A Regional Carbon Data Assimilation System and its Preliminary Evaluation in East Asia

Zhen Peng*, Meigen Zhang*, Xingxia Kou2,3, Xiangjun Tian4, and Xiaoguang Ma4

1 School of Atmospheric Sciences, Nanjing University, Nanjing 210093, China
2 State Key Laboratory of Atmospheric Boundary Layer Physics and Atmospheric Chemistry, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100029, China
3 University of Chinese Academy of Sciences, Beijing 100049, China
4 Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100029, China

*Corresponding author: pengzhen@nju.edu.cn; *Corresponding author: mgzhang@mail.iap.ac.cn;
ABSTRACT

In order to optimize surface CO$_2$ fluxes at grid scales, a regional surface CO$_2$ flux inversion system (Carbon Flux Inversion system and Community Multi-scale Air Quality, CFI-CMAQ) has been developed by applying the ensemble Kalman filter (EnKF) to constrain the CO$_2$ concentrations and applying the ensemble Kalman smoother (EnKS) to optimize the surface CO$_2$ fluxes. The smoothing operator is associated with the atmospheric transport model to constitute a persistence dynamical model to forecast the surface CO$_2$ flux scaling factors. In this implementation, the ‘signal-to-noise’ problem can be avoided; plus, any useful observed information achieved by the current assimilation cycle can be transferred into the next assimilation cycle. Thus, the surface CO$_2$ fluxes can be optimized as a whole at the grid scale in CFI-CMAQ. The performance of CFI-CMAQ was quantitatively evaluated through a set of Observing System Simulation Experiments (OSSEs) by assimilating CO$_2$ retrievals from GOSAT (Greenhouse Gases Observing Satellite). The results showed that the CO$_2$ concentration assimilation using EnKF could constrain the CO$_2$ concentration effectively, illustrating that the simultaneous assimilation of CO$_2$ concentrations can provide convincing CO$_2$ initial analysis fields for CO$_2$ flux inversion. In addition, the CO$_2$ flux optimization using EnKS demonstrated that CFI-CMAQ could in general reproduce true fluxes at grid scales with acceptable bias. Two further sets of numerical experiments were conducted to investigate the sensitivities of the inflation factor of scaling factors and the smoother window. The results showed that the ability of CFI-CMAQ to optimize CO$_2$ fluxes greatly relied on
the choice of the inflation factor. However, the smoother window had a slight
influence on the optimized results. CFI-CMAQ performed very well even with a short
lag-window (e.g. 3 days).

1 Introduction

Considerable progress has been made in recent years to reduce the uncertainties of
surface CO$_2$ flux estimates through the use of an advanced data assimilation technique
(e.g., Chevallier et al., 2005, 2007a and 2007b; Baker et al., 2006; Engelen et al., 2009;
Liu et al., 2012). Feng et al. (2009) showed that the uncertainties of surface CO$_2$ flux
estimates can be reduced significantly by assimilating OCO X$_{CO2}$ measurements.
Peters et al. (2005, 2007, 2009) developed a surface CO$_2$ flux inversion system,
CarbonTracker, by incorporating the ensemble square-root filter (EnSRF) into the
atmospheric transport TM5 model. And the inversion results obtained by assimilating
in situ surface CO$_2$ observations are in excellent agreement with a wide collection of
carbon inventories that form the basis of the first North American State of the Carbon
Cycle Report (SOCCR) (Peters et al., 2007). CarbonTracker is also well used to
constrain the surface CO$_2$ fluxes over Europe and Asia (e.g., Zhang et al., 2014a,
2014b). Kang et al. (2012) presented a simultaneous data assimilation of surface CO$_2$
fluxes and atmospheric CO$_2$ concentrations along with meteorological variables using
the Local Ensemble Transform Kalman Filter (LETKF). They indicated that an
accurate estimation of the evolving surface fluxes can be gained even without any a
priori information. Recently, Tian et al. (2013) developed a new surface CO$_2$ flux data
assimilation system, Tan-Tracker, by incorporating a joint PODEn4DVar assimilation
framework into the GEOS-Chem model on the basis of Peters et al. (2005, 2007) and
Kang et al. (2011, 2012). They discussed in detail that the assimilation of CO₂ surface
fluxes could be improved through the use of simultaneous assimilation of CO₂
concentrations and CO₂ surface fluxes. Despite the rigor of data assimilation theory,
current CO₂ flux-inversion methods still face many challenging scientific problems,
such as: (1) the well-known ‘signal-to-noise’ problem (NRC, 2010); (2) large
inaccuracies in chemical transport models (e.g., Prather et al., 2008); (3) vast
computational expenses (e.g., Feng et al., 2009); and (4) the sparseness of observation
data (e.g., Gurney et al., 2002).

The ‘signal-to-noise’ problem is one of the most challenging issue for an
ensemble-based CO₂ flux inversion system due to the fact that surface CO₂ fluxes are
the model forcing (or boundary condition), rather than model states (like CO₂
concentrations), of the chemistry transport model (CTM). In the absence of a suitable
dynamical model to describe the evolution of the surface CO₂ fluxes, most CO₂
flux-inversion studies have traditionally ignored the uncertainty of anthropogenic and
other CO₂ emissions and focused on the optimization of natural (i.e., biospheric and
oceanic) CO₂ emissions at the ecological scale (e.g., Deng et al., 2007; Feng et al.,
2009; Peters et al., 2005, 2007; Jiang et al., 2013; Peylin et al., 2013).

This compromise is acceptable to some extent. Indeed, the total amount of
anthropogenic CO₂ emissions can be estimated by relatively well-documented global
fuel-consumption data with a small degree of uncertainty (Boden et al., 2011). And
the uncertainties involved in the total amount of anthropogenic CO\textsubscript{2} emissions are much smaller than those related to natural emissions. However, their spatial distribution, strength and temporal development still remain elusive, because of their inherent non-uniformities (Andres et al., 2012; Gurney et al., 2009). Marland (2008) pointed out that even a tiny amount of uncertainty, i.e., 0.9%, in one of the leading emitter countries like the U.S. is equivalent to the total emissions of the smaller emitter countries in the world. Furthermore, the usual values of anthropogenic CO\textsubscript{2} emissions in chemical transport models have thus far been simply interpolated from very coarse monthly-mean fuel consumption data. Therefore, great uncertainty in the spatiotemporal distributions of anthropogenic emissions likely exists, which could reduce the accuracy of CO\textsubscript{2} concentration simulations and subsequently increase the inaccuracy of natural CO\textsubscript{2} flux inversion results. In addition, current research approaches tend only to assimilate natural CO\textsubscript{2} emissions at the ecological scale, which is far from sufficient. Therefore, surface CO\textsubscript{2} fluxes should be constrained as a whole at finer scale.

In CarbonTracker (Peters et al., 2007), a smoothing operator is innovatively applied as the persistence forecast model. In that application, the surface CO\textsubscript{2} fluxes can be treated as the model states and the observed information ingested by the current assimilation cycle can be used in the next assimilation cycle effectively. However, the ‘signal-to-noise’ problem is not yet resolved, and thus CarbonTracker also has to assimilate natural CO\textsubscript{2} emissions at the ecological scale only. In Tan-Tracker (Tian et al., 2013), a 4-D moving sampling strategy (Wang et al., 2010)
is used to generate the flux ensemble members, and so the surface CO$_2$ fluxes can be optimized as a whole at the grid scale. In the present reported work, the persistence dynamical model taken by Peters et al. (2005) was further developed for the purpose of resolving the ‘signal-to-noise’ problem to optimize the surface CO$_2$ fluxes as a whole at the grid scale. This process is described in detail in section 2 of this paper.

The surface CO$_2$ flux inversion system presented in this paper was developed by simultaneous optimizing the surface CO$_2$ fluxes and constraining the CO$_2$ concentrations. As we know, assimilating CO$_2$ observations from multiple sources can improve the accuracy of simulation results (e.g., Miyazaki, 2009; Liu et al., 2009, 2011, 2012; Tangborn et al, 2013; Huang et al., 2014). In addition, previous studies showed that the simultaneous assimilation of CO$_2$ concentrations and surface CO$_2$ fluxes can largely eliminate the uncertainty in initial CO$_2$ concentrations on the CO$_2$ evolution (Kang et al., 2012; Tian et al., 2013). Therefore, we also use the simultaneous assimilation framework and the ensemble Kalman filter (EnKF) was used to constrain CO$_2$ concentrations and the ensemble Kalman smoother (EnKS) was used to optimize surface CO$_2$ fluxes. Since the regional chemical transport models, compared to global models, have some advantages to reproduce the effects of meso–micro–scale transport on atmospheric CO$_2$ distributions (Ahmadov et al., 2009, Pillai et al., 2010; Kretschmer et al., 2011), we choose a regional model, Regional Atmospheric Modeling System and Community Multi-scale Air Quality (RAMS-CMAQ) (Zhang et al. 2002, 2003, 2007; Kou et al. 2013; Liu et al., 2013; Huang et al. 2014), to develop this inversion system. For simplicity, this system is
referred to as CFI-CMAQ (Carbon Flux Inversion system and Community Multi-scale Air Quality).

Since this is the first time of introducing CFI-CMAQ, we focus mainly on introducing the methodology in this paper. Nevertheless, in addition, Observing System Simulation Experiments (OSSEs) were designed to assess the system’s ability to optimize surface CO$_2$ fluxes. The retrieval information of GOSAT X$_{CO2}$ are used to generate artificial observations because of the sparseness and heterogeneity of ground-based measurements.

The remainder of the paper is organized as follows. Section 2 describes the details of the regional surface CO$_2$ flux inversion system, CFI-CMAQ, including the developed persistence dynamical model, a simple review of the EnKS and EnKF assimilation approaches, and the process involved. The experimental designs are then introduced and the assimilation results shown in Sect. 3. Finally, a summary and conclusions are provided in Sect. 4.

2 Framework of the regional surface CO$_2$ flux inversion system

Supposed we have the prescribed net CO$_2$ surface flux, $F^*(x, y, z, t)$, which can be released from a climate model or be generated by others methods, our ultimate goal is to optimize $F^*(x, y, z, t)$ by assimilating CO$_2$ observations from various platforms. As an ensemble-based assimilation system, CFI-CMAQ was also developed by applying a set of linear multiplication factors, similar to the approach by Peters et al. (2007) and Tian et al. (2013). The $i$th ensemble member of the surface
fluxes, \( F_i(x, y, z, t) \), from an \( N \)-member ensemble can be described by

\[
F_i(x, y, z, t) = \lambda_i(x, y, z, t) F^*(x, y, z, t), \quad (i = 1, \ldots, N),
\]

where \( \lambda_i(x, y, z, t) \) represents the \( i \)th ensemble member of the linear scaling factors (Peters et al., 2007; Tian et al., 2013) for each time and each grid to be optimized in the assimilation. The notations are standard: the subscript \( i \) refers to the \( i \)th ensemble member. In the following, \( \lambda_i(x, y, z, t) \) is referred to as \( \lambda_{i,j} \), \( F^*(x, y, z, t) \) is referred to as \( F_i^* \), and \( F_i(x, y, z, t) \) is referred to as \( F_{i,j} \) for simplicity.

At each optimization cycle, CFI-CMAQ includes the following four parts in turn (see Fig. 1): (1) forecasting of the linear scaling factors at time \( t \), \( \lambda_{i,|t-1|} \); (2) optimization of the scaling factors in the smoother window, \( \lambda_{i,t-M+1|t-1|}, \ldots, \lambda_{i,t-M+1|t-1|}, \lambda_{i,t-j|t-1|}, \lambda_{i,t-j|t-1|} \), by EnKS, Where \( \lambda_{i,j} \) \( (j = \text{t} - \text{M}, \ldots, \text{t} - 1) \) refer to analyzed quantities from the previous assimilation cycle at time \( j \) (see Fig. 1). \( \text{t} - 1 \) means that these factors have been updated by using observations before time \( \text{t} - 1 \), and the superscript \( \text{a} \) refers to the analyzed; (3) updating of the fluxes in the smoother window, \( F_{i,t-M+1|t-1|}, \ldots, F_{i,t-M+1|t-1|}, F_{i,t-j|t-1|}, F_{i,t-j|t-1|} \); and (4) assimilation of the forecast CO2 concentration fields at time \( t \), \( C_i^f(x, y, z, t) \) (referred to as \( C_{i,j}^f \), and the superscript \( f \) refers to the forecast or the background), by EnKF. A flowchart illustrating CFI-CMAQ is presented in Fig. 2. The assimilation procedure is addressed in detail below. In addition, the observation operator is introduced, particularly for use of the GOSAT XCO2 data in Sect. 2.4. Furthermore, covariance inflation and localization techniques applied in CFI-CMAQ are introduced briefly in Sect. 2.5.
2.1 Forecasting the linear scaling factors at time $t$, $\lambda_{v,t-1}$

In the previous assimilation cycle $t-1-M\sim t-1$ (see Fig. 1), the optimized scaling factors in the smoother window are $(\lambda_{v,j-M-1}^a, \lambda_{v,j-M+1}^a, \cdots, \lambda_{v,j-M+2j-1}^a)\cdots, \lambda_{v,j-1}^a, \cdots, \lambda_{v,j-M+2j-1}^a)$ and the assimilated $\mathrm{CO}_2$ concentration fields at time $t-1$ are $C_{v,i}^a(x,y,z,t-1)$ (referred to as $C_{v,i}^a$). In the current assimilation cycle $t-M\sim t$, the scaling factors in the current smoother window are $(\lambda_{v,j-M-1}^a, \lambda_{v,j-M+1}^a, \cdots, \lambda_{v,j-M+2j-1}^a)\cdots, \lambda_{v,j-1}^a, \cdots, \lambda_{v,j-M+2j-1}^a)$ and the forecast $\mathrm{CO}_2$ concentration fields at time $t$ are $C_{v,i}^f$.

In order to pass the useful observed information onto the next assimilation cycle effectively, following Peters et al. (2007) the smoothing operator is applied as part of the persistence dynamical model to calculate the linear scaling factors $\lambda_{v,t-1}^a$,

$$\lambda_{v,t-1}^s = \frac{\left( \sum_{j=-M}^{t-1} \lambda_{v,j-1}^a + \lambda_{v,j-1}^p \right)}{M+1}, \quad (i = 1, \cdots, N), \quad (2)$$

where $\lambda_{v,j-1}^p$ refers to the prior values of the linear scaling factors at time $t$. The superscript $p$ refers to the prior. This operation represents a smoothing over all the time steps in the smoother window (see Fig. 1), thus dampening variations in the forecast of $\lambda_{v,t-1}^a$ in time.

In order to generate $\lambda_{v,t-1}^p$, the atmospheric transport model (CMAQ) is applied and a set of ensemble forecast experiments are carried out. It integrates from time $t-1$ to $t$ to produce the $\mathrm{CO}_2$ concentration fields $\hat{C}_i^f(x,y,z,t)$ (referred to as $\hat{C}_i^f$ hereafter to distinguish from $C_{v,i}^f$) forced by the prescribed net $\mathrm{CO}_2$ surface flux $F_{i}^e$ with $C_{v,i}^s$ as initial conditions. Then, the ratio $\kappa_{i,t} = \hat{C}_{i,t}^f / \overline{C}_{i,t}^f$ is calculated, where $\overline{C}_{i,t} = \frac{1}{N} \sum_{j=1}^{N} \hat{C}_{i,t}^f$. Supposed that $\lambda_{v,t-1}^p = \kappa_{i,t}$ due to the fact that the surface
CO\textsubscript{2} fluxes correlate with its concentrations, the values for \( \lambda_{i,j|t-1}^p \) are obtained and then \( \lambda_{i,j|t-1}^a \) can finally be calculated (see the part with red arrows in the flowchart in Fig. 2).

The way the prior scaling factor \( \lambda_{i,j|t-1}^p \) is updated by associating with the atmospheric transport model is the main improvement over the one used in CarbonTracker (Peters et al., 2007). In CarbonTracker, all \( \lambda_{i,j|t-1}^p \) are set to 1 (Peters et al., 2007). The distribution of the ensemble members of the linear scaling factors at time \( t, \lambda_{i,j|t-1}^p \), are finally dependent on the distribution of the previous scaling factors because Eq. (2) is a linear smoothing operator. In this study, the values of \( \lambda_{i,j|t-1}^p \) are updated by associating with the atmospheric transport model. It is important to note that \( \lambda_{i,j|t-1}^p \) in this study are random fields with mean 1. However, the distribution of \( \lambda_{i,j|t-1}^a \) are dependent on the distribution of all the scaling factors in the smoother window. An OSSE was designed to illustrate the difference between our method and the one where \( \lambda_{i,j|t-1}^p \) are set to 1 in Sect. 3.

It is also important to note that, similar to Peters et al. (2007), this dynamical model equation still does not include an error term in the dynamical model, and the model error cannot yet be estimated. However, the covariance inflation is applied to compensate for model errors before optimization, which is addressed in section 2.5.

2.2 Optimizing the scaling factors in the smoother window by EnKS

Substituting \( \lambda_{i,j|t-1}^a \) into Eq. (1), the \( i \)th member of the surface fluxes at time \( t, F_{i,j|t-1}^a \), can be generated. Then forced by \( F_{i,j|t-1}^a \), CMAQ was run from time \( t-1 \) to \( t \) to produce the background concentration field \( C_{i,t}^f \) with \( C_{1,t-1}^a \) as initial...
conditions.

In the current assimilation cycle \( t-M\rightarrow t \) (see Fig. 1), the scaling factors to be optimized in the smoother window are \((\lambda^a_{j-M\rightarrow t-1}, \lambda^a_{j-M\rightarrow t-1}, \ldots, \lambda^a_{j-M\rightarrow t-1}, \lambda^a_{j-1})\), as stated in the first paragraph of Sect. 2.1. Using the EnKS analysis technique, these scaling factors are updated in turn via

\[
\lambda^a_{i,j,t} = \lambda^a_{i,j,t-1} + K^e_{i,j,t-1}(y_{i}^{\text{obs}} - y_{i}^{f}) + \nu_{i}, \quad (i = 1, \ldots, N, j = t-M, \ldots, t),
\]

(3)

\[
K^e_{i,j,t-1} = S^e_{i,j,t-1}H^T(HS^e_{i,j,t-1}H^T + R)^{-1},
\]

(4)

\[
S^e_{i,j,t-1} = \frac{1}{N-1} \sum_{i=1}^{N} [\lambda^a_{i,j,t-1} - \bar{\lambda}^a_{i,j,t-1}][\lambda^a_{i,j,t-1} - \bar{\lambda}^a_{i,j,t-1}]^T,
\]

(5)

\[
S^e_{i,t-1} = \frac{1}{N-1} \sum_{i=1}^{N} [\lambda^a_{i,j,t-1} - \bar{\lambda}^a_{i,j,t-1}][\lambda^a_{i,t-1} - \bar{\lambda}^a_{i,t-1}]^T,
\]

(6)

\[
y_{i,t}^f = H(\varphi_{i,t-1}, \bar{\lambda}^a_{i,t-1})) = H(C^f_{i,t}),
\]

(7)

where \( K^e_{i,j,t-1} \) is the Kalman gain matrix of EnKS, \( y_{i}^{\text{obs}} \) is the observation vector measured at time \( t \) and \( y_{i}^{f} \) is the simulated values, \( \nu_{i} \) is a random normal distribution perturbation field with zero mean, \( S^e_{i,j,t-1} \) is the background error cross-covariance between the state vector \( \lambda^a_{i,j,t-1} \) and \( \lambda^a_{i,j,t-1} \), \( S^e_{i,t-1} \) is the background error covariance of the state vector \( \lambda^a_{i,j,t-1} \), \( H(\cdot) \) is the observation operator that maps the state variable from model space into observation space, \( R \) standard deviation representing the measurement errors, and \( \varphi(\cdot) \) is the atmospheric transport model.

In actual implementations, it is unnecessary to calculate \( S^e_{i,j,t-1} \) and \( S^e_{i,t-1} \) separately. \( S^e_{i,j,t-1}H^T \) and \( HS^e_{i,j,t-1}H^T \) can be calculated as a whole by

\[
S^e_{i,j,t-1}H^T = \frac{1}{N-1} \sum_{i=1}^{N} [\lambda^a_{i,j,t-1} - \bar{\lambda}^a_{i,j,t-1}][y_{i,t}^f - \bar{y}_{i,t}^f]^T.
\]

(8)
\[
HS_{i,j,t-1}^{f} H^T = \frac{1}{N-1} \sum_{i=1}^{N} (y_{i,j}^f - \overline{y}_i^f) (y_{i,j}^f - \overline{y}_i^f)^T ,
\]  

(9)

\[
\overline{y}_i^f = H(C_i^f) = H\left(\frac{1}{N} \sum_{i=1}^{N} C_i^f\right) .
\]  

(10)

After EnKS, \((\lambda_{i-M+1;}, \lambda_{i-M+1;}, \ldots, \lambda_{i-M+1;}, \lambda_{i-M+1;})\) are gained. Then the corresponding fluxes in the smoother window \((F_{i-M+1;}, F_{i-M+1;}, \ldots, F_{i-M+1;}, F_{i-M+1;})\) can be gained (see the part with green arrows in the flowchart in Fig. 2) by substituting \((\lambda_{i-M+1;}, \lambda_{i-M+1;}, \ldots, \lambda_{i-M+1;}, \lambda_{i-M+1;})\) into Eq. (1).

Then the ensemble mean values of the assimilated fluxes in the smoother window can be calculated via,

\[
\overline{F}_{i,j,t}^a = \frac{1}{N} \sum_{i=1}^{N} F_{i,j,t}^a, \quad (j = t - M, \ldots, t) ,
\]  

(11)

Finally, those ensemble mean assimilated fluxes which are before the next smoother window and will not be updated by the succeeding observations are regarded as the final optimized fluxes. We referred them as \(\overline{F}_{i}^a\) for simplicity.

2.3 Assimilating the CO\(_2\) concentration fields at time \(t\) by EnKF

The analysis of CO\(_2\) concentrations fields at time \(t\) in the EnKF scheme is updated via

\[
C_{i,t}^a = C_{i,t}^f + K(y_{i,t}^{obs} - y_i^f + \delta_{i,t}) ,
\]  

(12)

\[
K = P^f H^T (HP^f H^T + R)^{-1} ,
\]  

(13)

where \(K\) is the Kalman gain matrix of EnKF, \(P^f\) is the background error covariance among the background CO\(_2\) concentration fields \(C_{i,t}^f\).

In actually application, \(P^f H^T\) and \(HP^f H^T\) can be calculated as a whole by
\[ P^f H^T = \frac{1}{N-1} \sum_{i=1}^{N} [C_{i}^f - \bar{C}_i^f] [y_i - \bar{y}_i]^T, \]  
(14)

\[ HP^f H^T = \frac{1}{N-1} \sum_{i=1}^{N} [y_i^f - \bar{y}_i^f]^T [y_i^f - \bar{y}_i^f]^T, \]  
(15)

\[ \bar{C}_i^f = \frac{1}{N} \sum_{i=1}^{N} C_{i,j}^f \]  
(16)

Finally, the ensemble mean values of the assimilated CO\textsubscript{2} concentrations fields can be gained via,

\[ \bar{C}_i^a = \frac{1}{N} \sum_{i=1}^{N} C_{i,j}^a \]  
(17)

where \( \bar{C}_i^a \) is regarded as the final analyzing concentration field.

2.4 The observation operator

As mentioned above, the observation operator \( H(\cdot) \) transforms the state variable from model space into observation space. Usually, it is the spatial bilinear interpolator for traditional ground-based observations. Since the GOSAT X\textsubscript{CO2} retrieval is a weighted CO\textsubscript{2} column average, the simulated X\textsubscript{CO2} should be calculated with the same weighted column average method (Connor et al., 2008; Crisp et al., 2010, 2012; O’Dell et al, 2012). So, the observation operator to assimilate the GOSAT X\textsubscript{CO2} retrieval is

\[ y_{i,j}^f = H(\varphi_{x_{j-1}x}(A_{x_{j-1}}^a)) = H(C_{i}^f) = y_{i}^{\text{priori}} + h^T a_{\text{CO2}}(S(C_{i}^f) - f_{\text{priori}}), \]  
(18)

where \( y_{i,j}^f \) is the simulated X\textsubscript{CO2}; \( y_{i}^{\text{priori}} \) is the a priori CO\textsubscript{2} column average used in the GOSAT X\textsubscript{CO2} retrieval process; \( S(\cdot) \) is the spatial bilinear interpolation operator that interpolates the simulated fields to the GOSAT X\textsubscript{CO2} locations to obtain the simulated CO\textsubscript{2} vertical profiles there; \( f_{\text{priori}} \) is the a priori CO\textsubscript{2} vertical profile used.
in the retrieval process; \( h \) is the pressure weighting function, which indicates the contribution of the retrieved value from each layer of the atmosphere; and \( a_{co2} \) is the normalized averaging kernel.

### 2.5 Covariance inflation and localization

In order to keep the ensemble spread of the CO\(_2\) concentrations at a certain level and compensate for transport model error to prevent filter divergence, covariance inflation is applied before updating the CO\(_2\) concentrations. So,

\[
(C_{i,t}^f)_{\text{new}} = \alpha (C_{i,t}^f - \overline{C_{i,t}^f}) + \overline{C_{i,t}^f},
\]

(19)

where \( \alpha \) is the inflation factor of CO\(_2\) concentrations and \((C_{i,t}^f)_{\text{new}}\) is the final field used for data assimilation.

Similarly, covariance inflation is also used to keep the ensemble spread of the prior scaling factors at a certain level and compensate for dynamical model error. So,

\[
(\lambda^p_{r;\text{new}}) = \beta (\lambda^p_{r;\text{old}} - \overline{\lambda^p_{r;\text{old}}}) + \overline{\lambda^p_{r;\text{old}}},
\]

(20)

where \( \beta \) is the inflation factor of scaling factors and \((\lambda^p_{r;\text{new}})\) is the final scaling factors used for data assimilation.

In addition, the Schur product is utilized to filter the remote correlation resulting from the spurious long-range correlations (Houtekamer and Mitchell 2001). So, the Kalman gain matrix \( K^e_{r;\text{new}} \) and \( K \) are updated via,

\[
K^e_{r;\text{new}} = [(\rho \circ S^e_{r;\text{new}})H^T(H/(\rho \circ P^e_{r;\text{new}})H^T + R)^{-1}],
\]

(21)

\[
K = [(\rho \circ P^f)H^T][(H/(\rho \circ P^f)H^T + R)^{-1}],
\]

(22)

where the filtering matrix \( \rho \) is calculated using the formula
\[
C_0(r, c) = \begin{cases}
-\frac{1}{4} \left( \frac{|r|}{c} \right)^5 + \frac{1}{2} \left( \frac{|r|}{c} \right)^4 + \frac{5}{8} \left( \frac{|r|}{c} \right)^3 - \frac{5}{3} \left( \frac{|r|}{c} \right)^2 + 1, & 0 \leq |r| \leq c \\
\frac{1}{12} \left( \frac{|r|}{c} \right)^5 - \frac{1}{2} \left( \frac{|r|}{c} \right)^4 + \frac{5}{8} \left( \frac{|r|}{c} \right)^3 + \frac{5}{3} \left( \frac{|r|}{c} \right)^2 - 5 \left( \frac{|r|}{c} \right) + 4 - \frac{2}{3} \left( \frac{c}{r} \right), & c \leq |r| \leq 2c \\
0, & c \leq |r| 
\end{cases}
\]

where \( c \) is the element of the localization Schur radius. The matrix \( \mathbf{\rho} \) can filter the small background error correlations associated with remote observations through the Schur product (Tian et al., 2011). And the Schur product tends to reduce the effect of those observations smoothly at intermediate distances due to the smooth and monotonically decreasing of the filtering matrix.

3 OSSEs for evaluation of CFI-CMAQ

A set of OSSEs were designed to quantitatively assess the performance of CFI-CMAQ. The setup of the experiments and the results are described in this section.

3.1 Experimental setup

The chemical transport model utilized was RAMS-CMAQ (Zhang et al., 2002), in which \( \text{CO}_2 \) was treated as an inert tracer. The model domain was \( 6654 \times 5440 \text{ km}^2 \) on a rotated polar stereographic map projection centered at \( (35.0^\circ \text{N}, 116.0^\circ \text{E}) \), with a horizontal grid resolution of \( 64 \times 64 \text{ km}^2 \) and 15 vertical layers in the \( \sigma_z \)-coordinate system, unequally spaced from the surface to approximately 23 km. The initial fields and boundary conditions of the \( \text{CO}_2 \) concentrations were interpolated from the simulated \( \text{CO}_2 \) fields of CarbonTracker 2011 (Peters, 2007). The prior surface \( \text{CO}_2 \)
fluxes included biosphere–atmosphere CO₂ fluxes, ocean–atmosphere CO₂ fluxes, anthropogenic emissions, and biomass-burning emissions (Kou et al., 2013),

\[ F^p(x, y, z, t) = F^p_{\text{bio}}(x, y, z, t) + F^p_{\text{oce}}(x, y, z, t) + F^p_{\text{ff}}(x, y, z, t) + F^p_{\text{inc}}(x, y, z, t), \]  

(24)

where \( F^p(x, y, z, t) \) (referred to as \( F^p_i \)) was the prior surface CO₂ flux; \( F^p_{\text{bio}}(x, y, z, t) \) and \( F^p_{\text{oce}}(x, y, z, t) \) were the biosphere–atmosphere and ocean–atmosphere CO₂ fluxes, respectively, which were obtained from the optimized results of CarbonTracker 2011 (Peters, 2007); \( F^p_{\text{ff}}(x, y, z, t) \) was fossil fuel emissions, adopted from the Regional Emission inventory in ASia (REAS, 2005 Asia monthly mean emission inventory) with a spatial resolution of \(0.5^\circ \times 0.5^\circ\) (Ohara et al., 2007); \( F^p_{\text{inc}}(x, y, z, t) \) was biomass–burning emissions, provided by the monthly mean inventory at a spatial resolution of \(0.5^\circ \times 0.5^\circ\) from the Global Fire Emissions Database, Version 3 (GFED v3) (Van der Werf et al., 2010). Among all these fluxes, \( F^p_{\text{bio}}(x, y, z, t) \), \( F^p_{\text{oce}}(x, y, z, t) \) and \( F^p_{\text{ff}}(x, y, z, t) \) had nonzero values at model level 1, while they all were zeros at other 14 levels. However, \( F^p_{\text{inc}}(x, y, z, t) \) had nonzero values at model level 1~5 and they were all zeros at other 10 levels. So, all fluxes in this paper were the function of \((x, y, z, t)\) for convenience.

Firstly, the prior flux \( F^p_i \) was assumed as the true surface CO₂ flux in all of the following OSSEs. Forced by \( F^p_i \), the RAMS-CMAQ model was run to produce the artificial true CO₂ concentration results \( C^p(x, y, z, t) \) (refer to as \( C^p_i \) in the following). Then, the artificial GOSAT observations \( y^\text{obs}_t \) (or \( X^p_{CO_2} \)) were generated by substituting \( C^p_i \) into the observation operator in Eq. (18). The retrieval information of GOSAT \( X_{CO_2}(y^\text{prior}_t, f^\text{prior}_t, h \text{ and } a_{CO_2}) \) needed in Eq. (18) were
gained from the v2.9 Atmospheric CO$_2$ Observations from Space (ACOS) Level 2 standard data products, which only utilized the SWIR observations. Only data classified into the “Good” category were utilized in this study. During the retrieval process, most of the soundings (such as data with a solar zenith angle greater than $85^\circ$, or data not in clear sky conditions, or data collected over ocean but not in glint, etc.) were not processed, so typically data products for the “Good” category contained only 10-100 soundings per satellite orbit (Osterman et al., 2011), and there were only 0~60 samples per orbit in the study model domain generally. Fig. 3 (a) also showed the total number of “good” GOSAT X$_{CO2}$ observations for each model grid in February in 2010. There were relatively more observations over most continental regions of the study domain except some regions in North-East and South China. The total numbers ranged from 1 to 8. However, there were almost no data over oceans of the study domain.

Secondly, the prescribed surface CO$_2$ fluxes series $F^*_t$ were created by

$$F^*_t = (1.8 + \delta(x, y, z, t))F^p_t,$$

(25)

where $\delta$ was a random number. They were standard normal distribution time series at each grid in the integration period of our numerical experiment. Driven by $F^*_t$, the RAMS-CMAQ model was integrated to obtain the CO$_2$ simulations $C^f_t(x, y, z, t)$ (referred to as $C^f_t$ hereafter). Then, the column-averaged concentrations $X^f_{CO2}$ were obtained using Eq. (18).

The performance of CFI-CMAQ was evaluated through a group of well-designed OSSEs. And the goal of each OSSE was to retrieve the true fluxes $F^p$ from given
true observations $X^p_{CO2}$ and “wrong” fluxes $F^*_t$. In all the OSSEs, we assimilated artificial observations $X^p_{CO2}$ about three times a day since GOSAT has about three orbits in the study model domain. If there were some observations, CFI-CMAQ paused to assimilate. Otherwise, it continued simulating. The default ensemble size $N$ was 48, the measurement errors were 1.5 ppmv, the standard localization Schur radius $c$ was 1280 km (20 grid spacing), and the covariance inflation factor of concentrations $\alpha$ was 1.1. The referenced lag-window was 9 days and the covariance inflation factor of the prior scaling factors $\beta$ was 70. Since the smoother window was very important for CO$_2$ transportation and $\beta$ was a newly introduced parameter, both these parameters were further investigated by several numerical sensitivity experiments. The primary focus of this paper was to describe the assimilation methodology, so all the numerical experiments started on 1 January 2010 and ended on 30 March 2010.

As for the initialization of CFI-CMAQ, only the ensemble of background concentration fields $C^t_{i,0}$ needed to be initialized at the time $t=0$ because the values of $\Lambda^a_{i,\delta-1}$ were updated by using the persistence dynamical model. In practice, the mean concentration fields at $t=0$ are interpolated from the simulated CO2 fields of CarbonTracker 2011 (Peters, 2007). The ensemble members of the background concentration fields were created by adding random vectors. The mean values of the random vectors were zero and the variances were 2.5 percent of the mean concentration fields. Then the atmospheric transport model integrated from time $t=0$ to $t=1$ driven by $F^*_t$ with $C^t_{i,0}$ as initial conditions to produce the CO$_2$
concentration fields $\hat{C}^f_{i,1}$. And then the first prior linear scaling factors, $\lambda_{i,10}^p$, could be calculated by applying $\hat{C}^f_{i,1}$. Assumed $\lambda_{i,0}^a = \lambda_{i,10}^p$, $\lambda_{i,10}^c$ are gained finally. For the first assimilation cycle, the lag-window was only one (that is, only $\lambda_{i,10}^a$ needed to be optimized in the first assimilation cycle). And it increased for the first dozens of assimilation cycles until it reached M+1 as CFI-CMAQ continued to assimilate observations. Once the system was initialized, all future scaling factors could be created using the persistence dynamical model, which was associated the smoothing operator with the atmospheric transport model.

In order to illustrate the limitation by only using the smoothing operator as the persistence dynamical model to generate all future scaling factors, another OSSE (referred to as the reference experiment to distinguish it from the above-mentioned CFI-CMAQ OSSEs) was designed to optimize the surface CO$_2$ fluxes at grid scale. The reference experiment was under the same assimilation framework as CFI-CMAQ except that all $\lambda_{i,p-1}^p$ were set to 1 (Peters et al., 2007). Besides, the initialization procedure of the reference experiment was different from that of the CFI-CMAQ. In practice, both the ensemble of background concentration fields at $t = 0$, $C^f_{i,0}$, and the ensemble members of the scaling factors at $t = 1$, $\lambda^a_{i,10}$, needed to be initialized because they could not generated by other ways (Peters et al., 2005). The initial concentration fields $C^f_{i,0}$ were created using the same method as that was used to generate $C^f_{i,0}$ for the CFI-CMAQ OSSEs. The ensemble members of the scaling factors $\lambda^a_{i,10}$ were rand fields. Their mean values were 1 and their variances were 0.1. In addition, in order to keep the ensemble spread of the scaling factors $\lambda^a_{i,p-1}$ at a

[Page 19]
certain level and compensate for dynamical model error, covariance inflation was also used and the covariance inflation factor of the scaling factors $\mathcal{A}_{p_{j-1}}$ was 1.6. All other parameters are the same as used in the CFI-CMAQ OSSEs. The ensemble size $N$ was 48, the measurement errors were 1.5 ppmv, the standard localization Schur radius $c$ was 1280 km, the covariance inflation factor of concentrations $\alpha$ was 1.1, and the lag-window was 9 days.

3.2 Experimental results

Essentially, the assimilation part of CFI-CMAQ includes two subsections: one for the CO$_2$ concentration assimilation with EnKF, which can provide a convincing CO$_2$ initial analysis fields for the next assimilation cycle; and the other for the CO$_2$ flux optimization with EnKS, which can provide better estimation of the scaling factors for the next time through the persistence dynamical model except for optimized CO$_2$ fluxes. The performance of the EnKF subsection will be greatly influenced by the validation of the EnKS subsection, or vice versa. Firstly, the performance of CFI-CMAQ will be quantitatively assessed in detail by using the assimilated results of a CFI-CMAQ OSSE, in which the lag-window was 9 days and $\beta$ was 70. Then the sensitivities of $\beta$ and the lag-window will be discussed in the following two paragraphs. And finally, the assimilation results of the reference experiment in which $\mathcal{A}_{p_{j-1}}$ were set to 1 will be described in brief at the end of this subsection.

We begin by describing the impacts of assimilating artificial observations $X_{CO_2}^p$ on CO$_2$ simulations by CFI-CMAQ. As shown in Figs. 4a, 4b and 4d, the monthly mean values of the background CO$_2$ concentrations $C_i^f$ produced by the magnified
surface CO$_2$ fluxes $F_i^*$ were much larger than those of the artificial true CO$_2$
concentrations $C_i^p$ produced by the prior surface CO$_2$ fluxes $F_i^p$ near the surface in
February 2010. In the east and south of China especially, the magnitude of the
difference between $C_i^p$ and $C_i^f$ was at least 6 ppmv. Also, as expected, the monthly
mean $X_{\text{CO}_2}^i$ was much larger than the monthly mean artificial observations $X_{\text{CO}_2}^p$, and the magnitude of the difference between $X_{\text{CO}_2}^p$ and $X_{\text{CO}_2}^f$ reached 2 ppmv in
the east and south of China (see Figs. 3b, 3c and 3e). However, the impact of
magnifying surface CO$_2$ fluxes on the CO$_2$ concentrations was primarily below the
model-level 10 (approximately 6 km), and especially below model-level 7
(approximately 1.6 km). And above model-level 10, the differences between $C_i^p$ and
$C_i^f$ fell to zero (see Fig. 5a and 5b). After assimilating $X_{\text{CO}_2}^p$, the analysis CO$_2$
concentrations $\overline{C_i^a}$ was much closer to $C_i^p$ (see Figs. 4c, 4e and 4f). The monthly
mean difference between $C_i^p$ and $\overline{C_i^a}$ ranged from $-2$ to 2 ppmv and the relative
error $(C_i^p - \overline{C_i^a}) / C_i^p$ ranged from $-1$ to 1% in almost the entire model domain at
model-level 1. The monthly mean differences between $C_i^p$ and $\overline{C_i^a}$ were negligible
above model-level 2 (see Fig. 5c and 5d). The monthly mean $X_{\text{CO}_2}^a$ was also closer
to $X_{\text{CO}_2}^p$ and the difference between $X_{\text{CO}_2}^p$ and $X_{\text{CO}_2}^a$ ranged from $-0.5$ to 0.5
ppmv. In order to evaluate the general impact of assimilating $X_{\text{CO}_2}^p$ in the surface
layer, time series of the daily mean CO$_2$ concentration extracted from the background
simulations and the assimilations were compared with the artificial true simulations at
four national background stations in China and their nearest large cities. As shown in
Fig. 3a, Waliguan is 150 km away from Xining, Longfengshan is 180 km away from
Haerbin, Shangdianzi is 150 km away from Beijing, and Linan is 50 km away from Hangzhou. The assimilated results are shown in Fig. 6. The background time series were much larger than the artificial true time series, especially at Shangdianzi, Beijing, Linan and Hangzhou, which are strongly influenced by local anthropogenic CO$_2$ emissions. After assimilating $X_{\text{CO}_2}^p$, the assimilated time series were very close to the true time series with negligible bias, as expected, at Waliguan, Xining, Shangdianzi, Beijing, Linan and Hangzhou, especially after the first 10 days, which can be considered the spin-up period. Meanwhile, the improvements at Longfengshan and Haerbin were limited due to the absence of observation data at those locations (see Fig. 3a). Nevertheless, in general, the substantial benefits to the CO$_2$ concentrations in the surface layer of assimilating GOSAT $X_{\text{CO}_2}$ with EnKF are clear. All the results illustrated that CFI-CMAQ can provide a convincing CO$_2$ initial analysis fields for CO$_2$ flux inversion.

The impacts of assimilating $X_{\text{CO}_2}^p$ on surface CO$_2$ fluxes were also highly impressive by CFI-CMAQ. On the whole, the prescribed CO$_2$ surface fluxes $F_i^p$ were much larger than the true surface CO$_2$ fluxes $F_i^p$ in February 2010, especially in the east and south of China. The monthly mean difference between $F_i^*$ and $F_i^p$ reached 5 μmole m$^{-2}$ s$^{-1}$ in Jing-Jin-Ji, the Yangtze River Delta, and Pearl River Delta Urban Circle because of the strong local anthropogenic CO$_2$ emissions (see Figs. 7a, 7b and 7d). After assimilating $X_{\text{CO}_2}^p$, the ensemble mean of the assimilated surface CO$_2$ fluxes $\overline{F_i^a}$ decreased sharply. Thus, the monthly mean values of $\overline{F_i^a}$ were much smaller than $F_i^*$ in most of the model domain in February 2010. The pattern of
the difference between $F_t^a$ and $F_t^*$ was similar to that of the difference between $F_t^a$ and $F_t^*$ (see Fig. 7d). The ensemble mean of the assimilated surface CO$_2$ fluxes $F_t^a$ were also compared to the artificial true fluxes $F_t^p$, revealing that $F_t^a$ was equivalent to $F_t^p$ in most of the model domain. The monthly mean difference between $F_t^a$ and $F_t^p$ ranged from $-0.1$ to $0.1$ μmole m$^{-2}$ s$^{-1}$ only (see Fig. 7e). In addition, the root-mean-square errors (RMSEs) of the assimilated flux members were analyzed. As shown in Fig. 8, the monthly mean RMSE was less than 0.5 μmole m$^{-2}$ s$^{-1}$ in most of the model domain, except in areas near to large cities such as Beijing, Shanghai and Guangzhou, indicating that the assimilated CO$_2$ fluxes were reliable.

In order to evaluate the ability of CFI-CMAQ to optimize the surface CO$_2$ fluxes comprehensively, the ratios of the monthly mean $F_t^*$ to the monthly mean $F_t^p$ were analyzed. In actual implementation, we only analyzed the ratios where the absolute values of the monthly mean $F_t^p$ were larger than 0.1, to avoid random noise. As shown in Fig. 9a, the ratios of the monthly mean $F_t^*$ to the monthly mean $F_t^p$ are about 1.8 in most of China, except in the Qinghai–Tibet Plateau, where the absolute values of the monthly mean $F_t^p$ in February were very small and we did not analyze. In addition, the ratios of the monthly mean $F_t^a$ to the monthly mean $F_t^p$ are shown in Fig. 9b. This figure demonstrates that the impact of the assimilation of $X_{CO2}^p$ by CFI-CMAQ on CO$_2$ fluxes was great in the east and south of China in general, but the influence was negligible in Northeast China due to the lack of observation data.

Time series of daily mean surface CO$_2$ fluxes extracted from $F_t^*$ and $F_t^a$ were also compared with that from $F_t^p$ at four national background stations in China and
their nearest large cities, similar to the CO$_2$ concentration assimilation. The results are shown in Fig.10. The background time series were much larger than the artificial true time series, especially at Haerbin, Shangdianzi, Beijing, Linan and Hangzhou, which are strongly influenced by local anthropogenic CO$_2$ emissions. After assimilating $X_{CO_2}^p$, the assimilated time series were near to the true time series with acceptable bias, as expected, at Waliguan, Xining, Shangdianzi, Linan and Hangzhou after the 10-day spin-up period. However, the improvements at Longfengshan and Haerbin were negligible because of a lack of observations at these locations. Also, this inversion system failed to show improvements at Beijing. One of the possible reasons was that the values of the ensemble spread of $\Lambda_{i j l-1}^2$ in Beijing area are too large (see Fig. 11c). Beijing was located in Jing-Jin-Ji Urban Circle, which had strong local anthropogenic CO$_2$ emissions during January to March. So the values of the ensemble spread of $C_{i j}^f$ in Beijing area at model-level 1 could be much larger than those in other areas, which had weak local anthropogenic CO$_2$ emissions (see Fig. 11a). As a result, the values of the ensemble spread of $\Lambda_{i j l-1}^p$ before inflating in Beijing area are much larger than those in other areas with small local anthropogenic CO$_2$ emissions (see Fig. 11b). After inflating, the ensemble spread of $\Lambda_{i j l-1}^p$ in Beijing area could be too large, compare to those in other areas with small local anthropogenic CO$_2$ emissions (see Fig. 11c), which lead to the failure to reproduce the true fluxes in Beijing area. Later, CFI-CMAQ will be improved by optimizing the covariance inflation method.

Since the impact of assimilation $X_{CO_2}^p$ by CFI-CMAQ on CO$_2$ fluxes was in
general greater in the east and south of China than other model areas (see Figs. 7 and 9), the time series of the daily mean CO$_2$ fluxes in that area averaged from $F^a_t$ was compared with those from $F^*_t$ and $F^p_t$ (see Fig. 12). This figure indicates that CFI-CMAQ could in general reproduce the true fluxes with acceptable bias.

As stated in the above section, $\beta$ was a newly introduced parameter. The prior scaling factors should have been inflated indirectly through the inflated CO$_2$ concentration forecast. However, the values of the ensemble spread of $\lambda^p_{r,jr-1}$ before inflating were very small (ranging from 0 to 0.08 in most area at model-level 1, see Fig. 11b), though the values of the ensemble spread of $C^i_{r,j}$ after inflating could reach 1 to 14 ppmv in most area at model-level 1 (see Fig. 11a). So we had to inflate them again before using them into Eq. (2). Fig. 11c showed the distribution of the ensemble spread of $\lambda^a_{r,jr-1}$ at model-level 1 at 00 UT on 1 March 2010 when $\beta = 70$. It showed that the values of the ensemble spread of $\lambda^a_{r,jr-1}$ ranged from 0.1 to 0.8 in most area. In order to investigate the sensitivity of the inflation factor of the scaling factors $\beta$, a series of numerical experiments were conducted. As shown in Fig. 12, CFI-CMAQ worked rather well for $\beta = 60, 70, 75, 80$. However, if $\beta$ was much smaller than 50 (e.g. $\beta = 10$), the impact of assimilation was small due to the small ensemble spread; or if $\beta$ was much larger than 80 (e.g. $\beta = 100$), the assimilated CO$_2$ fluxes deviated markedly from the “true” CO$_2$ fluxes. In other words, the performance of CFI-CMAQ greatly relies on the choice of $\beta$.

From the perspective of the lag-window, the differences among the four assimilation sensitivity experiments with lag-windows of 3, 6, 9 and 12 days were
very small (see Fig. 13). Although Peters et al. (2007) indicated that the lag-window should be more than five weeks, it seemed that the smoother window had a slight influence on the assimilated results for CFI-CMAQ. It was clear that the assimilated results with a larger lag-window were better than those with a smaller lag-window; however, CFI-CMAQ performed very well even with a short lag-window (e.g. 3 days).

At the end of this subsection, the assimilation results of the reference experiment in which $\lambda_{r,0}^p$ were set to 1 will be addressed briefly. The impact of assimilation $X_{\text{CO2}}^p$ on CO$_2$ fluxes was disordered. The monthly mean values of the difference between the prior true surface CO$_2$ fluxes and the ensemble mean values of the assimilated surface CO$_2$ fluxes were irregular noise (see Fig. 14). The main reason is that all the elements of the scaling factors to be optimized in the smoother window are only random numbers. As stated in the above section, only $\lambda_{r,1|0}^a$ needed to be optimized in the first assimilation cycle. However, $\lambda_{r,0}^a$ were rand fields (in other words, all the elements of $\lambda_{r,0}^a$ are only random numbers) because they could not generated by other ways at the first time. So their spatial correlations were too small. The correlations between the scaling factors and the observations were also too small. Therefore it was impossible to systematically change the values of $\lambda_{r,0}^a$ in large areas where the observations located after assimilating observations at $t=1$. Thus the signal-to-noise problem arose. So the elements of $\lambda_{r,1|0}^a$ are only random numbers too. Though $\lambda_{r,2|1}^p$ could be generated automatically by the smoothing operator when all $\lambda_{r,2|1}^p$ were set to 1, the elements of $\lambda_{r,2|1}^a$ are random numbers too since the smoothing
operator is only a linear operator. Similarly, it was impossible to systematically
change the values of $\mathbf{A}_{\text{a},1|1}$ and $\mathbf{A}_{\text{a},2|1}$ in large areas after assimilating observations at
$t = 2$. As this inversion system continued assimilating observations, all future scaling
factors could be created by the smoothing operator and then updated. But this
inversion system could not ingest the observations effectively because all the elements
of the scaling factors were always random numbers. Though the 9 days lag-window in
the reference experiment is too short compared to the 5 weeks lag-window
recommended by Peters et al(2007), this reference experiment could illustrate the
limitation by only using the smoothing operator as the persistence dynamical model.
If the lag-window was around 5 weeks, we could get better results because there were
more observations in every assimilation cycle. However, the results could not be
better than those obtained by CFI-CMAQ because most grids have no observations
(refer to Fig. 3a) and the signal-to-noise problem still remained.

4 Summary and conclusions

A regional surface CO$_2$ flux inversion system, CFI-CMAQ, has been developed
to optimize CO$_2$ fluxes at grid scales. It operates under a joint data assimilation
framework by applying EnKF to constrain the CO$_2$ concentrations and applying EnKS
to optimize the surface CO$_2$ flux, which is similar to Kang et al. (2011, 2012) and
Tian et al. (2013). The persistence dynamical model, which was first introduced by
Peters et al. (2007) by applying the smoothing operator to transport the useful
observed information onto the next assimilation cycle, is further developed. We
associated the smoothing operator with the atmospheric transport model to constitute the persistence dynamical model to forecast the surface CO2 flux scaling factors for the purpose of resolving the ‘signal-to-noise’ problem, as well as transporting the useful observed information onto the next assimilation cycle. In this application, the scaling factors to be optimized in the flux inversion system can be forecast at the grid scale without random noise. The OSSEs showed that the performance of CFI-CMAQ is effective and promising. In general, it could reproduce the true fluxes at the grid scale with acceptable bias.

This study represents the first step in developing a regional surface CO2 flux inversion system to optimize CO2 fluxes over East Asia, particularly over China. In future, we intend to further develop the covariance localization techniques and inflation techniques to improve the performance of CFI-CMAQ. Furthermore, the uncertainty of the boundary conditions should be considered to improve the effectiveness of regional CO2 flux optimization.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (Grant No. 41130528), the Strategic Priority Research Program–Climate Change: Carbon Budget and Relevant Issues (XDA05040404), the National High Technology Research and Development Program of China (2013AA122022). CarbonTracker results used to generate the initial condition are provided by NOAA ESRL, Boulder, Colorado, USA from the website at http://carbontracker.noaa.gov. The numerical calculations in this paper have been done on the IBM Blade cluster system in the High Performance Computing Center (HPCC) of Nanjing University.

References

Andres, R. J., Boden, T. A., Bréon, F. M., Ciais, P., Davis, S., Erickson, D., Gregg, J. S., Jacobson, A., Marland, G., Miller, J., Oda, T., Olivier, J. G. J., Raupach, M. R., Rayner, P. and...


Pillai, D., Gerbig, C., Ahmadov, R., Rödenbeck, C., Kretschmer, R., Koch, T., Thompson, R., Neininger, B., and Lavrié, J. V.: High-resolution simulations of atmospheric CO2 over complex terrain – representing the Ochsenkopf mountain tall tower, Atmos. Chem. Phys., 11,


List of Figures

Fig. 1. Schematic diagram of the smoother window.

$$(\lambda_{a,j-1-M-1}, \lambda_{a,j-M-1}, \lambda_{a,j-M+1-M-1}, \ldots, \lambda_{a,j-M+1}),$$ are the optimized scaling factors in the smoother window and $C_{i,t-1}$ are the assimilated CO$_2$ concentrations fields at time $t-1$ in the previous assimilation cycle $t-1-M-t-1$.

$$(\lambda_{a,j-M-1}, \lambda_{a,j-M+1-M-1}, \ldots, \lambda_{a,j-M+1}, \lambda_{a,j-M})$$ are the scaling factors in the smoother window and $C_{i,t}$ are the forecast CO$_2$ concentrations fields at time $t$ which need to be optimized in the current assimilation cycle $t-M-t$.

Fig. 2. Flowchart of the CFI-CMAQ system used to optimize surface CO$_2$ fluxes at each assimilation cycle. The system includes the following four parts in turn: (1) forecasting of the linear scaling factors $\lambda_{a,j-1}$ (red arrows); (2) optimization of the scaling factors in the smoother window by EnKS (see Fig. 1) (blue arrows); (3) updating of the flux in the smoother window (green arrows); and (4) assimilation of the CO$_2$ concentration fields at time $t$ by EnKF (black arrows).

Fig. 3. (a) Total number of observations in February 2010 in the model grid. Each symbol indicates the total number of all GOSAT X$_{CO2}$ measurements in the corresponding model grid. Monthly mean values in February 2010 of (b) $X_{CO2}^f$, column mixing ratio of $C_{i,t}$; (c) $X_{CO2}^a$, column mixing ratio of $C_{i,t}$; (d) $\bar{X}_{CO2}^a$, column mixing ratio of $\bar{C}_{i,t}$; (e) $X_{CO2}^p - X_{CO2}^f$; and (f) $X_{CO2}^p - \bar{X}_{CO2}^a$. All column mixing ratios are column-averaged with real GOSAT X$_{CO2}$ averaging kernels at GOSAT X$_{CO2}$ locations. Each symbol indicates the monthly average value of all X$_{CO2}$
estimates in the model grid. $\overline{C_i^a}$ are the ensemble mean values of the assimilated CO$_2$ concentrations fields of a CFI-CMAQ OSSE, in which the lag-window was 9 days and $\beta$ was 70. And they are the same OSSE in Fig. 3 to Fig. 6.

Fig. 4. Monthly mean values of (a) $C_i^p$, the artificial true simulations driven by the prior surface CO$_2$ fluxes $F_i^p$; (b) $C_i^f$, the background simulations driven by magnified surface CO$_2$ fluxes $F_i^f = (1.8 + \delta(x, y, z, t))F_i^p$; (c) $\overline{C_i^a}$, the ensemble mean values of the assimilated CO$_2$ concentrations fields; (d) $C_i^p - C_i^f$; (e) $C_i^p - \overline{C_i^a}$; and (f) $100*(C_i^p - \overline{C_i^a})/C_i^p$ at model-level 1 in February 2010. Black lines EF and GH indicate the positions of the cross sections shown in Fig. 5.

Fig. 5. Monthly mean cross sections of $C_i^p - C_i^f$ along line (a) EF and (b) GH, and monthly mean cross sections of $C_i^p - \overline{C_i^a}$ along line (c) EF and (d) GH (cross section lines shown in Fig. 4d) in February 2010.

Fig. 6. Daily mean time series of CO$_2$ concentrations at national background stations in China and their nearest large cities from 1 Jan. to 20 Mar. 2010 extracted from the artificial true simulations $C_i^p$ (black), background simulations $C_i^f$ (red), and the ensemble mean values of the assimilated CO$_2$ concentrations fields $\overline{C_i^a}$ (blue). All time series were interpolated to the observation locations by the spatial bilinear interpolator method. The sites used are (a) Waliguan (36.28°N, 100.91°E), (b) Xining (36.56°N, 101.74°E), (c) Longfengshan (44.73°N, 127.6°E), (d) Haerbin (45.75°N, 126.63°E), (e) Shangdianzi (40.65°N, 117.12°E), (f) Beijing (39.92°N, 116.46°E), (g)
Linan (30.3°N, 119.73°E), and (h) Hangzhou (30.3°N, 120.2°E).

Fig. 7. Monthly mean values in February 2010 of (a) $F_t^p$, the prior true surface CO$_2$ fluxes; (b) $F_t^*$, the prescribed CO$_2$ surface fluxes, $F_t^* = (1.8 + \delta(x, y, z, t))F_t^p$; (c) $\overline{F_t^a}$, the ensemble mean values of the assimilated surface CO$_2$ fluxes; (d) $F_t^p - F_t^*$; and (e) $F_t^p - \overline{F_t^a}$ (units: μmole m$^{-2}$ s$^{-1}$). $\overline{F_t^a}$ are the assimilated results of an CFI-CMAQ OSSE, in which the lag-window was 9 days and $\beta$ was 70. And they are the same in Fig. 7 to Fig. 10.

Fig. 8. Monthly mean RMSEs of $\overline{F_t^a}$ in February 2010 (units: μmole m$^{-2}$ s$^{-1}$).

Fig. 9. (a) Ratios of monthly mean $F_t^*$ to monthly mean $F_t^p$; and (b) ratios of monthly mean $\overline{F_t^a}$ to monthly mean $F_t^p$ in Feb. 2010. The white part indicates the ratios where the absolute values of monthly mean $F_t^p$ are larger than 0.1, not analyzed in this study. The black square labeled I indicates the domain where surface CO$_2$ fluxes were used for the results presented in Fig. 12 and 13.

Fig. 10. Daily mean time series of CO$_2$ fluxes at national background stations in China and their nearest large cities from 1 Jan to 20 Mar. 2010 extracted from the prior true surface CO$_2$ fluxes $F_t^p$ (black), the prescribed CO$_2$ surface fluxes $F_t^*$ (red), and the assimilated CO$_2$ fluxes $\overline{F_t^a}$ (blue). All time series were interpolated to the observation locations by the spatial bilinear interpolator method. The sites used are (a) Waliguan, (b) Xining, (c) Longfengshan, (d) Haerbin, (e) Shangdianzi, (f)
Beijing, (g) Linan, and (h) Hangzhou.

Fig. 11. (a) Ensemble spread of $C_{i,t}$ after inflating; (b) ensemble spread of $\lambda_{i,j,t-1}^p$ before inflating; (c) ensemble spread of $\lambda_{i,j,t-1}^a$ at model-level 1 at 00 UT on 1 March 2010 when $\beta = 70$.

Fig. 12. Time series of daily mean CO$_2$ fluxes averaged in domain I (shown in Fig. 9a) from 1 Jan. to 20 Mar. 2010 with the inflation factor of scaling factors $\beta = 50, 60, 70, 75$ and 80. The black dashed line is the time series averaged from $F_t^*$ and the black solid line is the time series averaged from $F_t^p$.

Fig. 13. Time series of daily mean CO$_2$ fluxes averaged in domain I (shown in Fig. 9a) from 1 Jan. to 20 Mar 2010 with different smoother windows (3, 6, 9 and 12 days). The black dashed line is the time series averaged from $F_t^*$ and the black solid line is the time series averaged from $F_t^p$.

Fig. 14. Monthly mean values of the difference between the prior true surface CO$_2$ fluxes and the ensemble mean values of the assimilated surface CO$_2$ fluxes (units: $\mu$mole m$^{-2}$ s$^{-1}$) of the reference experiment in which $\lambda_{i,j,t-1}^p$ were set to 1.
Fig. 1. Schematic diagram of the smoother window.

\[
\begin{align*}
\{\lambda_{t-M-1}^a, \lambda_{t-M+1}^a, \lambda_{t-M+2}^a, \ldots, \lambda_{t-M+|I|-1}^a\} \quad & \text{are the optimized scaling factors in the smoother window,} \\
\{\lambda_{t-M+1}^a, \lambda_{t-M+2}^a, \lambda_{t-M+3}^a, \ldots, \lambda_{t-M+|I|-1}^a\} \quad & \text{are the scaling factors in the smoother window,} \\
\{C_{i,t}^a\} \quad & \text{are the assimilated CO}_2 \text{ concentrations fields at time } t-1 \text{ in the previous assimilation cycle } t-1-M-t. \\
\{C_{i,t}^f\} \quad & \text{are the forecast CO}_2 \text{ concentrations fields at time } t \text{ which need to be optimized in the current assimilation cycle } t-M-t.
\end{align*}
\]
Fig. 2. Flowchart of the CFI-CMAQ system used to optimize surface CO₂ fluxes at each assimilation cycle. The system includes the following four parts in turn: (1) forecasting of the linear scaling factors \( \lambda_{ijt}^a \) (red arrows); (2) optimization of the scaling factors in the smoother window by EnKS (see Fig. 1) (blue arrows); (3) updating of the flux in the smoother window (green arrows); and (4) assimilation of the CO₂ concentration fields at time \( t \) by EnKF (black arrows).
Fig. 3. (a) Total number of observations in February 2010 in the model grid. Each symbol indicates the total number of all GOSAT $X_{CO2}$ measurements in the corresponding model grid.

Monthly mean values in February 2010 of (b) $X_{CO2}^p$, column mixing ratio of $C_i^p$; (c) $X_{CO2}^l$, column mixing ratio of $C_i^l$; (d) $X_{CO2}^{p^*}$, column mixing ratio of $C_i^{p^*}$; (e) $X_{CO2}^p - X_{CO2}^l$; and (f) $X_{CO2}^p - X_{CO2}^{p^*}$. All column mixing ratios are column-averaged with real GOSAT $X_{CO2}$ averaging kernels at GOSAT $X_{CO2}$ locations. Each symbol indicates the monthly average value of all $X_{CO2}$ estimates in the model grid. $C_i^{p^*}$ are the ensemble mean values of the assimilated CO$_2$ concentrations fields of a CFI-CMAQ OSSE, in which the lag-window was 9 days and $\beta$ was 70. And they are the same OSSE in Fig. 3 to Fig. 6.
Fig. 4. Monthly mean values of (a) $C_{t}^{p}$, the artificial true simulations driven by the prior surface CO$_2$ fluxes $F_{t}^{p}$; (b) $C_{t}^{l}$, the background simulations driven by magnified surface CO$_2$ fluxes $\delta F_{t}^{p}$; (c) $\overline{C}_{t}^{a}$, the ensemble mean values of the assimilated CO$_2$ concentrations fields; (d) $C_{t}^{p} - C_{t}^{l}$; (e) $C_{t}^{p} - \overline{C}_{t}^{a}$; and (f) $100 \times (C_{t}^{p} - \overline{C}_{t}^{a})/C_{t}^{p}$ at model-level 1 in February 2010. Black lines EF and GH indicate the positions of the cross sections shown in Fig. 5.
Fig. 5. Monthly mean cross sections of $C_i^p - C_i^f$ along line (a) EF and (b) GH, and monthly mean cross sections of $C_i^p - C_i^a$ along line (c) EF and (d) GH (cross section lines shown in Fig. 4d) in February 2010.
Fig. 6. Daily mean time series of CO$_2$ concentrations at national background stations in China and their nearest large cities from 1 Jan. to 20 Mar. 2010 extracted from the artificial true simulations $C^p_i$ (black), background simulations $C^f_i$ (red), and the ensemble mean values of the assimilated CO$_2$ concentrations fields $C^a_i$ (blue). All time series were interpolated to the observation locations by the spatial bilinear interpolator method. The sites used are (a) Waliguan (36.28°N, 100.91°E), (b) Xining (36.56°N, 101.74°E), (c) Longfengshan (44.73°N, 127.6°E), (d) Haerbin (45.75°N, 126.63°E), (e) Shangdianzi (40.65°N, 117.12°E), (f) Beijing (39.92°N, 116.46°E), (g) Linan (30.3°N, 119.73°E), and (h) Hangzhou (30.3°N, 120.2°E).
Fig. 7. Monthly mean values in February 2010 of (a) $F^p_t$, the prior true surface CO$_2$ fluxes; (b) $F^*_t$, the prescribed CO$_2$ surface fluxes, $F^*_t = (1.8 + \delta(x, y, z, t))F^p_t$; (c) $\bar{F}^a_t$, the ensemble mean values of the assimilated surface CO$_2$ fluxes; (d) $F^p_t - F^*_t$; and (e) $F^p_t - \bar{F}^a_t$ (units: $\mu$ mole m$^{-2}$ s$^{-1}$). $\bar{F}^a_t$ are the assimilated results of an CFI-CMAQ OSSE, in which the lag-window was 9 days and $\beta$ was 70. And they are the same in Fig. 7 to Fig. 10.
Fig. 8. Monthly mean RMSEs of $\overline{F_i}$ in February 2010 (units: μmole m$^{-2}$ s$^{-1}$).
Fig. 9. (a) Ratios of monthly mean $F_i^*$ to monthly mean $F_i^p$; and (b) ratios of monthly mean $F_i^*$ to monthly mean $F_i^p$ in Feb. 2010. The white part indicates the ratios where the absolute values of monthly mean $F_i^p$ are larger than 0.1, not analyzed in this study. The black square labeled I indicates the domain where surface CO$_2$ fluxes were used for the results presented in Fig. 12 and 13.
Fig. 10. Daily mean time series of CO$_2$ fluxes at national background stations in China and their nearest large cities from 1 Jan to 20 Mar. 2010 extracted from the prior true surface CO$_2$ fluxes $F_t^p$ (black), the prescribed CO$_2$ surface fluxes $F_t^*$ (red), and the assimilated CO$_2$ fluxes $\overline{F}_t^\text{ass}$ (blue). All time series were interpolated to the observation locations by the spatial bilinear interpolator method. The sites used are (a) Waliguan, (b) Xining, (c) Longfengshan, (d) Haerbin, (e) Shangdianzi, (f) Beijing, (g) Linan, and (h) Hangzhou.
Fig. 11. (a) Ensemble spread of $C^f_{1,t}$ after inflating; (b) ensemble spread of $\mathbf{A}^p_{n,t-1}$ before inflating; (c) ensemble spread of $\mathbf{A}^a_{n,t-1}$ at model-level 1 at 00 UT on 1 March 2010 when $\beta = 70$. 

$49$
Fig. 12. Time series of daily mean CO$_2$ fluxes averaged in domain I (shown in Fig. 9a) from 1 Jan. to 20 Mar. 2010 with the inflation factor of scaling factors $\beta = 50, 60, 70, 75$ and 80. The black dashed line is the time series averaged from $F^*_i$ and the black solid line is the time series averaged from $F^{p}_i$. 

1
2
3
4
5
6
7
Fig. 13. Time series of daily mean CO$_2$ fluxes averaged in domain I (shown in Fig. 9a) from 1 Jan. to 20 Mar 2010 with different smoother windows (3, 6, 9 and 12 days). The black dashed line is the time series averaged from $F_t^*$ and the black solid line is the time series averaged from $F_t^p$. 
Fig. 14. Monthly mean values of the difference between the prior true surface CO₂ fluxes and the ensemble mean values of the assimilated surface CO₂ fluxes (units: μmole m⁻² s⁻¹) of the reference experiment in which $\lambda_{t,j,p}^{P}$ were set to 1.