Environ Ecol Stat (2013) 20:129–146
DOI 10.1007/s10651-012-0211-6

A constrained least-squares approach to combine bottom-up and top-down CO_2 flux estimates

Daniel Cooley · F. Jay Breidt · Stephen M. Ogle · Andrew E. Schuh · Thomas Lauvaux

Received: 26 September 2011 / Revised: 11 April 2012 / Published online: 4 July 2012

Abstract Terrestrial CO_2 flux estimates are obtained from two fundamentally different methods generally termed bottom-up and top-down approaches. Inventory methods are one type of bottom-up approach which uses various sources of information such as crop production surveys and forest monitoring data to estimate the annual CO_2 flux at locations covering a study region. Top-down approaches are various types of atmospheric inversion methods which use CO_2 concentration measurements from monitoring towers and atmospheric transport models to estimate CO_2 flux over a study region. Both methods can also quantify the uncertainty associated with their estimates. Historically, these two approaches have produced estimates that differ considerably. The goal of this work is to construct a statistical model which sensibly combines estimates from the two approaches to produce a new estimate of CO_2 flux for our study region. The two approaches have complementary strengths and weaknesses, and our results show that certain aspects of the uncertainty associated with each of the approaches are greatly reduced by combining the methods. Our model is purposefully simple and designed to take the two approaches' estimates and measures of uncertainty at 'face value'. Specifically, we use a constrained least-squares approach to appropriately weigh the estimates by the inverse of their variance, and the constraint

D. Cooley (✉) · F. J. Breidt
Department of Statistics, Colorado State University, Fort Collins, CO, USA
e-mail: cooleyd@stat.colostate.edu

S. M. Ogle
Natural Resources Ecology Laboratory, Colorado State University, Fort Collins, CO, USA

A. E. Schuh
Department of Atmospheric Sciences, Colorado State University, Fort Collins, CO, USA

T. Lauvaux
Department of Meteorology, Pennsylvania State University, State College, University Park, PA, USA

imposes agreement between the two sources. Our application involves nearly 18,000 flux estimates for the upper midwest United States. The constrained dependencies result in a non-sparse covariance matrix, but computation requires only minutes due to the structure of the model.

Keywords Atmospheric inversion · Carbon inventory · Climate change

1 Introduction

There is general consensus in the scientific community that changes to the Earth's climate are accelerating due to anthropogenic greenhouse gas emissions (IPCC 2007). Currently, there is much discussion about reducing carbon emissions to curb the increase of atmospheric CO_2, one of the primary greenhouse gases. Of the anthropogenic CO_2 emissions, approximately 45 % remain in the atmosphere (Le Quéré et al. 2009) and the remainder is absorbed by the oceans and terrestrial environment; thus understanding net exchange of CO_2 between the terrestrial biosphere and atmosphere is critical to predict future CO_2 levels and consequently future climate change. An aspect of the global carbon cycle that has been a challenge to fully quantify is the exchange between regional sources and sinks of CO_2 (Schimel et al. 2001). There are two fundamentally different approaches for estimating terrestrial CO_2 fluxes over a study region and these are generally referred to as bottom-up and top-down methods. The goal of this research is to better quantify terrestrial CO_2 flux in a particular study region by combining bottom-up and top-down estimates while accurately accounting for their uncertainties. We employ a constrained least-squares approach to combine the estimates and produce reconciled CO_2 flux estimates at locations which cover the region.

The mid-continent intensive (MCI) campaign is a scientific effort to study CO_2 flux in the upper-Midwest region of the United States.[1,2] The MCI was selected as a test region because the drivers of terrestrial CO_2 fluxes are relatively well-understood. The region is largely agricultural and tracked in national agricultural databases, and modeling the atmospheric transport of CO_2 is relatively simple as the region's orography is not complex. The underlying scientific question of this project could be summarized as "How well can we estimate the CO_2 flux for the MCI region?"

Inventories begin with measurements of quantities that influence carbon cycling such as crop yields, forest growth and timber harvest, and fossil fuel consumption (IPCC 2006). Because they begin from ground-based measurements, inventories are referred to as bottom-up approaches. The data driving inventories are not direct measurements of CO_2 flux and moreover are from only a sample of locations which may be either point-referenced (such as soil carbon measurements at study sites) or areal-referenced (such as reported crop yield for a county). Thus, it is necessary to convert these measurements into CO_2 fluxes estimates by modeling (Ogle et al. 2010). Inventories aim to provide a CO_2 flux estimate for every location in a study region. Many

[1] For contributing groups see http://www.nacarbon.org/cgi-nacp/web/investigations/inv_ic_profiles.pl.

[2] http://www.nacarbon.org/cgi-nacp/web/investigations/inv_pgp.pl?pgid=108.

inventories produce estimates on an annual timescale, partly because some of the underlying data such as crop yields (USDA-NASS 2010) are only provided annually.

In contrast, atmospheric inversions begin with atmospheric CO_2 concentration measurements from monitoring towers and are thus deemed top-down approaches. Atmospheric inversions are typically deterministic numerical models which use observed meteorology, estimated boundary conditions, and often a "prior" estimate of CO_2 flux based on land-use characteristics to backwards-project CO_2 flux estimates for each location of the inversion's domain (Tarantola 2005). Inversions can produce estimates on a daily or sub-daily timescale, but these estimates are usually aggregated up to weekly timescales or longer.

Both inventory and inversion approaches generally provide a measure of the uncertainty associated with their point estimates. Recent updating of the North America Carbon Plan has highlighted accurate accounting of the uncertainty as one of the key shortcomings of previous work.[3] Uncertainty quantification has been a primary focus of the greater reconciliation project, and will continue to be studied by both the inventory and inversion communities.

It is not clear whether the bottom-up or top-down approach provides a better or more appropriate estimate of the true carbon flux, which is unmeasurable. At a fundamental level, the inventories and inversions model different processes, make assumptions about different things, and draw on different data sources. A combined estimate draws on the two approaches' complementary strengths.

Additionally, the two estimation procedures provide different information about the carbon flux in the region. Inventories have information about different flux sources such as fossil fuel emissions, cropland, and forest fluxes while the inversions are generally unable to distinguish between distinct CO_2 flux sources. Quantifying the influence of different sources is important for addressing potential mitigation for policy development. Because they arise from densely sampled ground-based measurements, it is generally believed that inventories have richer spatial information than inversions. Inventories can express spatially-concentrated CO_2 plumes from sources such as powerplants and feedlots, whereas inversions have difficulty capturing these plumes as their signal is dispersed when CO_2 from these sources reaches the monitoring towers which drive the inversions. In contrast, the region-wide CO_2 flux is likely better constrained by the inversion than the inventory. From its tower measurements, the inversion has a measurement constraint on the total flux for the region, whereas model uncertainty common to all locations in the inventory leads to greater uncertainty for the aggregated regionwide flux. Inversions are able to capture temporal effects not seen by the inventories which generally provide information only on an annual timescale. Knowledge about changes in CO_2 flux due to seasonality or effects on a subannual timescale due to floods or droughts can be obtained from the inversion. Finally, the inventories never "see" the atmosphere, but instead estimate atmospheric fluxes indirectly. For this reason, inversions are important both for verifying inventory estimates and for determining compliance under regulatory policies should carbon-emission agreements be enacted.

[3] http://www.carboncyclescience.gov/USCarbonCycleSciencePlan-August2011.pdf.

Although there is considerable interaction between the bottom-up and top-down research communities, there has been little prior effort to combine or reconcile the two sources of information. The work most closely related to ours is that of Chan and Lin (2011) who combined agriculture census data for 20 counties in southwest Ontario with a satellite-driven biospheric agriculture CO_2-flux model via a normal-normal Bayesian framework. Our scope is broader than Chan and Lin as we attempt to account for all sources of CO_2 flux rather than focusing only on agriculture, and the Chan and Lin approach has no direct measurements of atmospheric CO_2.

The objective of this study is to propose a statistical model which combines inventory estimates with inversion estimates, and which appropriately accounts for the uncertainty associated with each set of estimates. The 'data' we will analyze is actually model output from both an inversion and from inventories for several sources. In this sense, our statistical model can be thought of as a meta-model. Combining the output from the two approaches into one estimate is attractive because the resulting estimate will draw on the strengths of the two sources. Combining the two sources of information is also of scientific interest because, historically, flux estimates from inventory sources and atmospheric inversions have shown large differences, although the results from these two approaches have not necessarily been considered inconsistent due to the large uncertainties (Pacala et al. 2001). Furthermore, it is hoped that combining the two sources of information will provide additional insight about areas to focus on for future research in CO_2 flux estimation for both the bottom-up and top-down methods.

We propose a constrained least squares approach to produce a unified estimate from the inventory and inversion estimates. Although the inversion and inventory represent different aspects of CO_2 flux and on different time scales, they should agree on the total annual flux from all sources at each location, and this is the constraint we impose. The appeal of our constrained least squares approach is its simplicity, especially given that CO_2 flux estimates could be contentious if national or international commitments to reduce carbon emissions are agreed upon and subsequent regulations require verification of emissions reductions. Our statistical model does not require additional prior information or assume any sort of dynamics in the weekly flux estimates. Our constrained least-squares approach is interesting statistically because it is used in a manner different from typical constrained least-squares applications.

The remainder of the paper is structured as follows. In Sect. 2 we provide details about the inventory estimates and the inversion which we integrate into our statistical model. In Sect. 3.1 we describe our constrained model and inference. In Sect. 4 we provide the combined estimate for CO_2 flux and its associated uncertainties for our study region. We conclude in Sect. 5 with a discussion.

2 Two sources of CO_2 flux model output

In this section we provide the details about the specific atmospheric inversion and inventory used in our analysis. Throughout the paper, our fluxes are from the atmosphere's perspective: a positive flux implies a net release of CO_2 from the surface to the atmosphere, and a negative flux means there is net uptake of CO_2 from the atmosphere to the surface.

2.1 Inventory

We include key carbon sources in the inventory: forest biomass and associated harvested woody products, CO_2 uptake in agricultural crop and associated harvest, livestock respiration, human respiration, and agricultural soil carbon changes. CO_2 emissions from combustion of fossil fuels is also included in this project's inventory (Gurney et al. 2009), but are omitted from the combined model since they are removed prior to running the inversion for reasons we explain below. Each of the source inventories employ survey data on the activity (such as forest plot level data or crop yields from agricultural lands) as inputs into models which produce CO_2 flux estimates for the entire region (EPA 2011). Many of the measurements such as livestock population or harvest yield are recorded at the county level, and thus the inventory data is compiled by county. This data is then interpolated to a common 0.5 degree lattice also used for the inversions. Details for each of these sources are given below.

The forest inventory comes from the Forest Inventory and Analysis (FIA) program of the US Forest Service (Smith et al. 2007). Negative fluxes result from forest establishment and growth, and are maintained in the terrestrial pool when this carbon is incorporated into harvested woody products. Positive fluxes result from disturbance, such as fire and pest outbreaks, as well as the decay of harvested woody products.

By its nature, harvest flux must be negative. Although the CO_2 uptake in croplands is mostly lost during a single year because the majority of crops in the region are short-lived annuals, crop commodities resulting from harvest can be transported out of the region leading to an atmospheric CO_2 sink in the region (West et al. 2011). In terms of CO_2 sources and sinks, it is important to note that harvest represents an "apparent" sink because this carbon will be returned to the atmosphere within a few years as the commodities are used.

The human respiration data were derived from population data in the region (West et al. 2009), as were livestock estimates (West et al. 2011). Both the human respiration and livestock fluxes represent respired CO_2 and are therefore positive.

The agricultural soil carbon estimates are based on land use and management activity data for cropland and grasslands in the region (EPA 2011; Ogle et al. 2010). Although it is the smallest of the fluxes at an annual timescale, soil carbon is an important quantity for mitigation as it (along with forest) represents a longer-term carbon sequestration source/sink (Paustian et al. 1997).

Even though this study region is relatively sparsely populated, fossil fuel emissions are by far the largest contributer to CO_2 fluxes. Fossil fuels are a source of 170 Tg C for the region, while the annual flux estimates from forest, harvest, human respiration, livestock, and soils are -23, -100, 5, 16, and -0.015 Tg C respectively. Because CO_2 from fossil fuels is highly concentrated around population centers and power plants, and because data from tracers unique to fossil fuels are limited, the fossil fuel signal is hard to resolve by the atmospheric inversions. Fossil fuel consumption is also relatively well known (Gurney et al. 2009), and thus the atmospheric inversions

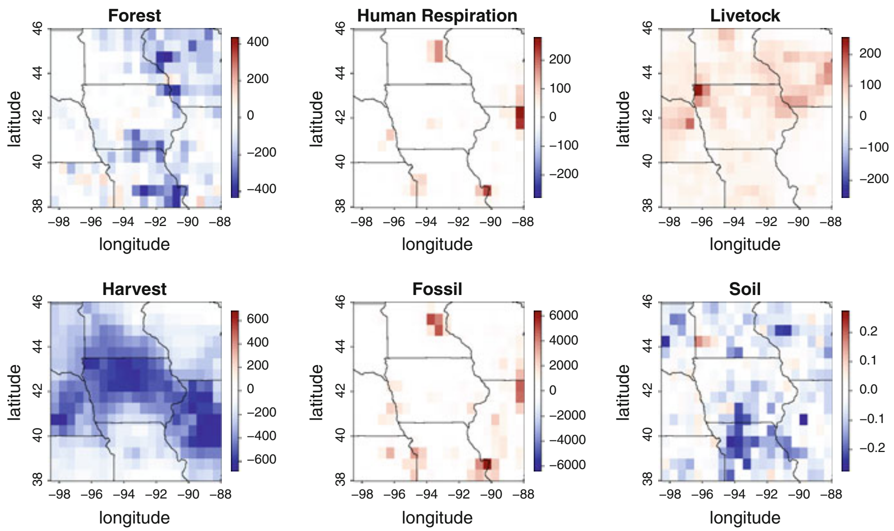

Fig. 1 Point estimates of annual CO_2 flux from the inventory sources for the MCI region consisting of Iowa (*center*) and parts of six other states (clockwise, from *top left*) South Dakota, Minnesota, Wisconsin, Illinois, Missouri, Kansas, and Nebraska. Units are Gg C per grid cell from CO_2 flux. Note that the *scales* differ for each source

subtract it out before running their models. As our aim is to combine the atmospheric inversion estimates with the inventory estimates, we leave fossil fuels out of our analysis. However, it is important to note that any quantification of regional CO_2 flux will require an accounting of fossil fuel consumption.

Figure 1 gives the point estimates for all inventory sources. The estimates for CO_2 flux due to forest are mostly negative with the largest sinks in Wisconsin and Missouri; however a few cells show a positive flux estimate. The CO_2 flux estimates from harvest show the deepest sinks in the largely corn and soybean-producing regions of Iowa and Illinois. Human respiration and fossil fuel fluxes are largely concentrated around the largest metropolitan centers in this region: Minneapolis/St. Paul, St. Louis, and Chicago (although the grid cell which contains the city of Chicago lies just east of our study region). Livestock is less concentrated, but a few cells indicate large livestock concentrations, for instance in the extreme northwest corner of Iowa. Finally the CO_2 flux estimates from soils are both positive and negative, but the actual fluxes are considerably smaller than the other sources and therefore have less influence on the overall inventory.

2.2 Inversion

Although there have been other methods proposed for atmospheric inversions (Bocquet 2008; Chevallier et al. 2010), most atmospheric inversions can be understood as Gaussian state-space or data assimilation models (Schuh et al. 2010; Lauvaux et al. 2009; Göckede et al. 2010; Gourdji et al. 2012). Inversion modelers often use Bayesian terminology when describing their models. An inversion estimate at a given time begins with a "prior" flux estimate for all grid cells based on known land-use characteristics, plant phenology, and meteorology for this time. Then, given tower-based CO_2 concentration measurements for this time period and an atmospheric transport model, a "posterior" flux estimate is produced for each location in the study region. In addition to the point estimate, the state space model produces a covariance matrix describing the uncertainty of the flux estimates for each time period.

The inversion for this study is a high resolution inversion that produces estimates for the MCI study area (Lauvaux et al. 2011). The boundary conditions are based on CO_2 concentration 3D fields from the CarbonTracker system (Peters et al. 2007) corrected by weekly aircraft profiles from the NOAA aircraft program (Sweeney et al. 2011). The prior fluxes were computed with the SiBcrop vegetation model (Lokupitiya et al. 2009) enhanced to describe crop phenology over the region. The atmospheric transport model WRF-Chem was run at 10 km resolution over the domain and coupled to a Lagrangian Dispersion Model in backward mode (Uliasz 1994) to generate the linearized adjoint of the transport. The inversion produces estimates on a 20 km grid using an analytical matrix inversion technique, but these results are then interpolated to a standard 0.5-degree grid for easy comparison. We analyze model output on the region shown in Fig. 2, corresponding to the 0.5 degree cells that are completely covered by the inversion's native grid.

The Lauvaux et al. (2011) inversion is the most data-constrained CO_2 inversion thus far as tower measurements were provided by the "ring of towers".[4] Historically, there have only been a few highly-calibrated CO_2 concentration towers scattered across the continental United States, making high-resolution atmospheric inversions ill-conditioned. The ring of towers was a short-term project whose aim was to understand what information a dense network of towers over a small region could provide, and the MCI was selected as a logical study region for the aforementioned reasons. The ring of towers became operational in the summer of 2007 and was decommissioned at the end of 2009.

For this paper we work with preliminary inversion results which have been produced for 2007. A final inversion for 2007 and 2008 will be produced as part of the greater MCI project. Although the inversion produces flux estimates on a sub-daily time scale, in this study we use flux estimates that are compiled weekly. In fact, for the inversion we use, a "week" is equal to 7.5 days, and the year is broken into 48 periods which we will continue to refer to as "weeks". Because the ring of towers was not operational until the summer of 2007, the first 20 weeks of inversion estimates and covariance matrices are actually taken from the prior.

The total CO_2 flux estimates for weeks 12, 25, and 35 shown in Fig. 2 illustrate the annual carbon cycle. Week 12 is prior to the growing season, and the decay from previous years' plant litter results in a small positive flux for most of the region. Week 25 is the height of the growing season and there is a large negative flux region-wide, with deep sinks in the corn-producing regions of Iowa and Illinois. Week 35 shows

[4] http://ring2.psu.edu/.

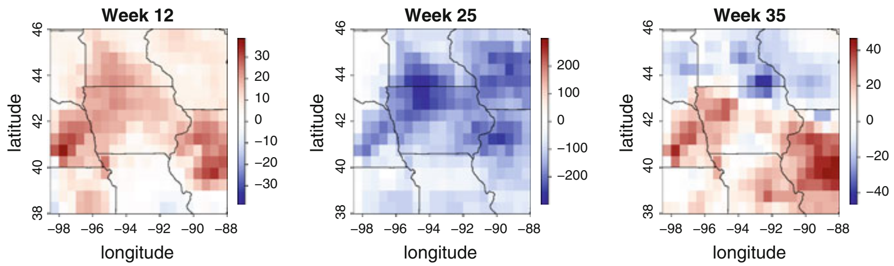

Fig. 2 Point estimates of CO_2 flux from all sources from the inversion for weeks 12 (*left*), 25 (*center*), and 35 (*right*). Units are Gg C per grid cell. Note that the *scales* for each figure differ

a small positive flux in the south and negative flux of similar magnitude in the north, consistent with conditions at the end of the growing season.

2.3 Uncertainty estimates provided by the inversion and inventory

Uncertainty estimation is an area of active research for both the inversion and inventory communities. The left panels of Fig. 3 show both point-wise standard deviation and spatial correlation for week 25 of the inversion. The week 25 inversion estimate has a fairly uniform level of uncertainty across the region, with slightly larger standard deviations at locations with larger estimated fluxes such as in Iowa and Illinois. The spatial correlation map shows correlations with the marked location in Northern Iowa, and the correlations range from $(-0.11, 0.57)$ over the study region. The correlation is positive at short distances and drops off to become slightly negative at longer distances. The negative correlation at longer distances constrains the level of uncertainty for the inversion's estimate of regional flux, even though individual locations' measurements are not as well constrained.

A Monte Carlo approach was employed by each of the inventory contributors to obtain the covariance matrices for the inventory sources. Density functions for measured quantities, model parameters, and other model inputs were created, drawn from, and then used within the inventory models to produce 100 realizations of possible inventory estimates. Because the inventories for different sources are produced using different methods, there is some difference in the uncertainties associated with the sources. These differences largely reflect that some sources are better understood and quantified than others. We used the Monte Carlo realizations to produce empirical covariance matrices for each inventory source. As the number of locations was greater than the number of draws, we used a standard shrinkage covariance estimator.

The center and right panels show point-wise standard deviation and spatial correlation for the forest and harvest inventories. The standard deviation for the forest inventory is much larger than that for the harvest inventory. Measuring the annual CO_2 flux from forest growth and change is difficult due to the 5-year return sampling of the FIA as well as significant measurement uncertainty. In contrast, because it is linked to the nation's food supply, crop harvest data are relatively well-known at the county level and the uncertainty associated with the CO_2 flux is largely due to

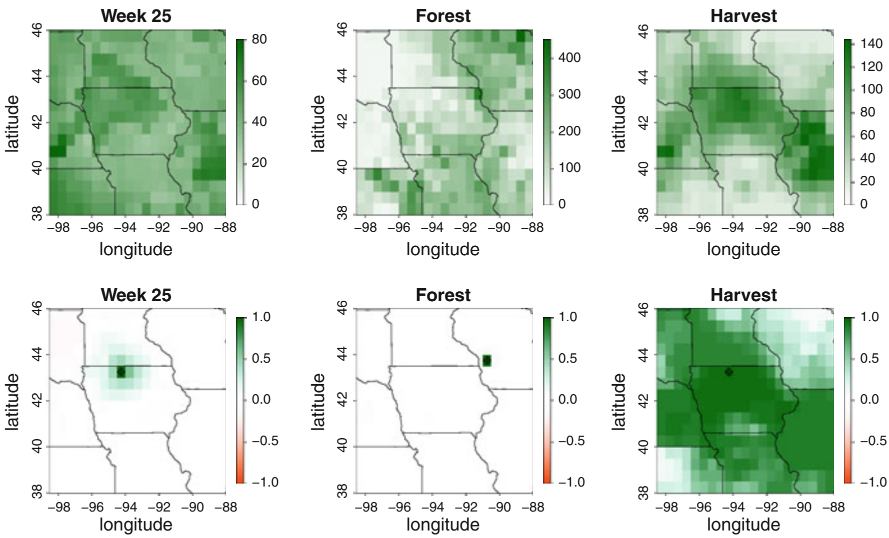

Fig. 3 *Top row*: Standard deviations associated with the estimates. *Bottom row*: Spatial correlation for the estimates with the marked location. *Left column* is for the inversion estimate for week 25, *center column* is for the forest inventory, and *right column* is for the harvest inventory

uncertainty in the parameterizations which convert harvest data into CO_2 flux. The spatial correlation of these two inventories also reflects their differences. The forest's spatial correlation is consistent with a process dominated by measurement uncertainty; the spatial correlation is nearly negligible with values ranging from (-0.03 to 0.07). The large spatial correlation in the harvest estimates results from the variation in model parameters which are reflected throughout the region.

3 Statistical model and inference

3.1 A constrained model for estimation

Constrained least squares is a well-known statistical approach, and obtaining inference for such problems requires solving a Lagrange multiplier problem. Constrained least squares problems regularly arise in hypothesis-testing problems for linear models, see for instance sections 6.3 and 6.11 of Graybill (1976). Constrained least squares optimization also arises in universal kriging, a spatial prediction method described in most spatial statistics books such as Section 3.2 of Cressie (1993) or Section 5.2.2 of Schabenberger and Gotway (2005). However, the setting in which we employ constrained least squares differs from these typical applications. Here we do not aim to test a hypothesis or perform spatial prediction, rather our goal is estimation. Our constraints impose agreement between the inversion and inventory flux estimates at each location in the study region, and this differs from the small number of constraints from a set of hypotheses to be tested or single constraint in ordinary kriging due to requiring an unbiased estimator.

Assume the N locations in our study region are denoted by $z_1, \ldots, z_N$. Let $\boldsymbol{Y}_{\cdot t} = (Y_{\cdot t}(z_1), \ldots, Y_{\cdot t}(z_N))^T$, where $Y_{\cdot t}(z_j)$ denotes the inversion flux estimate for location z_j and week $t = 1, \ldots, T$. The $\cdot$ in the subscript denotes that the inversion estimate is the total of all sources $s = 1, \ldots, S$. We assume that $\boldsymbol{Y}_{\cdot t} = \boldsymbol{X}_{\cdot t} + \boldsymbol{W}_{\cdot t}$, where $\boldsymbol{X}_{\cdot t} = (X_{\cdot t}(z_1), \ldots, X_{\cdot t}(z_N))^T$ is the vector of true fluxes from all sources for all locations in study region $\mathcal{D}$ for week t. $\boldsymbol{W}_{\cdot t} \sim N(\boldsymbol{0}, R_{\cdot t})$ is an $N \times 1$ random vector which represents the error associated with the inversion estimate for week t and $R_{\cdot t}$ is the error covariance matrix provided by the inversion.

Likewise, let $\boldsymbol{Y}_{s\cdot}$ denote the vector of annual inventory estimates at all locations for source $s = 1, \ldots, S$. We order the sources alphabetically: $s = 1, 2, 3, 4, 5$ correspond to forest, harvest, human respiration, livestock, and soils respectively. We similarly assume that $\boldsymbol{Y}_{s\cdot} = \boldsymbol{X}_{s\cdot} + \boldsymbol{W}_{s\cdot}$, where $\boldsymbol{X}_{s\cdot}$ is the true annual CO_2 flux from source s, $\boldsymbol{W}_{s\cdot} \sim N(\boldsymbol{0}, R_{s\cdot})$ is the error associated with the inventory estimate, and $R_{s\cdot}$ is the covariance matrix obtained from the Monte Carlo uncertainty simulation for source s. Finally, assume that $\boldsymbol{W}_{\cdot t}$, $t = 1, \ldots, T$ and $\boldsymbol{W}_{s\cdot}$, $s = 1, \ldots, S$, are mutually uncorrelated.

Let $\boldsymbol{Y} = (\boldsymbol{Y}_{\cdot 1}^T, \ldots, \boldsymbol{Y}_{\cdot T}^T, \boldsymbol{Y}_{1\cdot}^T, \ldots, \boldsymbol{Y}_{S\cdot}^T)^T$, $\boldsymbol{X} = (\boldsymbol{X}_{\cdot 1}^T, \ldots, \boldsymbol{X}_{\cdot T}^T, \boldsymbol{X}_{1\cdot}^T, \ldots, \boldsymbol{X}_{S\cdot}^T)^T$, $\boldsymbol{W} = (\boldsymbol{W}_{\cdot 1}^T, \ldots, \boldsymbol{W}_{\cdot T}^T, \boldsymbol{W}_{1\cdot}^T, \ldots, \boldsymbol{W}_{S\cdot}^T)^T$, and R be the block diagonal matrix $\text{diag}(R_{\cdot 1}, \ldots, R_{\cdot T}, R_{1\cdot}, \ldots, R_{\cdot S})$. Then, the unconstrained model can be written as

$$\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{W}, \quad \text{where } E(\boldsymbol{W}) = \boldsymbol{0}, \quad \text{and} \quad \text{Cov}(\boldsymbol{W}) = R. \tag{1}$$

We impose the N-dimensional constraint

$$\sum_{t=1}^{T} \boldsymbol{X}_{\cdot t} = \sum_{s=1}^{S} \boldsymbol{X}_{s\cdot} \tag{2}$$

which implies the total of the weekly fluxes (from all sources) equals the total of the annual source fluxes at each location. This constraint can be written as

$$A\boldsymbol{X} = \boldsymbol{0}, \tag{3}$$

where A is the $N \times N(T+S)$ matrix

$$A = \left[\underbrace{I \ldots I}_{T} \underbrace{-I \ldots -I}_{S} \right],$$

and I is the $N \times N$ identity matrix.

3.2 Inference

We use the method of Lagrange multipliers to minimize the weighted least squares

$$(\boldsymbol{Y} - \boldsymbol{X})^T R^{-1} (\boldsymbol{Y} - \boldsymbol{X}) \tag{4}$$

subject to the constraint (3). Let $\boldsymbol{\lambda}$ be the $N \times 1$ vector of Lagrange multipliers. From (4) and (3), we obtain the Lagrange function

$$(\boldsymbol{Y} - \boldsymbol{X})^T R^{-1} (\boldsymbol{Y} - \boldsymbol{X}) + 2\boldsymbol{\lambda}^T A \boldsymbol{X}, \tag{5}$$

where the factor 2 is chosen to allow cancellation below. Taking derivatives with respect to the elements of $\boldsymbol{X}$ and $\boldsymbol{\lambda}$ and setting equal to zero yields the system of equations

$$\begin{bmatrix} R^{-1} & A^T \\ A & 0 \end{bmatrix} \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} R^{-1}\boldsymbol{Y} \\ \boldsymbol{0} \end{pmatrix}. \tag{6}$$

Inverting the 2×2 block matrix on the left-hand-side of (6) yields

$$\begin{bmatrix} R^{-1} & A^T \\ A & 0 \end{bmatrix}^{-1} = \begin{bmatrix} R - RA^T(ARA^T)^{-1}AR & RA^T(ARA^T)^{-1} \\ (ARA^T)^{-1}AR & -(ARA^T)^{-1} \end{bmatrix}. \tag{7}$$

Thus,

$$\hat{\boldsymbol{X}} = \boldsymbol{Y} - RA^T(ARA^T)^{-1}A\boldsymbol{Y}, \tag{8}$$

and furthermore, it is straightforward to show that

$$\mathrm{Var}(\hat{\boldsymbol{X}}) = R - RA^T(ARA^T)^{-1}AR. \tag{9}$$

Notice $\hat{\boldsymbol{X}}$ is unbiased, since $\hat{\boldsymbol{X}} = (\boldsymbol{X} + \boldsymbol{W}) - RA^T(ARA^T)^{-1}A(\boldsymbol{X} + \boldsymbol{W})$, $E[\boldsymbol{W}] = 0$, and $A\boldsymbol{X} = 0$. Furthermore, consider any estimator of the form $\boldsymbol{v}^T\boldsymbol{Y}$; i.e., an estimator which is a linear combination of the raw data $\boldsymbol{Y}$. Note that

$$\mathrm{Var}(\boldsymbol{v}^T\hat{\boldsymbol{X}}) = \boldsymbol{v}^T R\boldsymbol{v} - \boldsymbol{v}^T RA^T(ARA^T)^{-1}AR\boldsymbol{v} \le \boldsymbol{v}^T R\boldsymbol{v} = \mathrm{Var}(\boldsymbol{v}^T\boldsymbol{Y}),$$

since $RA^T(ARA^T)^{-1}AR$ is a non-negative definite matrix.

4 Results

4.1 Computation

The number of locations in our dataset is $N = 336$, the number of weeks is $T = 48$, and the number of sources is $S = 5$. Thus, $\boldsymbol{Y}$, $\boldsymbol{X}$, and R are of dimension $N(T + S) = 17808$. Our constraint matrix A is of dimension $N \times N(T + S) = 336 \times 17808$.

As R is block diagonal and both it and A are mostly sparse, we use the sparse matrix package `spam` in `R` (Furrer 2008). As sparse matrices, R is 68.5 MB, and A is 0.2 MB. Ultimately, we need to compute $RA^T(ARA^T)^{-1}A$ and $RA^T(ARA^T)^{-1}AR$. The matrix ARA^T is dimension 336×336 and is easily inverted. Let $M = (ARA^T)^{-1}$.

However, $RA^T M^{-1} A$ and $RA^T M^{-1} AR$ are large and not sparse; if constructed completely these matrices are over 3 GB in size. However, due to the block-diagonal nature of R and structure of A, both of these matrices have a simple block pattern and do not need to be constructed entirely. We refer to the block-rows and block-columns of these matrices by indexes $(\cdot, t)$ for $t = 1, \ldots, T$ and $(s, \cdot)$ for $s = 1, \ldots, S$. $RA^T M^{-1} A$ is such that the $(\cdot t)$th row (of blocks) is given by

$$\underbrace{R_{\cdot t} M \ldots R_{\cdot t} M}_{T} \underbrace{-R_{\cdot t} M \ldots - R_{\cdot t} M}_{S},$$

and the $(s, \cdot)$th row is given by

$$\underbrace{-R_{s\cdot} M \ldots - R_{s\cdot} M}_{T} \underbrace{R_{s\cdot} M \ldots R_{s\cdot} M}_{S}.$$

The matrix $RA^T M^{-1} AR$ has $R_{\cdot t} M R_{\cdot t'}$, $-R_{\cdot t} M R_{s\cdot}$, and $R_{s\cdot} M R_{s'\cdot}$ for the $((\cdot t), (\cdot t'))$, $((\cdot t), (s\cdot))$, and $((s\cdot), (s'\cdot))$ blocks respectively. Working with these blocks rather than the entire matrices tremendously reduces the computational burden. On a laptop with a 2.8 GHz processor and 4 GB of memory, the most difficult of the computations required to produce the results in the following section took only minutes, and most were practically instantaneous.

4.2 Carbon flux estimates

Figure 4 shows maps of the annual flux estimates from all sources from the inversion, inventory, and constrained least squares (combined) estimate. Because the inventory estimates generally have less uncertainty at the grid-cell level, the combined estimate more closely resembles the inventory map than the inversion map. However, some smoothing of the inventory estimates is evident, most notably in Wisconsin and central Missouri locations dominated by forest which has the most uncertainty of the inventory sources. The combined pointwise standard deviation estimate shows the expected decrease in uncertainty. Interestingly, the combined estimate's standard deviation is remarkably uniform, as the agricultural areas in Iowa and Illinois with the largest standard deviations in the inversions are the among the most constrained areas in the inventory.

A primary quantity of interest is the total annual carbon flux estimate for the entire study region from all sources. The left panel of Fig. 5 shows the inversion estimate, the inventory estimate, and the combined estimate for the regional annual flux as well as 95 % confidence intervals for each of these estimates. The point estimates for these three quantities are respectively −132, −102, and −112 TgC. The right panel shows the inversion, inventory, and combined estimate for a single grid cell; the selected cell is the location in Iowa shown in the week 25 and cropland harvest spatial correlation maps in Fig. 3. It is interesting that the combined total estimate in the left panel is closer to the inventory total estimate than the inversion estimate, despite the fact that the inversion estimate has a smaller confidence region for the regional flux. The right

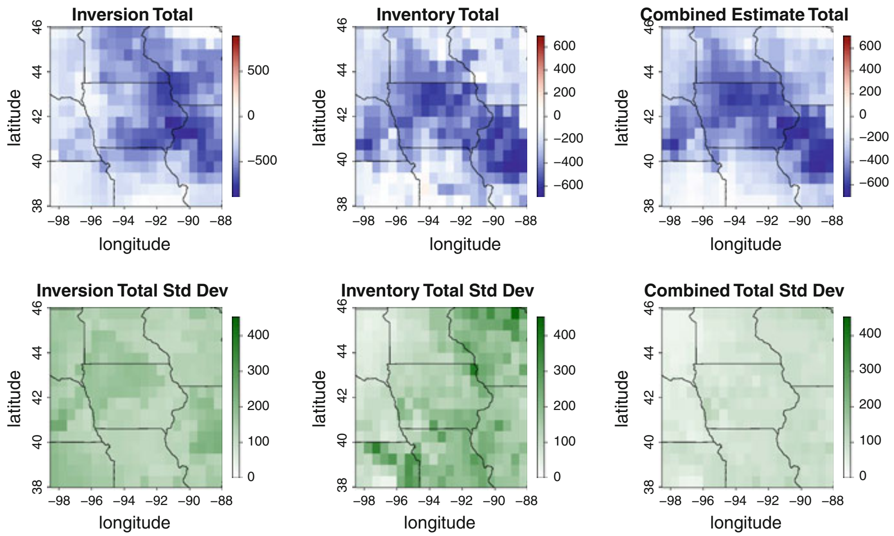

Fig. 4 (*Top row*) Maps of the inversion, inventory, and combined annual flux estimates from all sources, units are GgC per grid cell. (*Bottom row*) Maps of the inversion, inventory, and combined pointwise standard deviations

panel helps to explain this counterintuitive behavior. For the regionwide flux estimate in the left panel, the inventory has greater uncertainty than the inversion largely due to the strong positive spatial correlation shown for harvest in Fig. 3 and due to the fact that the total flux for the region is well-constrained by the tower measurements that drive the inversion. However, at an individual location, the inventory has much less uncertainty than the inversion. As the regional combined estimate is calculated as the sum of the flux estimates at each location (shown in Fig. 4), and the combined estimate at each location will tend to favor the inventory, the regional estimate ends up favoring the inventory.

Figure 6 shows weekly point estimates and 95 % confidence intervals for total carbon flux for the study region, along with the weekly estimates and 95 % confidence intervals provided by the inversion. Although there is not a dramatic difference in the weekly total flux estimates, the constrained least squares estimates tend to be more positive (most noticeably in weeks 12–18) due to the influence of the inventory, and that the confidence intervals are more narrow.

Figure 7 shows the point estimates and 95 % confidence intervals for the regional CO_2 flux for each source, along with the original estimates provided by the inventory. Both the forest and cropland harvest estimates have become more negative to better agree with the inversion which estimated a larger sink for the study region than did the inventories. The confidence interval for the cropland harvest estimate is also greatly reduced. Recall the large spatial correlation in the harvest estimate due to the model uncertainty, which contributes to the large confidence region for the harvest inventory estimate. The inversion helps to constrain the regionwide total flux estimate which, in turn, helps to restrict the cropland harvest estimate uncertainty. The better-constrained cropland harvest estimates could provide some insight about values for model parameters for this important source.

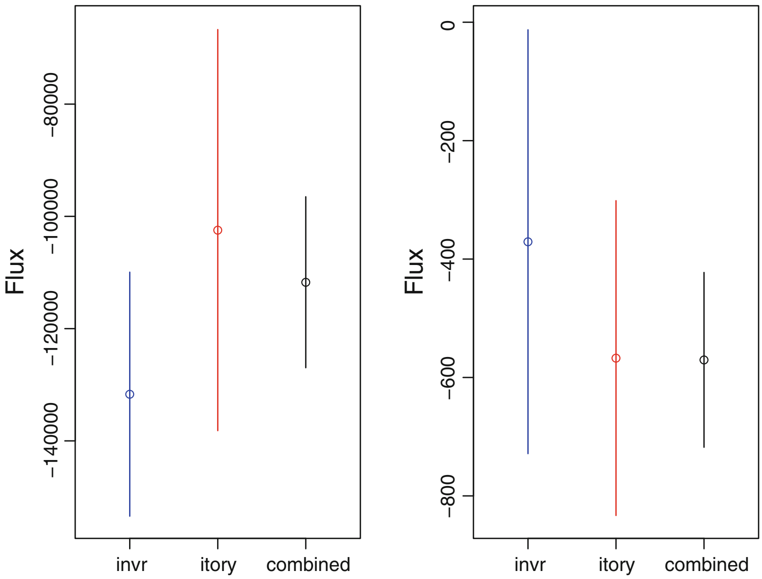

Fig. 5 (*Left*) Estimates for the total annual CO_2 flux for the entire study region from all sources. (*Right*) Estimates for the total annual carbon flux from one location in Iowa from all sources. Units are Gg C

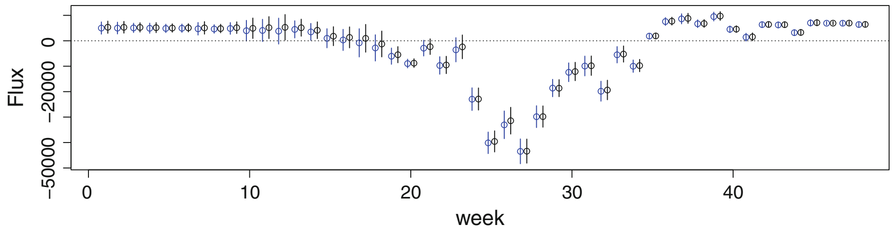

Fig. 6 Weekly flux estimates and 95 % confidence intervals from the original inversion (*left/blue*) and the constrained least squares estimate (*right/black*). Units are Gg C

To better see the differences between the combined estimates and the inversion output, Fig. 8 shows the difference $\boldsymbol{X}_{\cdot t} - \boldsymbol{Y}_{\cdot t}$ for $t = 12$, 25, and 35. Week 12 shows that the combined estimate is greater (less negative) than the original inversion estimate, which primarily reflects that the inventory total flux estimate is not as negative as that from the inversion. However, both week 25 and 35 show a negative difference in western Iowa, indicating that the combined estimate has a deeper sink in this region than the original inversion. Likewise, we produce maps of $\boldsymbol{X}_{s\cdot} - \boldsymbol{Y}_{s\cdot}$, to better understand the differences between the combined estimate and the inventories. The difference in forest estimates generally represents a spatial smoothing done to this source, probably largely due to the smoothness found in the inversion estimates and the large uncertainty associated with the forest estimates at each location. The large difference in the cropland harvest estimate along the Mississippi River near the border between Iowa, Minnesota and Wisconsin is interesting. The inversion estimates a much larger sink in

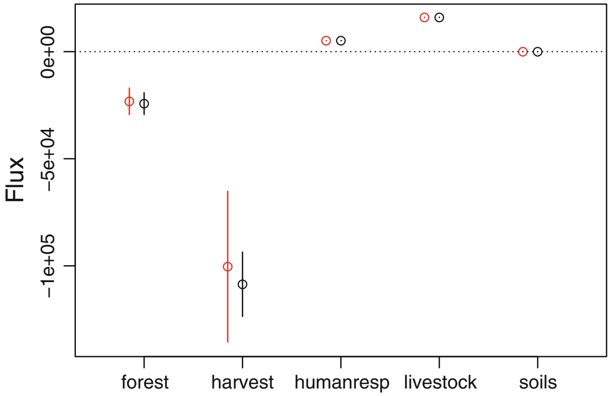

Fig. 7 Flux estimates and 95 % confidence intervals from the different inventory sources for the entire study region. *Left* are the original inventory estimates, *right* are the estimates from the combined model. Units are Gg C

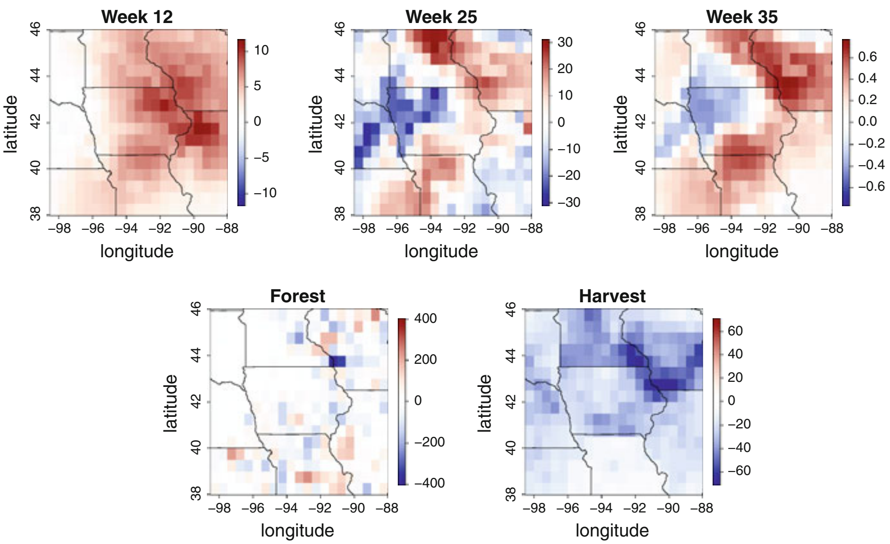

Fig. 8 Differences between the combined model estimates and (*top row*) inversion estimates for weeks 12, 25, and 35, and (*bottom row*) between the combined model estimates for the forest and harvest inventories. Units are Gg C per grid cell

this area than does the inventory, but why this result is attributed to cropland harvest rather than forest is unclear.

5 Conclusions

We have used a constrained least squares approach to combine inventory and inversion CO_2 flux estimates. The new estimate draws on the strengths of both the inventory and inversion, and provides information about how the flux evolves over the course of the year and about the various sources of CO_2 flux. The appeal of the constrained least-squares approach is its simplicity. It takes the uncertainty provided with the two types of estimates at face value, and the combined estimate properly weighs these uncertainties. The imposed constraint arises naturally because both the inventory and inversion aim to estimate the same CO_2 flux using different but complementary methods.

We did consider other statistical approaches for combining these two sources of information. We investigated a state-space formulation which accounted for the fact that we have "observations" (actually model output) that occur at two different time scales: weekly and annual. However, such an approach requires that we impose apriori week-to-week state dynamics, and some concerns about the numerical stability of the algorithm arose during initial investigations. Another approach we considered was a constrained Bayesian model. However, a Bayesian formulation requires selection of priors at some level, and one would need to carefully consider how priors on the weekly and source fluxes would effectively weigh the two sources of information. The advantage of the constrained least squares approach is that it did not require additional apriori information.

Acknowledgments This work has been funded by the National Aeronautics and Space Agency, Earth Sciences Division through the North American Carbon Program, Mid-Continent Intensive Campaign (Agreement No. NNX08AK08G). We would like to thank Linda Heath (US Forest Service) who compiled the forest inventory, Tris West (Pacific Northwest National Laboratories) who compiled the livestock and human respiration inventories, and Ram Gurung (Colorado State University) who compiled the cropland harvest inventory.

References

Bocquet M (2008) Inverse modelling of atmospheric tracers: non-Gaussian methods and second-order sensitivity analysis. Nonlinear Process Geophys 15(1):127–143

Chan E, Lin J (2011) What is the value of agricultural census data in carbon cycle studies?. J Geophys Res 116(G03012):G03012

Chevallier F, Ciais P, Conway TJ, Aalto T, Anderson BE, Bousquet P, Brunke EG, Ciattaglia L, Esaki Y, Fröhlich M et al (2010) CO_2 surface fluxes at grid point scale estimated from a global 21 year reanalysis of atmospheric measurements. J Geophys Res 115(D21):D21307

Cressie N (1993) Statistics for spatial data. Wiley, New York

EPA (2011) Inventory of U.S. greenhouse gas emissions and sinks: 1990–2009. Technical report, U.S. Environmental Protection Agency, Washington, DC

Furrer R (2008) *Spam: SPArse Matrix*. R package version 0.14-1

Göckede M, Michalak AM, Vickers D, Turner DP, Law BE (2010) Atmospheric inverse modeling to constrain regional-scale CO_2 budgets at high spatial and temporal resolution. J Geophys Res 115:D15113

Gourdji S, Mueller K, Yadav V, Huntzinger D, Andrews A, Trudeau M, Petron G, Nehrkorn T, Eluszkiewicz J, Henderson J et al (2012) North american CO_2 exchange: inter-comparison of modeled estimates with results from a fine-scale atmospheric inversion. Biogeosciences 9:457–475

Graybill F (1976) Theory and application of the linear model. Duxbury, North Scituate, Massachusetts

Gurney KR, Mendoza DL, Zhou Y, Fischer ML, Miller CC, Geethakumar S, de la Ruedu Can S (2009) High resolution fossil fuel combustion CO_2 emission fluxes for the United States. Environ Sci Technol 43(14):5535–5541

IPCC (2006) 2006 IPCC guidelines for national greenhouse gas inventories. In: Eggleston HS, Buendia L, Miwa K, Ngara T, Tanabe K (eds) Institute for Global Environmental Strategies, Kanagawa, Japan.

IPCC (2007) Climate change 2007: the physical science basis. In: Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL (eds) Contribution of working group I to the fourth assessment report of the intergovernmental panel on climate change. Cambridge University Press, Cambridge

Lauvaux T, Pannekoucke O, Sarrat C, Chevallier F, Ciais P, Noilhan J, Rayner PJ (2009) Structure of the transport uncertainty in mesoscale inversions of CO_2 sources and sinks using ensemble model simulations. Biogeosciences 6(6):1089–1102

Lauvaux T, Schuh AE, Uliasz M, Richardson S, Miles N, Andrews AE, Sweeney C, Diaz LI, Martins D, Shepson PB, Davis KJ (2011) Constraining the CO_2 budget of the corn belt: exploring uncertainties from the assumptions in a mesoscale inverse system. Atmos Chem Phys 12:337–354. doi:10.5194/acpd-11-20855-2011

Le Quéré C, Raupach M, Canadell J, Marland G et al (2009) Trends in the sources and sinks of carbon dioxide. Nat Geosci 2(12):831–836

Lokupitiya E, Denning S, Paustian K, Baker I, Schaefer K, Verma S, Meyers T, Bernacchi CJ, Suyker A, Fischer M (2009) Incorporation of crop phenology in simple biosphere model (sibcrop) to improve land-atmosphere carbon exchanges from croplands. Biogeosciences 6:1103

Ogle S, Breidt F, Easter M, Williams S, Killian K, Paustian K (2010) Scale and uncertainty in modeled soil organic carbon stock changes for US croplands using a process-based model. Glob Chang Biol 16:810–822

Pacala SW, Hurtt G, Baker D, Peylin P, Houghton R, Heath L, Sundquist RB, Stallard E, Ciais R, Moorcroft P, Caspersen P, Shevliakova J, Moore E, Kohlmaier B, Holland G, Gloor E, Harmon M, Fan M, Sarmiento S, Goodale J, Schimel C, Field DC (2001) Consistent land- and atmosphere-based U.S. carbon sink estimates. Science 292:2316–2320

Paustian K, Andren O, Janzen HH, Lal R, Smith P, Tian G, Tiessen H, Noordwijk MV, Woomer PL (1997) Agricultural soils as a sink to mitigate CO_2 emissions. Soil Use Manag 13:230–244

Peters W, Jacobson AR, Sweeney C, Andrews AE, Conway TJ, Masarie K, Miller JB, Bruhwiler LMP, Pétron G, Hirsch AI et al (2007) An atmospheric perspective on North American carbon dioxide exchange: carbontracker. Proc Natl Acad Sci 104(48):18925

Schabenberger O, Gotway CA (2005) Statistical methods for spatial data analysis. Texts in statistical science. Chapman and Hall/CRC, Boca Raton

Schimel D, House JI, Hibbard KA, Bousquet P, Ciais P, Peylin P, Braswell BH, Apps MJ, Baker D, Bondeau A, Canadell J, Churkina G, Cramer W, Denning AS, Field CB, Friedlingstein P, Goodale C, Heimann M, Houghton RA, Melillo JM, Moore B, Murdiyarso D, Noble I, Pacala SW, Prentice IC, Raupach MR, Rayner PJ, Scholes RJ, Steffen WL, Wirth C (2001) Recent patterns and mechanisms of carbon exchange by terrestrial ecosystems. Nature 414:169–172

Schuh AE, Denning AS, Corbin KD, Baker IT, Uliasz M, Parazoo N, Andrews AE, Worthy DEJ (2010) A regional high-resolution carbon flux inversion of North America for 2004. Biogeosciences 7:1625–1644

Smith J, Heath L, Nichols M (2007) U.S. forest carbon calculation tool: forestland carbon stocks and net annual stock change. Technical report, US Forest Service, Newtown, Pennsylvania. General Technical Report NRS13. Northern Research Station

Sweeney C, Karion A, Wolter S, Neff D, Higgs JA, Heller M, Guenther D, Miller B, Montzka S, Miller J, Conway T, Dlugokencky E, Novelli P, Masarie K, Oltman S, Tans P (2011) Carbon dioxide climatology of the NOAA/ESRL greenhouse gas aircraft network. J Geophys Res (in prep). 20858, 20868

Tarantola A (2005) Inverse problem theory and methods for model parameter estimation. Society for Industrial Mathematics

Uliasz M (1994) Lagrangian particle dispersion modeling in mesoscale applications. Environ Model 2: 71–101

USDA-NASS (2010) Data and statistics. Technical report, United States Department of Agriculture, National Agriculture Statistics Service, Washington, DC

West T, Bandaru V, Brandt C, Schuh A, Ogle S (2011) Regional uptake and release of crop carbon in the United States. Biogeosciences 8:631–654

West T, Singh N, Marland G, Bhaduri B, Roddy A (2009) The human carbon budget: an estimate of the spatial distribution of metabolic carbon consumption and release in the United States. Biogeochemistry 94:29–41

Author Biographies

Daniel Cooley is an associate professor in the department of statistics at Colorado State University. He has research interests in statistical methods for atmospheric sciences, extreme values, and spatial statistics.

F. Jay Breidt is a Professor in the Department of Statistics at Colorado State University. He has research interests in time series, survey sampling, and environmental statistics.

Stephen M. Ogle is a research scientist and Associate Professor in the Natural Resource Ecology Laboratory and department of Ecosystem Science and Sustainability at Colorado State University. Dr. Ogle focuses on research dealing with biogeochemical processes related to greenhouse gas emissions from Agriculture, Forestry and Other Land Uses (AFOLU), and is an expert on research and application of greenhouse gas inventory methods.

Andrew E. Schuh is a research scientist with the Cooperative Institute for Research in the Atmosphere (CIRA) at Colorado State University in Fort Collins, Colorado. He has research interests in the terrestrial carbon cycle, atmospheric transport modeling, and data assimilation.

Thomas Lauvaux research Faculty at the Department of Meteorology of the Pennsylvania State University, works on atmospheric inversions of carbon sources and sinks at regional and local scales, mesoscale modeling, and data assimilation.