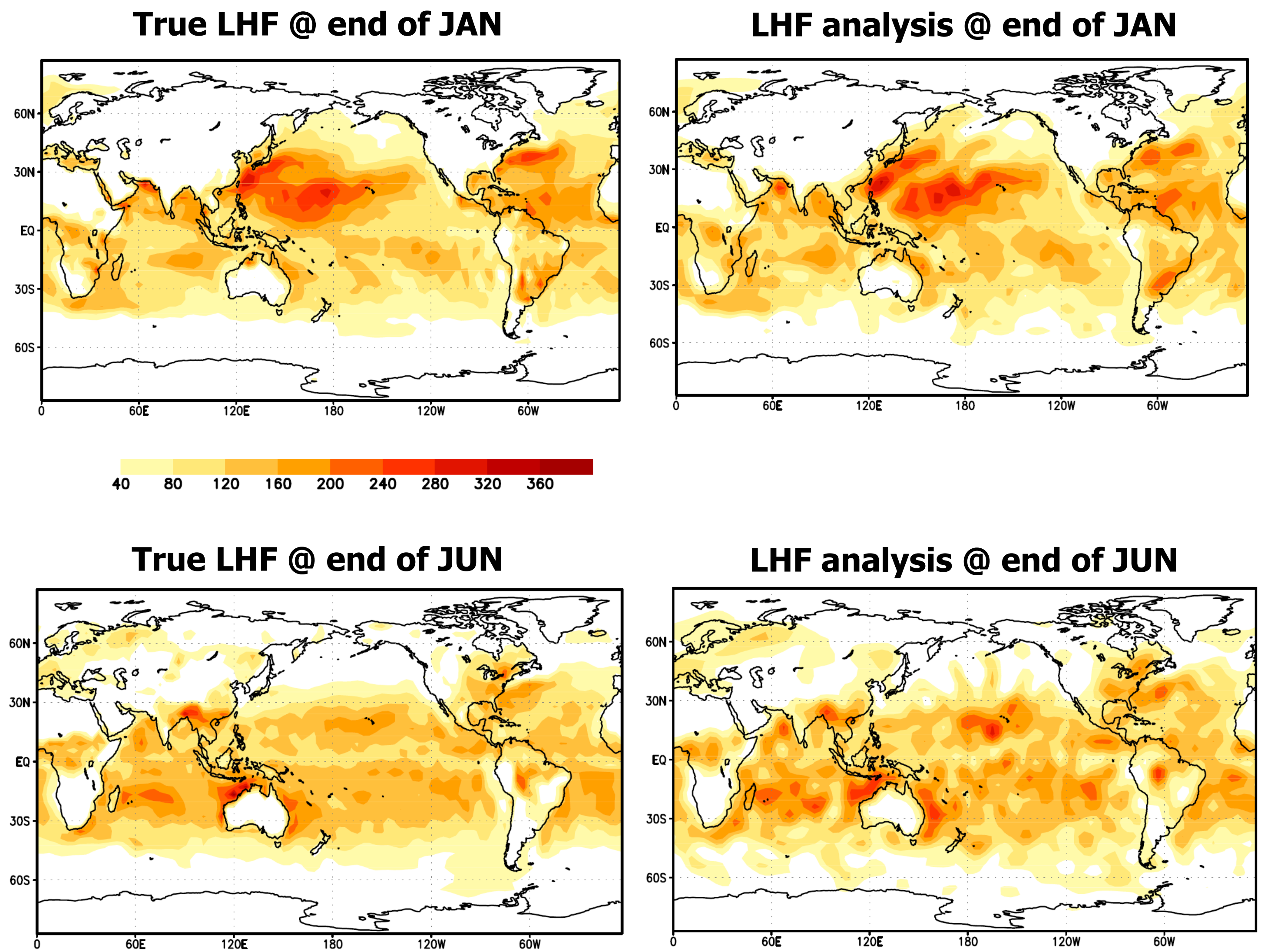
Results: SHF

(first, assimilating winds, but assuming perfect wind stress parameterization)

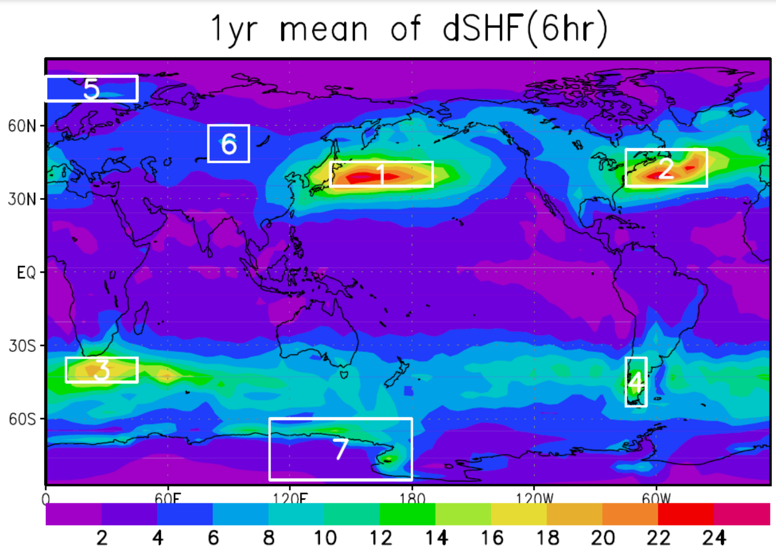


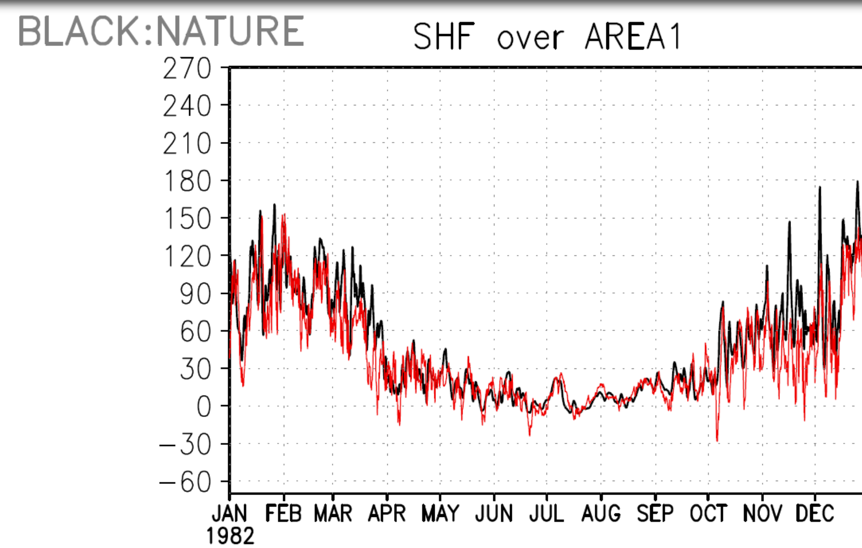
Results: LHF

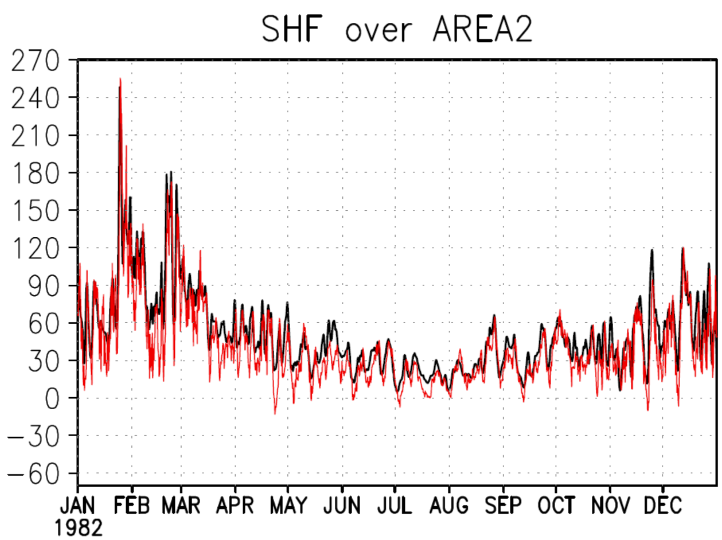
(assimilating wind, but assuming perfect wind stress parameterization)

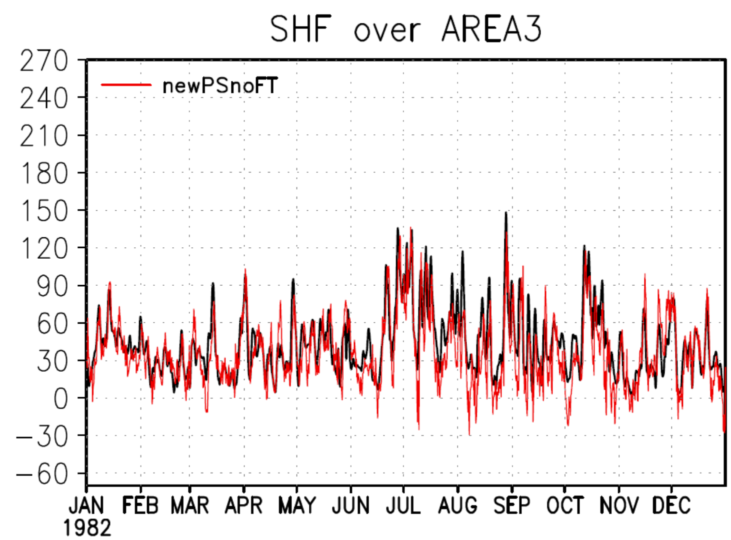


Time series of SHF (perfect wind stress parameterization)

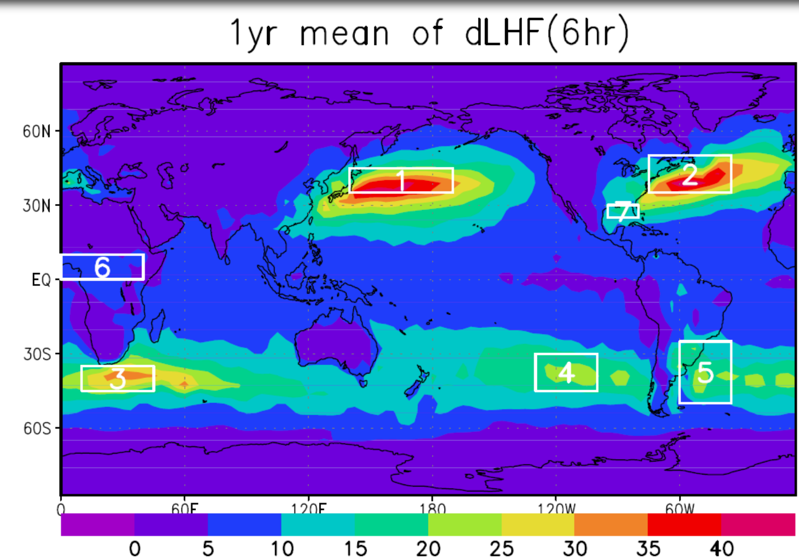


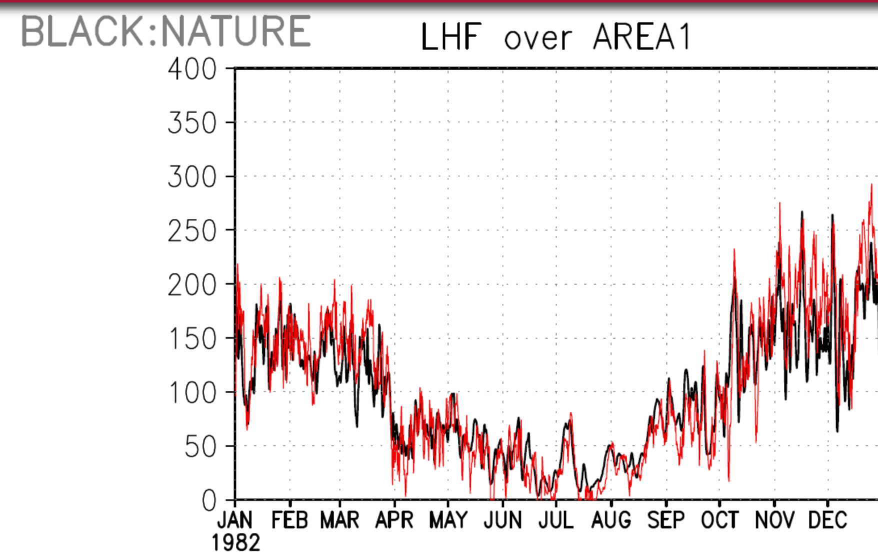


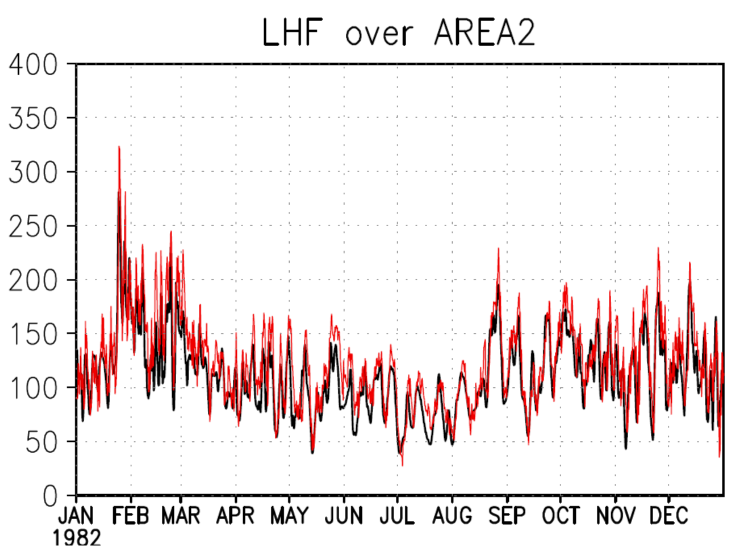


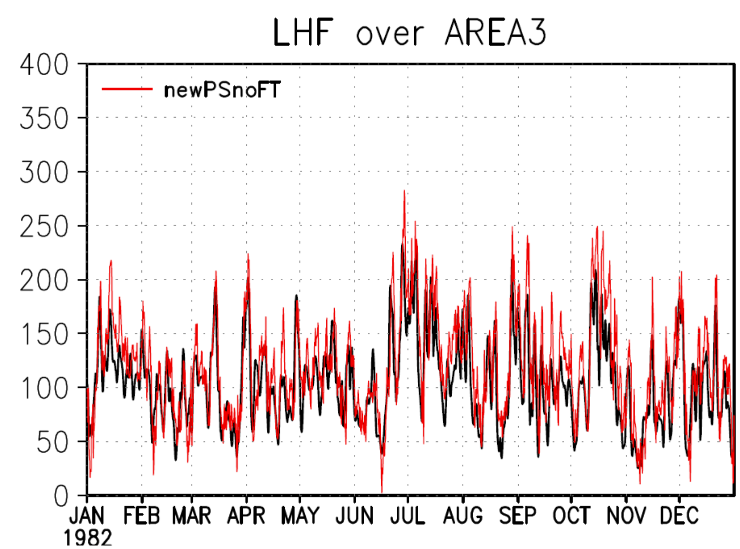


Time series of LHF (perfect wind stress parameterization)









Can we also estimate wind stress?

- OSSEs
  - Nature: SPEEDY
  - Forecast model: SPEEDY with **persistence forecast of Sensible/Latent heat fluxes (SHF/ LHF) <u>and wind stress</u> (USTR, VSTR).**
  - Observations: conventional observations of (U, V, T, q, Ps), AIRS retrievals of (T, q), and ASCAT ocean surface wind observations
    - **Observation error of ASCAT: 3.5m/s (not as good as AIRS data)**
    - ASCAT covers the global ocean every 12 hours, but with little overlap with AIRS.
    - Analysis: U, V, T, q, Ps + SHF, LHF, USTR, VSTR
- Fully multivariate data assimilation

Results diverge unless we increase the ensemble size

RMSE:

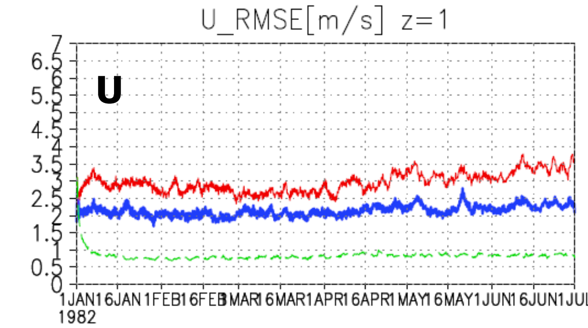
Blue: 80 ensembles

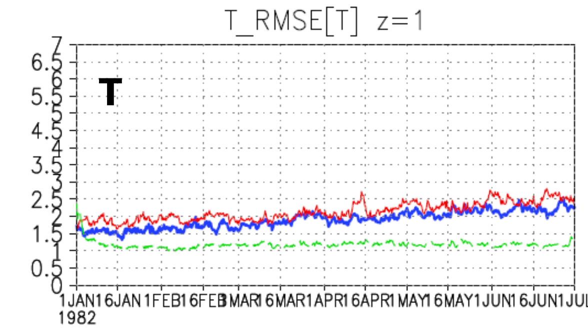
Red: 40 ensembles

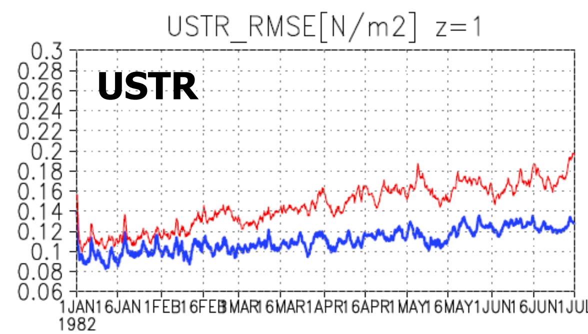
Green: perfect WSTR with 40 ensembles

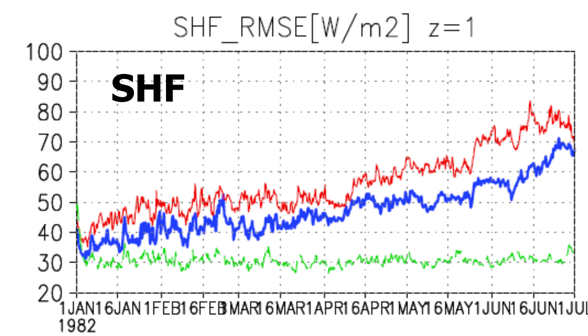
Doubling ensemble size reduces error but not enough to produce stable estimation of parameters throughout the analysis period.

Stress is reasonable but SHF are underestimated and LHF are overestimated over oceans. Over land, with more observations, they seem good.









Summary

- We have shown the feasibility of simultaneous analysis of meteorological and carbon variables within LETKF framework through OSSEs and short windows.
- The system LETKF-C has been tested in a intermediate-complexity model SPEEDY-C with good results.
  - Multivariate assimilation with localization of variables.
  - Physical vertical localization.
  - Additive and adaptive multiplicative inflation.
- Implementation of the LETKF-C to NCAR CAM 3.5 model: Analysis shows good performance in OSSEs with real observation coverage
- Application to estimation of surface fluxes of heat, moisture and momentum.
  - Preliminary results are encouraging, although slowly divergent. Need more ensemble members.