Agricultural and Forest Meteorology 149 (2009) 1829–1842

Contents lists available at ScienceDirect

Agricultural and Forest Meteorology

journal homepage: www.elsevier.com/locate/agrformet

Review

A review of applications of model–data fusion to studies of terrestrial carbon fluxes at different scales

Ying-Ping Wang [a,*], Cathy M. Trudinger [a], Ian G. Enting [b]

[a] CSIRO Marine and Atmospheric Research, Centre for Australian Weather and Climate Research, Private Bag #1, Aspendale, Victoria 3195, Australia
[b] MASCOS, 139 Barry St, The University of Melbourne, Victoria 3010, Australia

ARTICLE INFO

Article history:
Received 20 January 2009
Received in revised form 16 April 2009
Accepted 14 July 2009

Keywords:
Model–data fusion
Statistical estimation
Carbon flux
Model error
Measurement error
Parameter estimation
Data assimilation
Kalman filter

ABSTRACT

Model–data fusion is defined as matching model prediction and observations by varying model parameters or states using statistical estimation. In this paper, we review the history of applications of various model–data fusion techniques in studies of terrestrial carbon fluxes in two approaches: top-down approaches that use measurements of global CO_2 concentration and sometimes other atmospheric constituents to infer carbon fluxes from the land surface, and bottom-up approaches that estimate carbon fluxes using process-based models. We consider applications of model–data fusion in flux estimation, parameter estimation, model error analysis, experimental design and forecasting. Significant progress has been made by systematically studying the discrepancies between the predictions by different models and observations. As a result, some major controversies in global carbon cycle studies have been resolved, robust estimates of continental and global carbon fluxes over the last two decades have been obtained, and major deficiencies in the atmospheric models for tracer transport have been identified. In the bottom-up approaches, various optimization techniques have been used for a range of process-based models. Model–data fusion techniques have been successfully used to improve model predictions, and quantify the information content of carbon flux measurements and identify what other measurements are needed to further constrain model predictions. However, we found that very few studies in both top-down and bottom-up approaches have quantified the errors in the observations, model parameters and model structure systematically and consistently. We therefore suggest that future research will focus on developing an integrated Bayesian framework to study both model and measurement errors systematically.

Contents

* Corresponding author. Tel.: +61 3 9239 4577; fax: +61 3 9239 4444.
E-mail address: yingping.wang@csiro.au (Y.-P. Wang).

0168-1923/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.agrformet.2009.07.009

1. Introduction

Feedback between the terrestrial carbon cycle and climate change has been shown to be positive at the global scale, and has significant impact on the predictions of climate change in the future, but the magnitude of this positive feedback and its regional variation remains quite uncertain (Friedlingstein et al., 2006). To improve the understanding of this positive feedback, we need to constrain our model simulations and predictions using biophysical and atmospheric observations available at different time and spatial scales. Two approaches have traditionally been used to estimate carbon fluxes. The top-down approach uses the observed surface atmospheric CO_2 concentration, other atmospheric constituents such as $^{13}CO_2$ or O_2/N_2, and an atmospheric transport model to infer surface CO_2 sinks and sources in time and space (Ciais et al., 1995; Rayner et al., 1999; Bousquet et al., 2000; Gurney et al., 2002). The bottom-up approach uses a process-based model developed using observations at smaller scales to estimate the exchange of carbon between the land biosphere and the atmosphere at larger scale (Cramer et al., 1999). The process-based models can be quite complex, including detailed descriptions of physical, chemical and biological processes at leaf to ecosystem scales. Improving the representation of these processes is essential for using process-based models in a predictive mode.

Over time, there has been increasing sophistication of both the models and the techniques for integrating model and data in these studies (Keeling et al., 1989; Tans et al., 1990; Enting et al., 1995; Enting, 2002; Rödenbeck et al., 2003). Model and data integration, also called model–data fusion or model–data synthesis, is defined as combining models and observations by varying some properties of the model, to give the optimal combination of both (Raupach et al., 2005). Model–data fusion therefore encompasses both parameter estimation and data assimilation. Mathieu and O'Neill (2008) describe the combination of model and data in data assimilation as "a carefully constructed procedure that brings to bear all our knowledge of the system and measurement process as well as the known inaccuracies in (i) measurements (e.g. instrumental error), (ii) governing equations of the system (e.g. parameterisations error of sub-grid scale processes), and (iii) numerical representation of these equations (i.e. discretization and computational errors)". Early examples combining models and data are not as all-encompassing as in this description, but are precursors to such an approach. The description given by Mathieu and O'Neill (2008) can be seen as the goal for model–data fusion. Here we review the application of model–data fusion techniques, and their precursors, in top-down and bottom-up approaches to estimate terrestrial carbon fluxes. As we will see, the top-down and bottom-up approaches address the same problem by different routes, but many features of the techniques used to solve them are common, so a combined review is appropriate.

In the top-down approach, early studies of CO_2 flux inversions used ad hoc scaling of flux patterns to fit available CO_2 concentration data (Keeling et al., 1989; Tans et al., 1990). Enting et al. (1995) introduced a systematic method, known as 'Bayesian synthesis inversion', for estimating fluxes and, importantly, the uncertainties in these fluxes, which was further developed by Rayner et al. (1999) for a time dependent Bayesian synthesis inversion. Since then there have been numerous flux inversions, some considering only CO_2 observations (e.g. Bousquet et al., 2000; Rödenbeck et al., 2003; Peylin et al., 2005), some considering $\delta^{13}CO_2$ in addition to CO_2 to help distinguish fluxes to and from the land and oceans (Rayner et al., 2008). Early studies were reviewed by Enting (2002) and much subsequent work has been undertaken within the TransCom intercomparison community (e.g. Gurney et al., 2003; Baker et al., 2006a). Apart from the matrix inversion approach used in early Bayesian synthesis inversion studies, a range of other computational techniques have been applied to flux inversions, including sequential estimation with the Kalman filter (Baker, 2001; Bruhwiler et al., 2005; Peters et al., 2005) and variational data assimilation (Baker et al., 2006b).

In the bottom-up approach, precursors to model–data fusion in studies of carbon exchange between the atmosphere and the land biosphere by plant scientists before the 1970s were largely limited to fitting the response functions of photosynthesis or respiration to environmental variables, such as light, temperature, atmospheric vapor pressure or soil moisture using data collected from controlled environments (Harley et al., 1992). Questions were raised how useful those complex models really were if many of their parameters and underlying assumptions cannot be verified (Finnigan and Raupach, 1988). Because of the complex response of processes to multiple environmental variables, traditional methods of curve fitting became quite limited for extracting information from the data and therefore for improving the accuracies of model predictions.

Technology advances in the late 1980s made the continuous measurement of carbon fluxes in the field possible. As a result, the predictions of process-based models can be compared directly with field measurements over multiple years at different sites. Some early examples of parameter estimation in terrestrial carbon models using these measurements were Wang et al. (2001) and White and Luo (2002) using gradient-based parameter estimation methods. Other methods were later used for estimating parameters, such as Kalman filter techniques (Williams et al., 2005; Gove and Hollinger, 2006); Markov Chain Monte Carlo sampling method (Braswell et al., 2005; Richardson and Hollinger, 2005), Generalized Likelihood Uncertainty Estimation (Mo and Beven, 2004) and so on.

The top-down approach for estimating CO_2 fluxes has a number of limitations: the inversion is poorly constrained and is diagnostic, therefore does not readily allow for prediction. Furthermore, including other kinds of observations is difficult in synthesis inversions except as prior constraints on fluxes (formally this is not a limitation, but in practice calculations have been restricted to linear relations with Gaussian statistics). Much more information can be readily used in the bottom-up approaches using process models, such as biomass inventory, eddy fluxes and remote sensing. However, the bottom-up approach is generally used to calibrate models at individual sites, and it is difficult to know how representative these sites are, and therefore how accurate fluxes are when integrated over larger regions. A series of studies combining the top-down and bottom-up approaches, known as the Carbon Cycle Data Assimilation System (CCDAS), addresses these

issues by using atmospheric CO_2 observations to constrain parameters in a biosphere model (Kaminski et al., 2002; Rayner et al., 2005). This work has been an important step forward in terrestrial carbon model–data fusion. It allows prediction of fluxes, as well as calculation of uncertainties in the estimated fluxes (Scholze et al., 2007).

In this review we shall focus on the integration of models and data in the studies of carbon fluxes between the land biosphere and the atmosphere and the significant progress made over the last two decades. We will not discuss the errors of measurements due to instrument precision and representation, as this has been comprehensively reviewed by Raupach et al. (2005). The outline of the paper is as follows: In Section 2 we present the basic concepts involved in model–data fusion and precursor techniques and highlight some of the issues. In Section 3, we give examples of model–data fusion applied to studies of carbon exchange between land surface and atmosphere. We then discuss future directions of model–data fusion in the studies of CO_2 exchange between land and atmosphere. Finally we draw conclusions.

2. Model–data fusion—some basic concepts

Model–data fusion can be characterized as both an inverse problem, analyzing a system from observations, and as statistical estimation. 'Calibration' and 'Deconvolution' are two classic inverse problems. Calibration involves determining the behaviour of a model from observations of its outputs for specified inputs. 'Deconvolution' refers to determining the inputs of a given model from observations of the outputs. A variation on deconvolution is data assimilation, where observations are used to refine estimates of the evolving model state. In model–data fusion, we are addressing calibration (through estimation of parameters) and deconvolution (as data assimilation) simultaneously.

Model–data fusion brings together four components: (i) external forcing, (ii) a model that relates model parameters, state and external forcing to observations, (iii) observations and (iv) an optimization technique. An optimization technique will be used to find the statistically optimal match between model predictions and observations for given errors in the model and the observations. Errors in the external forcing are not considered in most applications, but should be considered in forecast applications.

When applying an optimization technique, we usually construct a cost function that quantifies the mismatch between model predictions and observations. A common form, termed an M-estimator in statistics, has the cost function expressed as a weighted sum of model–data mismatch (Garthwaite et al., 2002). The cost function may also reflect the statistical characteristics of the errors in the observations. For example, minimization of the absolute errors rather than the error squared should be used when the error in the observation is Laplacian (see Richardson and Hollinger, 2005).

Optimization methods, that seek the minimum in the cost function, can be classified into batch and sequential techniques, depending whether the data are processed all at once (batch) or in groups or possibly even one at a time (sequential). In the following we will discuss both batch and sequential methods used to estimate carbon fluxes and model parameters.

2.1. Batch methods

In batch techniques, the cost function Θ is treated as a single function to be minimized. Batch methods can be based on iterative numerical methods or closed-form matrix expressions (i.e. direct solution of the matrix equations for the unknowns). Iterative batch methods include both gradient-based methods and global search methods, and process all data together. Closed-form matrix expressions are possible when the dependence of observations on target variables is linear, as in the synthesis inversions where the (K) target variables, x_k, are regional fluxes that are related to (N) surface CO_2 concentrations $m_n(\mathbf{x})$. That is

$$m_n(\mathbf{x}) = \sum_{k=1}^{K} H_{nk} x_k \tag{1}$$

where H_{nk} represent the sensitivity of surface concentration at location n to regional flux x_k, and is independent of x_k. In matrix form, the above equation can be written as

$$\mathbf{m}(\mathbf{x}) = \mathbf{H}\mathbf{x} \tag{2}$$

where $\mathbf{H}$ (with dimension N by K) is a matrix relating the target variables, **x,** to observations and is called observational operator. (We follow the common convention of using bold letters for vectors and matrices, and the same letter but italic and not bold for their elements.)

In weighted least squares, the cost function is constructed as

$$\Theta = \sum_{n=1}^{N} \left(\frac{m_n(\mathbf{x}) - z_n}{r_{n:\mathrm{obs}}} \right)^2 \tag{3}$$

where $r_{n:\mathrm{obs}}$ are weights for the observations, or in matrix form as

$$\begin{aligned} \Theta &= (\mathbf{m}(\mathbf{x}) - \mathbf{z})^{\mathrm{T}} \mathbf{R}^{-1} (\mathbf{m}(\mathbf{x}) - \mathbf{z}) \\ &= (\mathbf{H}\mathbf{x} - \mathbf{z})^{\mathrm{T}} \mathbf{R}^{-1} (\mathbf{H}\mathbf{x} - \mathbf{z}) \end{aligned} \tag{4}$$

where $\mathbf{R}$ is the error covariance matrix for observations (with dimensions N by N). Eq. (3) and (4) are equivalent if $\mathbf{R}$ is diagonal (i.e. if the errors of the data are uncorrelated). The assumption of uncorrelated errors is common and is often used as a starting point in the absence of better information. It may be appropriate, but ideally should be tested to ensure this.

The goal of model–data fusion is to find a set of $\mathbf{x}$ that minimizes Θ. The difference between model predictions and observations, $m_n(\mathbf{x}) - z_n$, is often called the mismatch or innovation, and represents the total errors in the model, target variables and observations. Therefore, in batch methods, data errors are not considered separately from errors in the model and target variables.

The least squares form (Eq. (4)) is often used for reasons of computational simplicity. Least squares fitting is equivalent to another popular method for estimating the optimal combination of model and data, maximum likelihood estimation (Todling, 2000), only if the probability distribution of errors is independent and Gaussian, and the uncertainties are known and identical (weighted least squares is required if uncertainties are not identical). Other choices of cost function arise from maximum likelihood estimation with non-Gaussian distributions, e.g. the sum of absolute deviations gives maximum likelihood estimates for the Laplace error distribution. Alternatively cost functions can be chosen so as to give desirable statistical behaviour (e.g. robustness) of the parameter estimates. In cases where the uncertainties are unknown and treated as statistical parameters to be estimated, cost functions will include additional terms involving these statistical parameters (Garthwaite et al., 2002).

If prior information about target variables is included in the cost function, the form (3) generalizes to:

$$\begin{aligned} \Theta &= \sum_{n=1}^{N} \left(\frac{m_n(\mathbf{x}) - z_n}{r_{n:\mathrm{obs}}} \right)^2 + \sum_{k=1}^{K} \left(\frac{x_k - x_{k:\mathrm{prior}}}{w_{k:\mathrm{prior}}} \right)^2 \\ &= (\mathbf{H}\mathbf{x} - \mathbf{z})^{\mathrm{T}} \mathbf{R}^{-1} (\mathbf{H}\mathbf{x} - \mathbf{z}) + (\mathbf{x} - \mathbf{x}_{\mathrm{prior}})^{\mathrm{T}} \mathbf{W}^{-1} (\mathbf{x} - \mathbf{x}_{\mathrm{prior}}) \end{aligned} \tag{5}$$

The second term on the right-hand side of Eq. (5) represents prior information. $x_{k:\mathrm{prior}}$ is the prior estimate of target variable x_k ;

$w_{k:\text{prior}}$ is the weight of prior information. When prior information is included, the calculation is referred to as 'Bayesian'.

When errors in the observations and prior information are Gaussian and independent, the least squares solution is given by (see Enting, 2002)

$$\hat{\mathbf{x}} = \mathbf{x}_{\text{prior}} + [\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H} + \mathbf{W}^{-1}]^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}[\mathbf{z} - \mathbf{H}\mathbf{x}_{\text{prior}}] \quad (6)$$

And the estimate of covariance of $\mathbf{x}$, cov($\mathbf{x}$,$\mathbf{x}$) is given by

$$cov(\mathbf{x}, \mathbf{x}) = [\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H} + \mathbf{W}^{-1}]^{-1} \quad (7)$$

Therefore, the difference between estimate and prior, $\hat{\mathbf{x}} - \mathbf{x}_{\text{prior}}$, is proportional to the mismatch between observation ($\mathbf{z}$) and model prediction using the prior estimate ($\mathbf{x}_{\text{prior}}$). Without prior estimates of the target variables, we would be required to invert the matrix $\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}$ for calculating $\hat{\mathbf{x}}$. However, for underdetermined problems (i.e. where there are more unknowns than independent constraints), the matrix $\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}$ will be singular (i.e. has no inverse). By adding prior information with weights $\mathbf{W}$, we now require the inverse of $\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}+\mathbf{W}^{-1}$, thereby avoiding the singularity.

When the model, m, is a nonlinear function of target variable $\mathbf{x}$, $\mathbf{H}$ is a function of $\mathbf{x}$, and closed-form expressions do not exist. At the minimum of a cost function, the following relationships exist approximately:

$$\frac{\partial \Theta}{\partial \mathbf{x}} = 0 \quad (8)$$

and

$$\frac{\partial^2 \Theta}{\partial \mathbf{x}^2} > 0 \quad (9)$$

Using these relations, iterative methods can find the minimum in the cost function.

Non-gradient based methods, sometimes called 'global search' methods, such as genetic algorithms and simulated annealing, are usually used for observations from a limited number of target variables, because many more model runs are usually required for these methods than gradient-based methods. Markov Chain Monte Carlo techniques iteratively produce a sample from a Bayesian posterior distribution of parameters until certain convergence criteria are met. The major advantage of global search methods is that they are more likely to find the global minimum for cost functions with multiple minima than gradient-based methods.

2.2. Sequential methods

The Kalman filter is the most popular example of a sequential method, although other examples do exist (e.g. particle filters). The Kalman filter is a recursive algorithm for estimating the state of a system at each time using a state-space model and (noisy) measurements (Kalman, 1960; Kalman and Bucy, 1961; Gelb, 1974). It is called a filter because it aims to reduce the influence of noise in the measurements. It allows convenient representation of model errors, data errors and parameter errors.

The Kalman filter is based on the following equations

$$\mathbf{x}_t = \mathbf{\Phi}_{t-1}\mathbf{x}_{t-1} + \mathbf{u}_{t-1} + \mathbf{w}_{t-1} \quad (10)$$

$$\mathbf{z}_t = \mathbf{H}_t\mathbf{x}_t + \mathbf{v}_t \quad (11)$$

that describe evolution of the state ($\mathbf{x}$) with time t and relate the state variables to observations ($\mathbf{z}$), where Φ_t is the evolution matrix, $\mathbf{u}_t$ represents deterministic forcing and $\mathbf{w}_t$ represents stochastic (random) forcing with zero mean and covariance matrix $\mathbf{Q}$. $\mathbf{H}_t$ is the observational operator at time t, and $\mathbf{v}_t$ is measurement error with zero mean and covariance $\mathbf{R}$. The state vector $\mathbf{x}$ should contain any model variables required to describe evolution of the dynamical model, such as carbon pools, carbon fluxes and soil water content in a terrestrial carbon model, and should include all target variables (including parameters) and any other quantities for which updated estimates are required.

The Kalman filter performs estimation at each time in two steps. The first step is to project forward from time $t-1$ to time t using the dynamical model (Eq. (10)) to estimate state $\mathbf{x}_t$ and its covariance, $\mathbf{P}_t$:

$$\tilde{\mathbf{x}}_t = \mathbf{\Phi}_{t-1}\hat{\mathbf{x}}_{t-1} + \mathbf{u}_{t-1} \quad (12)$$

$$\tilde{\mathbf{P}}_t = \mathbf{\Phi}_{t-1}\hat{\mathbf{P}}_{t-1}\mathbf{\Phi}_{t-1}^{\mathrm{T}} + \mathbf{Q}_{t-1} \quad (13)$$

where $\tilde{\mathbf{x}}_t$, with covariance $\tilde{\mathbf{P}}_t$, is the first estimate of the state at time t. In the second step, measurements $\mathbf{z}_t$ are used to update these estimates:

$$\hat{\mathbf{x}}_t = \tilde{\mathbf{x}}_t + \mathbf{L}_t(\mathbf{z}_t - \mathbf{H}_t\tilde{\mathbf{x}}_t) \quad (14)$$

and

$$\hat{\mathbf{P}}_t = (\mathbf{I} - \mathbf{L}_t\mathbf{H}_t)\tilde{\mathbf{P}}_t(\mathbf{I} - \mathbf{L}_t\mathbf{H}_t)^T + \mathbf{L}_t\mathbf{R}_t\mathbf{L}_t^T \quad (15)$$

where $\mathbf{L}_t$ is the Kalman gain matrix. See Gelb (1974) for further details.

In the case of a model being nonlinear, Φ_{t-1} is a function of model state, $\mathbf{x}_{t-1}$, and the Extended Kalman filter (Gelb, 1974) or Ensemble Kalman filter (Evensen, 2003, 2007) can be used.

The Kalman filter (and variants like the Extended and Ensemble Kalman filters) uses only observations up to the current time step to improve the estimates of the state $\mathbf{x}$ at that time. A further improvement can be provided by the Kalman smoother, which involves a second pass through the data, often implemented backwards in time, so that the state estimates at all times depend on observations at all times (Gelb, 1974). In CO_2 flux inversions, as in other trace gas inversions, the influence of emissions at one time is usually seen at the observation locations at future times. Thus the Kalman filter, using observations only up to the current time to estimate current emissions, will not use key information to estimate emissions. Thus sequential estimation of CO_2 fluxes requires either some form of Kalman smoothing or sliding windows where fluxes from previous times are included in the state vector at the current time (e.g. Bruhwiler et al., 2005; Peters et al., 2005).

Eq. (14) and Eq. (6) are similar. Eq. (6) updates the estimate of $\mathbf{x}$ using the difference in model predictions between using the optimal estimate of $\mathbf{x}$ and using the prior estimates of $\mathbf{x}$, whereas Eq. (14) updates the model state using the mismatch $\mathbf{z}_t - \mathbf{H}_t\tilde{\mathbf{x}}_t$. The Kalman filter can also be considered as a form of Bayesian estimation; Eq. (10) is a prior model describing what we know about model state at time t before any measurements are taken. In addition, at each time t, $\tilde{\mathbf{x}}_t$ and $\tilde{\mathbf{P}}_t$ characterize the knowledge of that state (Eqs. (12) and (13)), prior to a Bayesian update, using $\mathbf{z}_t$ to obtain $\hat{\mathbf{x}}_t$ and $\hat{\mathbf{P}}_t$.

In theory, in the case of a linear model and without considering model error explicitly, the solution of least squares batch methods and the sequential Kalman filter are the same (Enting, 2000). However, in reality most applications will require simplifications that will make solutions differ.

The Kalman filter has most commonly been used for state estimation with model parameters being fixed, although more recently there has been a surge of interest in using it for both state and parameter estimation (see Trudinger et al., 2008 and references therein).

The batch and sequential optimization methods have some different advantages and disadvantages that can make one or other more suitable to a particular application. An advantage of batch

methods is that they are not restricted to the M-estimator form of the cost function. In addition, considering a single cost function allows efficient application of adjoint techniques (see Rayner et al., 2005). Batch methods might be the better choice when target variables are best constrained with observations at all times. Sequential methods are particularly well suited to the incorporation of model error separate from observation error. Sequential methods might be chosen instead of batch methods to reduce the computational size of a problem (e.g. Bruhwiler et al., 2005) or because the problem naturally lends itself to sequential estimation (e.g. with target variables at each time depending mainly on observations at current or past times).

2.3. Interpretation of optimization results

The following are a few important issues that can arise in optimization calculations, mostly described in terms of parameter estimation but also relating to other target variables such as carbon fluxes in the top-down approach.

2.3.1. Geometric interpretation

For a linear model with Gaussian and independent errors, we can interpret the optimal estimates of target variables and their error covariance geometrically (see Fig. 1). Consider the case of two model parameters being estimated, the geometric interpretation of optimal estimates is as follows: a 95% confidence region for the estimates of parameters p_1 and p_2 is shown as an ellipse with two major axes, where the projected length of the two axes are 2 times the 95% confidence of estimates of p_1 and p_2. The size of the ellipse will increase with data errors or a decrease in the prior parameter confidence intervals. In the case of nonlinear models, this interpretation of the error covariance matrix is valid only in the tangent region of optimal estimates of model parameters, and may not be applicable for a lower confidence interval, as the ellipse as shown in Fig. 1 may become asymmetric or even distorted (see Wang et al., 2001 for some examples).

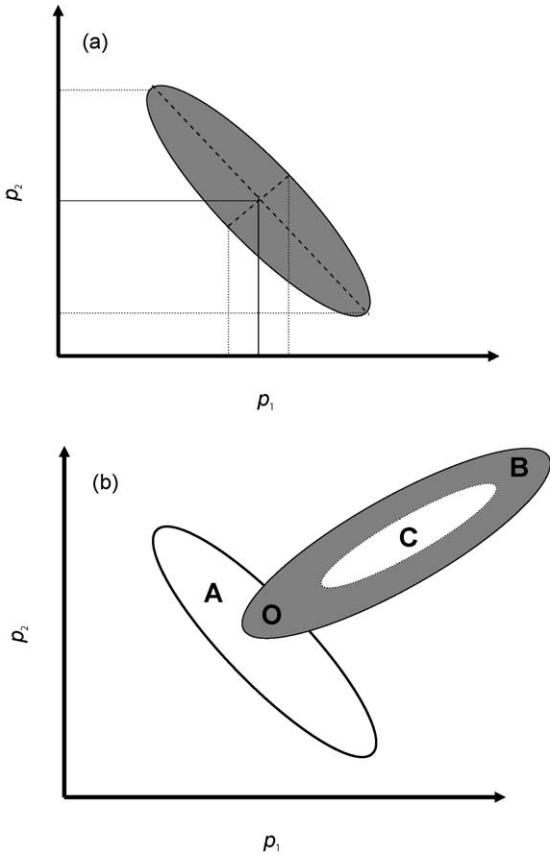

Fig. 1. (a) Geometric interpretation of the estimates and uncertainties of two negatively correlated model parameters, p_1 and p_2. The centre of the ellipse is the optimal estimates of p_1 and p_2, and the projected lengths of the two major axes onto x-axis and y-axis represents 2 times of 95% confidence intervals for the two parameters. The negative slope of the longer axis suggests a negative correlation between p_1 and p_2. (b) Ellipses A and B represent the estimates and uncertainties of p_1 and p_2 when they are estimated using two datasets (A and B). The overlapped region of ellipses A and B, region O represents the estimates of p_1 and p_2 consistent with both dataset. However, if the error of dataset B is smaller, ellipse C represents the estimate and uncertainties of two parameters when they are estimates using dataset B with smaller error only. There will be no overlap between ellipses A and C, therefore no estimates of p_1 and p_2 can be found to be consistent with both datasets.

When multiple datasets are used to estimate model parameters separately, it is very likely that optimal estimates of model parameters can be different for different datasets (see Fig. 1b) (also see Wang et al., 2001 for example). The overlapping region (region O in Fig. 1b) represents the optimal estimate of model parameters that can explain the information in both datasets consistently. If we feel more confident about one of two datasets, we may not find such an overlapping region for a given confidence level (in the case when ellipse B becomes C). That can be important for using multiple constraints in model–data fusion. In other words, a set of values of two parameters cannot be found to explain a given variance in both datasets consistent with the ascribed uncertainties. This suggests that some parts of the model may have significant errors, or that the error estimates of one or both of the two datasets are too optimistic. Weights given to different datasets can be quite subjective. If uncertainties within a class of observations can be regarded as identical, these uncertainties can be estimated as parameters in model–data fusion (see Michalak et al., 2005).

2.3.2. Equifinality and multiple minima

Equifinality or exchangeability (Aalderlink and Jovin, 1997; Franks et al., 1997) occurs when two or more model parameters have a similar effect on model outputs, so can be difficult to distinguish. This implies that quite different combinations of model parameters can give a similar match of model outputs to observations. When equifinality occurs, additional information (such as prior estimates for the parameters to keep them within physical bounds, or other kinds of observations, such as carbon stores) may help to distinguish the parameters. Alternatively, it might be possible to fix some parameters, or recast the model with fewer parameters. It is important to take equifinality into account when considering uncertainty analysis (Tang and Zhuang, 2008). While equifinality results in significant parameter uncertainty, it may or may not lead to significant prediction uncertainty. Broadly speaking, if the model is applied for conditions much like those used for calibration, as long as the correlations between parameters are taken into account the model is likely to do a good job at prediction. However, if the model is applied for conditions unlike those used for calibration, such as predicting into the future under global warming, there could be significant uncertainty if parameters describing different processes cannot be distinguished and the sensitivities to those parameters change into the future.

Another problem (that can be related to equifinality) is that the cost function may have multiple minima. This will pose a significant problem for gradient-based methods, as they can get trapped in local minima and not find the global minimum or find the minimum closest to the initial values of the estimated parameters or states. Global search methods, such Markov Chain Monte Carlo (MCMC) or genetic algorithm, will generally be more successful at avoiding these problems, and therefore can in principle be used to optimize many more parameters (although some of those parameters might be poorly resolved due to equifinality).

The Bayesian approach is often used to ensure that meaningful estimates of model parameters are obtained in the case of equifinality, or that optimization is numerically more stable, or both. Eq. (5) augments Eq. (4) by including the additional terms for our prior knowledge of the model parameters or state or both. Some caution needs to be exercised in assigning suitable weights to the observations and priors. If priors are used, the optimal estimates of parameter uncertainties will be equal to or less than the prior estimate if errors are Gaussian. The ratio of optimal and prior estimates of model parameter uncertainty is a measure of the amount of information in the data about that parameter, the lower the ratio, the more information data have on this parameter. A discussion of information content in carbon cycle modeling is given by White et al. (2005) and Enting (2008). As a precaution, it is useful to give lower weights or larger uncertainties to the priors so that the search for the optimal solution for Eq. (5) is not too constrained.

Another issue, which is somewhat related to the issue of weighting of priors, is how many model parameters should we vary? The two questions are somewhat related. In principal, we should vary all model parameters that have some degree of sensitivity to measurements. Attempting to vary too many parameters is likely to result in singularity of $\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}$ in Eqs. (6) and (7) (see Acton, 1970), due to insufficient information contained in the observations to uniquely constrain the parameters, and optimization will fail. One solution is to 'regularise' the estimation by taking a singular value decomposition of $\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}$ and, when inverting, replace the inverse of small singular values by zero (Press et al., 2001), since these singular values reflect combinations of parameters for which the data provide little, if any, information. Alternatively as noted above, by replacing $\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}$ by $\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H} + \mathbf{W}^{-1}$, Bayesian cases avoid such singularities. The prior estimates provide the information about parameters that is not available from the observations. Some model parameters are correlated, so fixing one parameter will improve the estimate of another parameter. In practice, we should vary the most sensitive parameters and use tighter constraints on those less sensitive parameters in the prior estimates. The number of model parameters that can be reliably estimated can be determined by trial and error, but often only a small number of model parameters in global land surface models can be well resolved using eddy covariance flux data alone (see Wang et al., 2001). Sensitivity analysis can also be used to estimate the number of identifiable parameters in a model (see Saltelli et al., 2004).

2.3.3. Resolution and variance

A final point is the tradeoff between model resolution and variance (see Menke, 1989; Enting, 2002). Suppose that we try to estimate the spatial distribution of global carbon fluxes using the surface CO_2 concentration measurements, we can discretize the earth's surface into K regions, and apply Bayesian synthesis to estimate the net flux over each region. As suitably calibrated time series of surface CO_2 concentrations are only available at a limited number of sites globally (currently <100), we may not be able to obtain independent estimates of fluxes over some regions where limited surface CO_2 concentration observations are available. If too fine discretization is used in the model, errors in the estimates of some regional fluxes will become quite large and their uncertainties will be highly correlated, as there is little information in the data to provide independent estimates of fluxes over those regions. If the discretization used is too coarse, the error of the flux estimate will be small, but we lose some information in the data about the finer structure of spatial variation of net C fluxes. Therefore, there is a trade off between model resolution and variance. Enting (2008) has used the example of digital filtering to illustrate the different aspects of model and data resolution and the implications for estimation.

2.4. Uncertainty analysis

Analysis of uncertainties is an integral part of model–data fusion, for two important reasons. Firstly, as neither the measurements nor the model are perfect, we seek the optimal fit of model predictions to measurements while taking account of the errors in both. Secondly, an estimate of a quantity, such as a carbon flux or a model parameter, is of much greater value if an estimate of its uncertainty is provided. Ideally a statistical model underlying a model–data fusion calculation needs to incorporate the uncertainties in measurements, initial conditions, prior parameter values, model inputs, as well as represent model error. These uncertainties must then be propagated through the calculations to the various output quantities including the target variables.

When the model is linear and errors are Gaussian and independent, the covariance of the estimates of the target variables, $\mathbf{x}$, is given by Eq. (7) as a function of errors in the data ($\mathbf{R}$) and in prior parameter estimates ($\mathbf{W}$) for the batch method, or by Eq. (15) for state estimation using the Kalman filter. When the model is nonlinear, or errors are non-Gaussian, the relationship between the covariance of the target variables and errors in the model and in the data may be analytically intractable. One approach is to use Monte Carlo techniques, whose output will be a sample from the estimated distribution of target variables.

Specifying the statistical model, while often critical to the success of a calculation, is by no means trivial. There are often physical reasons for particular choices of input uncertainties, such as the observation error, $\mathbf{R}$, being based on measurement and/or representativeness error. Enting (2002) discussed aspects of error modeling relevant to synthesis inversion. Raupach et al. (2005) identified many of the characteristics of data that will influence the form of the error distribution. If multiple independent data are believed to have similar characteristics, statistical estimation techniques can be used to estimate the observational uncertainties in a Bayesian synthesis inversion (see Michalak et al. (2005) for example).

Model error, which can be due to missing or mis-specified processes, incorrect values of model parameters and numerical representation of physical equations, is often the most difficult to quantify. It is important to recognize that model and data errors are likely to have very different covariance structures. For example, repeated instances of observations may have independent errors, while errors in repeated or related model calculations will usually be highly correlated. A critical point is that there needs to be consistency within the modeling process, such that all variability in the observations is modeled as either signal or noise (Enting, 2002).

There are ways to diagnose consistency in the statistical model. For normally distributed data, the chi-squared statistic is often used. This compares the sample covariance of the innovations $\boldsymbol{\nu} = \mathbf{z} - \mathbf{Hx}$ with the innovation covariance matrix $\boldsymbol{\Gamma} = \mathbf{HPH}^{\mathrm{T}} + \mathbf{R}$ (given here using the Kalman filter notation). The innovation covariance will vary for different choices of the observation error covariance $\mathbf{R}$ and the covariance of stochastic forcing $\mathbf{Q}$. Chi-squared is calculated as $\chi^2 = \boldsymbol{\nu}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\nu}$, and is compared to the number of observations to see whether the variation in the innovations is more (indicating a problem) or less (indicating a conservative approach) than is accommodated by the state and observation errors $\mathbf{P}$ and $\mathbf{R}$. Choices of different ratios for model and observation errors alter how much of the variability in the observations is treated as signal, and how much is treated as noise. Given that the total variance in the measurements is fixed, the more variance is assigned to data error by increasing $\mathbf{R}$, the less variance is required by model error, or vice versa. Ménard and Chang (2000) looked at ways to interpret chi-squared from a

Kalman filter, interpreting trends, offsets and initial values in chi-squared in terms of model, observation and prior errors, respectively. Uncertainties will refer to different things for different choices of model and observation errors (Trudinger et al., 2002, 2008).

When a cost function is constructed using Eq. (1) or (5), we assumed that all errors are Gaussian. If systematic variation exists in the model residual, a statistical model must be constructed to account for the non-Gaussian component, and used in the model–data fusion for estimating parameters or model states iteratively until all variance in the measurements are accounted for. The variance explained by the statistical model also contributes to the model error, and should be used to guide future model improvement (see Abramowitz et al., 2008).

Specifying uncertainties often involves quantifying and understanding correlations between estimated quantities. Measurements can have correlated errors (Wang and Barrett, 2003; Richardson et al., 2008), but many studies assume uncorrelated errors for simplicity. Estimated parameters will often be correlated. In this case both uncertainties and their correlations must be taken into account when models are used to make predictions. For example, considering the case of estimating two parameters that are strongly correlated, the combination of the two parameters might be well constrained, but the individual values not well constrained at all (see Fig. 1). In the extreme case (see Section 2.3.2), some combinations of parameters cannot be distinguished. In linear cases, the various issues of correlated target variables can be identified by standard techniques of matrix algebra. However, characterization of correlated estimates is significantly harder to quantify for non-linear problems.

3. Applications of model–data fusion techniques

Model–data fusion methods have been widely used in studies of carbon exchange between terrestrial biosphere and atmosphere. Broadly speaking, they can be divided into flux estimation, parameter estimation, analysis of model errors, experimental design and forecasting.

3.1. Flux estimation

Early work on estimating carbon fluxes in space and time using the top-down approach was based on ad hoc synthesis (Keeling et al., 1989; Tans et al., 1990). Essentially, the synthesis technique involves solving Eq. (2) for a small number of large-scale flux distributions (in space and/or time) with strengths x_k, where H_{nk} are determined from K calculations with an atmospheric transport model run with one source distribution at a time. The Bayesian synthesis inversion (Enting et al., 1995) extended the synthesis approach by (a) including more source components, (b) including prior estimates to ensure stability of the calculation in case $\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}$ being singular, (c) allowing for uncertainty in estimated fluxes by minimizing Eq. (4) rather than seeking to solve Eq. (2) exactly.

Within the general Bayesian framework, many different calculations have been undertaken. Increasing availability of $\delta^{13}C$ data provided additional constraints on partitioning of fluxes between land and ocean. Ciais et al. (1995) used latitudinal gradients of ^{13}C data in two-dimensional inversions, while Enting et al. (1995) used the long-term trend in $\delta^{13}C$ as an additional constraint. Later studies were able to make additional use of the increasing amount of ^{13}C data, and the subsequent availability of measurements of trends on the O_2:N_2 ratios (Rayner et al., 1999). One disappointing result from these studies was that inversions using surface CO_2 concentration measurements only provided little discrimination between land and ocean fluxes, with most of the discrimination obtained by including $^{13}CO_2$ (see Enting, 2002).

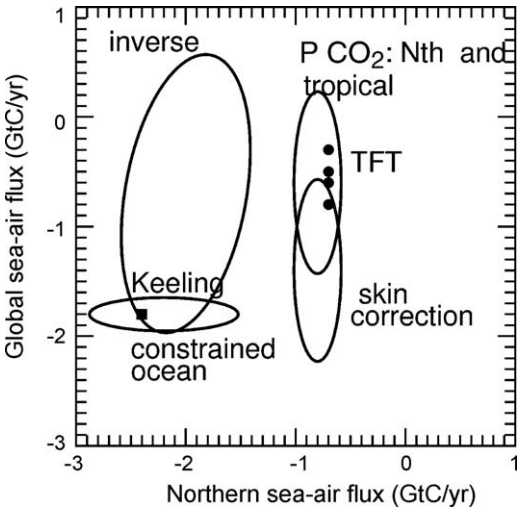

Fig. 2. shows the estimate of carbon uptake by the ocean in Northern Hemisphere against the estimate of carbon uptake by all oceans from inversions. The largest ellipse 'inverse' represents the estimates of two fluxes using measurements of surface CO_2 concentration only, the ellipse "constrained ocean" represents the estimates using the results from ocean box modeling. The square dot represents the estimates using both CO_2 concentration data and results from ocean box modeling as done by Keeling et al. (1989). The ellipses "TFT" and 'skin correction' represent the estimates of two fluxes using surface ocean water P_{CO_2} data with and without skin temperature correction, respectively. Results from Tans et al. (1990) who applied inverted both surface CO_2 concentration data and surface water P_{CO_2} without skin temperature correction are shown as black dots (adapted from Enting, 2002).

This was somewhat surprising given the large north–south asymmetry in the land–ocean distribution.

The evolving understanding of the carbon cycle provided the context within which the various developments took place. Enting (2002) reviewed some key steps, noting discrepancies in flux estimates and controversies regarding a strong terrestrial sink in North America. The discrepancies in estimates of the amount of oceanic CO_2 uptake were revealed by the first two three-dimensional synthesis inversions (Keeling et al., 1989; Tans et al., 1990) which differed by over 1 GtC/year. The respective high and low estimates were each supported by a global-scale analysis of ^{13}C distributions (Quay et al., 1992; Tans et al., 1993).

Fig. 2 shows re-analysis of the early synthesis inversion calculations from Enting (2002). Results are presented for aggregated ocean fluxes for the globe and for the northern hemisphere, but with the calculations performed with over 20 fluxes ($K > 20$) to reduce the discretization errors noted in Section 2. The large ellipse, 'inverse', represents the estimates of two fluxes from an inversion using measurements of surface CO_2 concentration only. If, as in the work of Keeling et al. (1989), a constraint on the total oceanic uptake is applied using results from ocean box modeling, the result is shown as the ellipse 'constrained ocean'. When two constraints are applied, the result as obtained by Keeling et al. (1989) is shown as the solid square symbol in Fig. 2. If the flux estimates from inversion of surface CO_2 concentration are constrained by estimates of tropical and northern fluxes using surface ocean water P_{CO_2} data, the result is shown as the 'P_{CO_2}' ellipse in Fig. 2. The solid points (TFT) give the results from Tans et al. (1990). Since these results use priors from chemically based air–sea flux measurements it was proposed that a correction be applied to reflect the difference between surface water temperatures as measured and the actual skin water temperature. This correction results in the estimates as indicated by the 'skin correction' ellipse in Fig. 2.

These discrepancies were ultimately resolved by Sarmiento and Sundquist (1992) by including the carbon export to ocean by rivers

and rainfall, differences between skin temperature and surface water temperature and the role of carbon monoxide transport in biasing inversion estimates. A consistent global isotopic budget reflecting these considerations was presented by Heimann and Maier-Reimer (1995).

The major contribution of the Bayesian synthesis approach was the capability of calculating uncertainties, by propagating uncertainty of the data and priors through the calculations. What was lacking was the capability for assessing the contributions from errors in the transport models. This problem was addressed through an 'intercomparison' (i.e. running multiple models with standardized calculations) known as TransCom. After initial comparisons of calculated responses for the prescribed CO_2 emissions and a study of transport of SF_6, TransCom produced a set of standardized synthesis inversions (Gurney et al., 2002, 2003). The main conclusion was that, in terms of the impact on estimated fluxes, transport model uncertainty was as important as data uncertainty. A major distinction in the results reflected the way in which models represented transport through the boundary layer. Most of the inversions did not resolve tropical fluxes well. This represented a combination of low density of observations and a poor signal-to-noise ratio since large-scale circulation moves air away from the tropical surface.

The availability of additional data and greater computing power allowed for more detailed calculations. In parallel with these developments, various methodological aspects were explored including: regularization that is used to fix or constrain more tightly some poorly constrained model parameters or states (Baker, 2000), spatial smoothing that is used to prescribe some spatial patterns of the target variables (Rödenbeck et al., 2003), estimating data uncertainties (Michalak et al., 2005) and the use of 'shrinkage estimators' that reduce mean-square error at the expense of a bias (Shaby and Field, 2006).

In parallel with methodological refinements, new computational techniques were explored. The most important of these were the adjoint techniques. These exploited the capability of efficient calculation of $\partial\Theta/\partial x_k$ without the need to evaluate each of the H_{nk} in (2), thus allowing the minimization of (3) by gradient techniques. Indeed, not only was calculation of the H_{nk} unnecessary, but even the existence of the linear relation (2) was not required. Adjoint techniques could be equally well applied to nonlinear dependence $m_n(\mathbf{x})$, so could be applied to the more general estimation based on minimizing the more general form (3).

Overall application of model–data fusion methods has improved the estimates of regional and global carbon fluxes over the last two decades. Robust estimates of decadal global carbon budgets were presented by Schimel et al. (2001) for the 1980s and 1990s and mean annual budget for 1992–1996 by Gurney et al. (2002). One of the most exciting recent advances is the development and implementation of an operational tool, CARBONTRACKER, using the Kalman filter to estimate surface fluxes and their uncertainties using surface concentration measurements in nearly real time (Peters et al., 2005; Peters et al., 2007).

3.2. Parameter estimation

Use of model–data fusion for parameter estimation is one of the most common applications. With the rapid expansion of eddy flux towers in the field over the last two decades, continuous measurements of surface fluxes have been made on a wide range of land surface types covering most parts of climate space experienced by terrestrial plants under the present climate conditions (Williams et al., 2009). Both gradient-based (e.g. Wang et al., 2001; Luo et al., 2003) and non gradient-based methods (e.g. Mo and Beven, 2004; Braswell et al., 2005; White et al., 2005) have been used. Model–data fusion methods were also used to diagnose the causes of variations of model parameters resulting from prolonged water stress (Reichstein et al., 2003; Zobitz et al., 2008) or leaf phenology (Sacks et al., 2007; Wang et al., 2007; Prihodko et al., 2008).

Examples of estimated model parameters include maximum leaf carboxylation rate (Wang et al., 2001, 2007), light use efficiency (Carvalhais et al., 2008), water use efficiency (Yuan et al., 2007) and soil carbon transfer coefficient (White and Luo, 2002) and so on. The number of parameters that are optimized using multi-year eddy flux data varies from 5 (see Wang et al., 2001) to 25 (Braswell et al., 2005). As more parameters are estimated, it becomes more likely that equifinality will occur in the optimization.

Whether some model parameters can be well constrained by the data depends on the amount and quality of information in the measurements and how that information is represented in the model, as quantified by the sensitivity of the model equivalents of the measurements ($\mathbf{Hx}$) to the optimized model parameters. Spectral analysis of eddy covariance data show that much of the information in the measurements is from daily to seasonal time scales (Katul et al., 2001). Diurnal variation of measured surface fluxes are strongly correlated with incident photosynthetically active radiation (PAR) and water vapor pressure deficit (Stoy et al., 2005), and therefore contains little information about model parameters that vary seasonally or inter-annually, such as leaf area index. At inter-annual time scales, the net carbon exchange between the land surface and the atmosphere is strongly influenced by site disturbance history (Schimel et al., 2000; Saleske et al., 2003), meteorological forcings (Schimel et al., 2005), and initial pool sizes (Carvalhais et al., 2008). When data on site disturbance history are not available or not used in the model, we may have to vary initial carbon pool sizes to explain the observed inter-annual variations of net ecosystem exchange of carbon (NEE).

The above results from spectral analysis of eddy flux measurements are also confirmed by model–data fusion studies. Santaren et al. (2007) found that model parameters related to photosynthesis and energy partitioning are well resolved by eddy flux measurements, whereas model parameters related to respiration are poorly resolved, as the estimates of respiration parameters can vary significantly with the estimates of initial carbon pool sizes. Another study by Carvalhais et al. (2008) showed that optimizing the initial carbon pool sizes as well as model parameters significantly improved the model fit to the measurements. Respiration parameters have often been found to be poorly constrained by NEE measurements alone. Complementary measurements, such as soil respiration and carbon pool sizes will provide useful constraints on those parameters (Williams et al., 2005).

Two recent studies compared a number of parameter estimation methods applied to common datasets and models (Trudinger et al., 2007; Fox et al., 2009). In the OptIC project (Trudinger et al., 2007), both batch methods and sequential methods were applied to synthetic data with different types of errors. The synthetic data were generated using a simple model that consists of two coupled nonlinear differential equations, where the model has properties similar to a terrestrial biosphere model. The largest variation in parameter estimates by different groups was due to the choice of cost function (how the minimum was defined), not the optimization method used (how the minimum was located). The REFLEX project (Fox et al., 2009) used both synthetic and real CO_2 flux and leaf area index data with the DALEC model (Williams et al., 2005). Model parameters and states were estimated using batch or sequential methods. Parameters linked to gross primary production (GPP) and ecosystem respiration were best constrained, while parameters describing slow processes (turnover of wood and allocation to roots) were poorly estimated, suggesting that

additional data on slow pools would be beneficial. The methods differed in their estimates of confidence intervals on parameter and state estimates. In both the OptIC and REFLEX projects, choices in the implementation of optimization methods, such as data weights, priors, initialization, and method-specific choices such as accept/reject criterion in Metropolis methods, were more important than the choice of optimization method itself. The results highlight the importance (and difficulty) of correctly representing uncertainties in both model and measurements in parameter estimation.

In the top-down approach, surface fluxes are often estimated. Only a few studies have estimated model parameters for different biomes using surface CO_2 concentration and remote sensing measurements (Wang and McGregor, 2003; Rayner et al., 2005). Rayner et al. (2005) developed the first carbon cycle data assimilation scheme (CCDAS), in which they estimated 58 model parameters using monthly surface CO_2 concentrations from about 40 monitoring stations from 1979 to 1999. They found that 500 CO_2 concentration measurements per year globally provided little information about model parameters related to leaf photosynthesis or soil carbon turnover rate. The best constrained model parameters were soil carbon storage efficiency. Interpretation of the results is that monthly CO_2 concentrations provide reasonable constraints on the net C exchange (i.e. the difference between net primary production and heterotrophic respiration), provided that fluxes from fire, land use change and so on are well quantified. Little information is available in the atmospheric CO_2 concentration measurements about model parameters related to net primary production and heterotrophic respiration, as the estimates of these two fluxes are often negatively correlated and poorly resolved by surface CO_2 concentration measurements only (Wang and McGregor, 2003).

Most applications of model–data fusion to parameter estimation have only presented the improved parameter estimates and model predictions, and did not further analyze whether any systematic model errors still remained and how the remaining error would influence the estimates of model parameters if removed. As found by Richardson and Hollinger (2005), the errors in the carbon fluxes as measured by eddy covariance technique were not Gaussian, and the standard deviation of the measurement uncertainty also varied seasonally, these error characteristics likely have significant impact on parameter estimates and model predictions (Trudinger et al., 2007; Tang and Zhuang, 2008).

3.3. Analysis of model error

The success of applying model–data fusion methods depends on how accurately the errors in the observations and model are described, in terms of both the magnitude and probability distributions. To some extent, errors in the observations can be quantified more easily than the errors in the model, particularly systematic model errors. Model–data fusion techniques can be used to identify the systematic model errors. To separate the systematic model errors from errors in model parameters, analysis of model errors can be carried out after the model parameters are optimized.

There are two types of model errors, random and systematic. Random errors, e.g. from coarse sampling of chaotic dynamics, are much easier to cope with, because many optimization methods assume errors are Gaussian. Systematic model errors that result from model structure probably are the most difficult to identify and quantify (Chatfield, 1995; Abramowitz et al., 2008). Systematic errors, if not treated correctly, can affect both the estimates and their uncertainties. Here we discuss applications of model–data fusion techniques in identifying and quantifying systematic model errors.

In the top-down approach, atmospheric transport models are used to calculate the **H**, which is the sensitivity of CO_2 concentration at observing sites to surface flux from a region. In a linear case, **H** is independent of **x**. When **H** is computed using atmospheric transport models, errors in **H** can significantly affect the estimates of surface fluxes (see Gurney et al., 2003, Baker et al., 2006a). Gurney et al. (2003) compared the estimates of surface fluxes from 26 regions using the modeled concentrations from 76 surface observation sites, and found that the errors in **H** can be as important as the number of measurements in estimating regional carbon fluxes.

Baker et al. (2006a) evaluated the errors in the atmospheric transport for CO_2 by comparing 13 different models, and found that errors in the atmospheric transport can have a significant impact on the partitioning of carbon fluxes between land and ocean in the northern hemisphere. A passive tracer with well known sources, sulfur hexafluoride (SF_6), was used to assess transport model errors. Denning et al. (1999) compared the simulated spatial and temporal distribution of SF_6 in the atmosphere with surface measurements, and found that the estimates of inter-hemispheric exchange time can vary from 0.8 to 2 years, with larger disagreement among the models for continental sites than for the sites near the marine boundary layer, suggesting that some models may have significant biases in the vertical transport and planetary boundary layer dynamics over land.

In the bottom-up approach, errors in the model can also be identified using model–data fusion approaches. This can be done by analyzing model residuals and comparing different models (see Meldyn et al., 2005; Stöckli et al., 2008; Williams et al., 2009). A more sophisticated method has recently been developed by Abramowitz et al. (2007) for analyzing systematic model errors in complex global land surface models. They used a neural network to predict each of three surface fluxes (latent and sensible heat and net ecosystem carbon exchange) as a function of four inputs (incoming short-wave radiation, air temperature, air humidity and wind speed) using the self-organizing linear output neural network (Hsu et al., 2002). First they divided the four inputs into 100 nodes, each node represents a particular combination of weather conditions, such as hot and dry day time, and wet and cold night time and so on, and a multiple piece-wise linear regression was fitted to the observed meteorological forcing and fluxes that belong to that node. By comparing the predictions of three major land surface models with those from the neural network, they found that all land surface models significantly underrepresented the information in the meteorological forcing about all three surface fluxes, and two of three model studies significantly underestimated the night-time ecosystem respiration under warm and humid conditions (see Abramowitz et al., 2008).

Spectral analysis of model residuals can provide quite a useful perspective on model errors (Braswell et al., 2005; Williams et al., 2009). By analyzing the model residuals using wavelet analysis, Braswell et al. (2005) found that eddy covariance flux measurements for 11 years from the Harvard Forest provided good constraints on the parameters related to the processes controlling net carbon exchange at daily to seasonal time scales, but not at inter-annual time scales at which the model's variance in NEE was significantly less than the variance in the measurements. Williams et al. (2009) used the orthonormal wavelet transformation to analyse model residuals in both the time and frequency domains (Stoy et al., 2005). They found that much of the model–data mismatch between the predictions by a land surface model, CABLE, and measurements resulted from the underestimation of variance at low frequency (seasonal to annual time scales) for an evergreen Eucalyptus forest site in south-east Australia. Both studies showed that the models poorly represented key processes influencing the variations of NEE at seasonal to annual time scales, such as canopy

leaf area index dynamics, disturbance or even the sequence of weather experienced by those forests in the past (Schimel et al., 2005).

Other model–data fusion methods can also be used, such as the generalized likelihood uncertainty estimation (Beven and Binley, 1992), Bayesian model averaging (Hoeting et al., 1999) or multimodel ensemble method (Palmer et al., 2000), but are rarely used in the studies of ecosystem carbon fluxes.

3.4. Evaluation of sampling strategies

Analysis of the sensitivity of model parameters or processes to measurements can be used to develop sampling strategies, that is, to determine what are the benefits of using different measurement types (Luo and Reynolds, 1999; Williams et al., 2005), sampling frequencies (Wang et al., 2006), sampling locations (Rayner et al., 1996) or sampling designs (Rayner and O'Brien, 2001). In linear problems, determining the extent to which new observations could reduce uncertainties is simplified by analyzing the variance as the exact expression for posterior covariance (see Eq. (7)) to **R** or **H**.

In the top-down approach, surface CO_2 concentrations are available from a limited number of sites globally. One question is whether we could improve the existing network by relocating some of the existing observation sites. Rayner et al. (1996) applied synthesis inversion to estimate the surface C fluxes from surface CO_2 concentration measurements using the existing network, and then applied a simulated annealing technique to find whether some of the existing stations could be relocated to where the sensitivities of flux estimates to the measurements, or **H**, are higher. This would maximally reduce the uncertainties of global C flux estimates.

In the top-down approach, most global inversion studies use weekly or monthly mean concentrations of background air. As a result, measurements from episodic high CO_2 concentration events resulting from dispersion of local land CO_2 sources are not used. However, measurements from those episodic high CO_2 concentration events may contain much information about regional land surface fluxes, biomass burning events and so on. For example, Law et al. (2004) estimated how well the regional C fluxes from the Australian continent could be constrained by hourly measurements of surface CO_2 concentrations. They found that measurements from the northern-west coastal region or central region of Australia would provide the best constraints on the estimates of regional C fluxes from the Australian continent.

In the bottom-up approach, model–data fusion can also be used to identify what measurements are needed to constrain certain model parameters. Xu et al. (2006) applied model–data fusion techniques to a model of ecosystem carbon dynamics and identified that commonly measured quantities of soil respiration, leaf biomass, woody biomass, litter fall and soil carbon pool size in most experiments could be used to provide effective constraints on estimates of four of the seven transfer coefficients in their model. Other measurements or different sampling strategies have to be used for constraining the remaining three transfer coefficients.

In the bottom-up approach, model–data fusion has been used to estimate parameters in a process-based model using eddy flux measurements. However, only a limited number of parameters are well constrained (Wang et al., 2001; Prihodko et al., 2008). The question is what other measurements are needed to constrain additional model parameters? Williams et al. (2005) found that dynamics of slow carbon pools, such as woody tissue and coarse woody debris were poorly constrained by flux measurements, and that direct measurements of slow carbon pools over time would provide better constraints on the simulated net carbon exchange rates.

Wang et al. (2006) applied a model–data fusion technique to address the question of what measurements are required to identify different feedback mechanisms of ecosystem responses to elevated CO_2 under nitrogen-limiting conditions. Three feedback mechanisms have been proposed to describe the influences of the soil nitrogen cycle on plant response to elevated CO_2 concentration: the litter quality feedback, litter quantity feedback and N mineralization feedback (Medlyn and McMurtrie, 2005). The first two feedbacks are negative and the third one is positive. It is uncertain which of these mechanisms will dominate in a given ecosystem. Wang et al. (2006) conducted a twin experiment (also known as an observing system simulation experiment) that has two parts: simulation and estimation. For the simulation, they used a process-based model of the carbon and nitrogen cycles, G'DAY (Comins and McMurtrie, 1993) to simulate the responses to elevated CO_2 for each of three mechanisms. Different mechanisms were simulated by choosing different values for four model parameters (see Fig. 3), and random noise was then added to the model output to mimic measurement uncertainty. For example, the litter quality negative feedback was simulated by reducing the fraction of leaf nitrogen translocated during senescence from 0.3 for the present CO_2 level to 0.1 for doubling present CO_2 level. For the estimation, they sampled different model outputs with random noise at different frequency to represent different "measurement" types, and a batch method was applied to the "measurements" to estimate those four model parameters and their covariance. When comparing the estimates of four model parameters with what were used in the simulation, they found that no single type of measurement could provide enough information to constrain all four model parameters, all three types of measurements were required over 5 years. On the other hand, the estimates of two

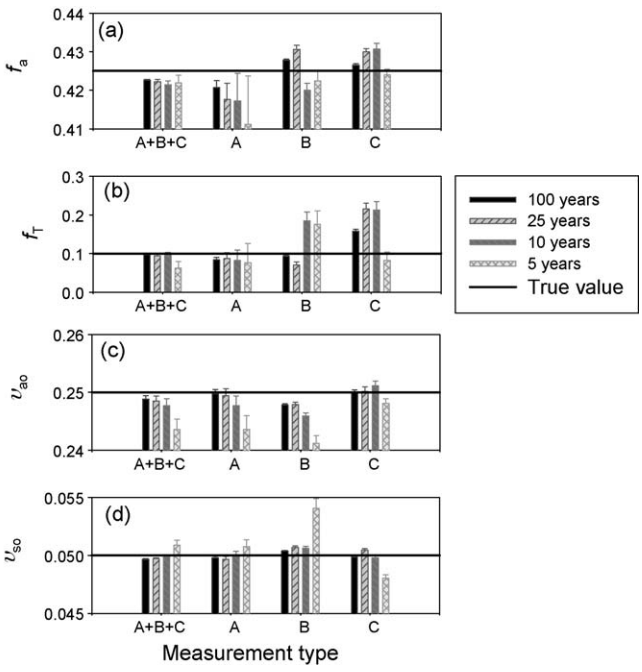

Fig. 3. Optimal estimates of all four parameters when 5, 10, 25 or 100 years' "measurements" (A, B, C or A + B + C) were used in the estimation. The error bar represents one standard error of the estimate. The horizontal line on each plot represents the "true" values of the parameters used in the simulations. "Measurements" are: monthly net N mineralization (A), yearly carbon and nitrogen pool sizes of foliage (B); yearly carbon and nitrogen pool sizes of fine roots and active soil organic matter (C). The four parameters are: $f_a = \eta_f/(\eta_f + \eta_r)$, where η_f and η_r are fractions of carbon allocated to leaf and fine root respectively, f_T: the fraction of nitrogen retranslocated from senescent foliage, υ_{ao} and υ_{so} are the maximal N:C ratio of the newly formed active and slow soil pools (adapted from Wang et al., 2006).

parameters (υ_{ao} and υ_{so}) remain highly negatively correlated even when all three measurements of 100 years were used in the estimation (see Fig. 3). Therefore, additional measurements are required to provide independent estimates of those two parameters.

3.5. Forecasting

Model–data fusion techniques can be used to systematically assess the uncertainties in model predictions. Perhaps the most common ones are Kalman-filter techniques and adjoint techniques that are often used in numerical weather forecasting but rarely used in studies of carbon fluxes. One exception is the work by Scholze et al. (2007). They used the observation data of surface CO_2 concentration from 1979 to 1999 to calibrate some model parameters of a global terrestrial ecosystem model and then used the calibrated model parameters and their uncertainties to hindcast net fluxes of CO_2 and their uncertainties between land surface and atmosphere and from 2000 to 2003.

The REFLEX model–data fusion project (Fox et al., 2009) involved forecasting carbon dynamics with confidence intervals, after model parameters had been estimated. In general, confidence intervals were larger for forecast periods than for periods with calibration data. A similar fraction of the 90% confidence intervals on flux estimates encompassed the truth in the calibration and forecast periods, suggesting that algorithms were able to make reasonable quantification of error propagation in forecasts. However, confidence intervals differed significantly between different implementations of the same methods, suggesting that further work was needed in the area of estimating (and verifying) uncertainties.

As discussed earlier, equifinality can occur when a complex model is calibrated. Forecasts may provide a useful way of separating multiple sets of model parameter estimates that give similar fits to the data during calibration, but quite different forecasts.

4. Future directions

Some of the past progress can be attributed to the rapid growth of available measurements. With the continuing growth of measurements and increasing model complexity in the studies of terrestrial carbon fluxes, model–data fusion methods will become essential for improving our understanding of CO_2 exchange and the feedback between the land biosphere and the atmosphere at all scales. Here we discuss a number of future directions.

4.1. An integrated framework for model–data fusion

As we reviewed in the previous section, model–data fusion methods can be applied to estimate model parameters and quantify model errors. When model parameters are estimated, we implicitly assume that model errors have been correctly accounted for. On the other hand, when model errors are assessed, we assume that all model parameters are reasonably correct. Neither is the case, as shown by Abramowitz et al. (2008) for a complex land surface model, some systematic model errors can be quite large and can not be accounted for by the variations of model parameters within their realistic ranges.

Model–data fusion techniques can be used to identify and quantify systematic errors in the model structure, as we reviewed in the previous section. For simple models, such as a model of soil respiration as an exponential function of soil temperature, measurements of soil carbon fluxes will be adequate. For complex models such as land surface models with significantly more parameters, we need multiple sets of observations from multiple sites for quantifying possible errors in different parts of the models (also see Williams et al., 2009). However, the errors in the data will be different for measurements at different time and spatial scales, and may be correlated in space and time (e.g. Wang and Barrett, 2003). For example, measurements of soil carbon pools are usually made once every 1–10 years, and measurements of C fluxes are made half-hourly using eddy covariance techniques, and their error distributions can be quite different. Therefore, we need to treat the errors and error correlation of different datasets consistently and systematically.

A number of international efforts have been made to compare different process-based models (e.g. Cramer et al., 1999), but the results were not analyzed using model–data fusion techniques. Therefore, the causes for the major discrepancies in the predictions by different models could not be identified and quantified systematically to guide future model development.

There is an urgent need for developing a consistent and integrated framework to account for the probability density functions of the errors in both model and measurements, this is particularly important when multiple datasets of different spatial and temporal resolution are used in the optimization. Recently Liu and Gupta (2007) advocated an integrated system framework for applying model–data fusion in hydrological modeling. Their framework explicitly includes uncertainties in the data for model calibration, model input, model output, model structural errors and model predictions in a systematic manner using the Bayesian framework. Their framework consists of four steps: system identification, parameter estimation, state estimation and model predictions. The most difficult step perhaps is system identification that compares the performance of different models using multiple datasets and selects better models for the subsequent steps. After calibrating the model parameters and states using measurements, the calibrated model is then used to make predictions with uncertainties, and the predictions are then compared with additional measurements. The four steps can be carried out sequentially and the whole process can be done iteratively as new measurements become available. The information learnt from the previous step is propagated into the next step using Bayes theorem.

4.2. Monte Carlo simulations and non-Gaussian statistics

Tarantola (2005) reviewed the progress in inversion studies over the two decades since his earlier book (Tarantola, 1987). He noted the increasing scope for using Monte Carlo techniques, made possible by increasing computing power.

In particular, he argued that presenting a set of results from a Monte Carlo sample could often be more informative than presenting a mean. This echoed his earlier assertion (Tarantola, 1987) that for any inverse problem, the solution is the posterior distribution from Bayes relation. As we discussed in the previous sections, the ability to propagate errors from data, model input, model structure and model predictions systematically and consistently is important for the applications of model–data fusion in the future, and an important advantage that Monte Carlo techniques bring to terrestrial carbon modeling is the ability to handle non-Gaussian distributions. Of course, exploiting this capability requires identification of appropriate distributions to make use of the advantages of Monte Carlo techniques. Furthermore, Markov Chain Monte Carlo approaches work with relative probabilities and so can be applied to Bayesian estimation. There remains considerable scope for deploying these techniques for model–data fusion studies of the terrestrial carbon system in the future.

4.3. Applying model–data fusion using satellite measurements of atmospheric CO_2 concentrations

Measurements of atmospheric CO_2 using satellites will offer unprecedented spatial and temporal resolutions, allowing greater insight into how terrestrial ecosystems and the ocean biosphere are functioning. Together with other satellite-based measurements of leaf phenology and absorbed photosynthetically active radiation and data from the expanding ecological networks throughout the world, we will be able to monitor C fluxes at regional and global scales and their response to short-term climate variation, such as drought and future global warming. Model–data fusion was successfully used to quantify the response of forest ecosystems to a recent heat wave in Europe (Ciais et al., 2005), to warming in the northern high latitude (Piao et al., 2008), and is expected to become an important diagnostic and prognostic tool for earth monitoring in the near future. As a result, we will develop better models at regional and global scale to identify and quantify key biophysical and biogeochemical processes that influence the feedbacks between land biosphere and climate in the present and future.

5. Conclusion

We reviewed the applications of the model–data fusion in two approaches: top-down and bottom-up in the studies of carbon exchange between terrestrial ecosystems and atmosphere. Top-down approaches that use measurements of global CO_2 concentration and sometimes other atmospheric constituents to infer carbon fluxes from the land surface, and bottom-up approaches that estimate model parameters using the process-based models. Significant progress has been made in estimating net carbon fluxes between terrestrial biosphere and atmosphere using top-down approaches over the last three decades. Robust estimates of regional and global carbon fluxes over the last two decades have been obtained, and major deficiencies in the atmospheric models for tracer transport have been identified. Some quite sophisticated assimilation and computational techniques have been introduced. In the bottom-up approach, model–data fusion has been used for improving the model predictions. Model–data fusion has also been applied in both approaches for analyzing model errors, developing optimal sampling strategies and forecasting. Common to both approaches is that model errors can affect the estimates as much as data error, and are difficult to quantify. More effort is required to identify and quantify model errors.

A few studies have combined the two approaches by applying model–data fusion to constrain key parameters in a process-based model using multiple datasets. We have discussed the importance of considering errors in different datasets as well as in the model consistently. An integrated framework for future model–data fusion application is recommended. Throughout this review, we have emphasized the importance of uncertainties and their propagation. For many applications in terrestrial ecology, the distribution of data errors and model errors are likely to differ significantly from Gaussian. Therefore, other estimation techniques, such as Monte Carlo simulations are expected to become more widely used in the future. They also allow presentation of model–data fusion results as posterior distributions, which are more informative than mean values.

Acknowledgements

We thank the financial support of this work by CSIRO for YPW, CMT and IGE, the Australian Department of Climate Change for YPW and the Australian Research Council to the Centre of Excellence for Mathematics and Statistics of Complex Systems (MASCOS) for IGE. Dr Ray Leuning and Professor Yiqi Luo provided constructive comments on the paper.

References

Aalderlink, R.H., Jovin, J., 1997. Identification of the parameters describing primary production from continuous oxygen signals. Water Science and Technology 36, 43–51.

Abramowitz, G., Pitman, A., Gupta, H., Kowalczyk, E., Wang, Y.P., 2007. Systematic bias in land surface models. Journal of Hydrometeorology 8, 989–1001, doi:10.1175/JHM628.1.

Abramowitz, G., Leuning, R., Clark, M., Pitman, A., 2008. Evaluating the performance of land surface models. Journal of Climate 21, 5468–5481, doi:10.1175/2008JCLI2378.1.

Acton, F.S., 1970. Numerical Methods That Work. Harper and Row, New York.

Baker, D., 2000. An inversion method for determining time-dependent CO_2 fluxes. Inverse methods in global biogeochemical cycles. American Geophysical Monograph, vol. 114. American Geophysical Union, Washington DC, pp. 279–293.

Baker, D. F. 2001. Sources and sinks of atmospheric CO_2 estimated from batch least squares inversions of CO_2 concentration measurements, PhD dissertation, 414 p., Program in Atmospheric and Oceanic Sciences, Princeton University, Princeton, New Jersey.

Baker, D.F., Law, R.M., Gurney, K.R., Rayner, P., Peylin, P., Denning, A.S., Bousquet, P., Bruhwiler, L., Chen, Y.-H., Ciais, P., Fung, I.Y., Heimann, M., John, J., Maki, T., Maksyutov, S., Masarie, K., Prather, M., Pak, B., Taguchi, S., Zhu, Z., 2006a. TransCom 3 inversion intercomparison: impact of transport model errors on the interannual variability of regional CO_2 fluxes, 1983-2003. Global Biogeochemical Cycles 20, GB1002, doi:10.1029/2004GB002439.

Baker, D.F., Doney, S.C., Schimel, D.S., 2006b. Variational data assimilation for atmospheric CO_2. Tellus 58B, 359–365.

Beven, K., Binley, A., 1992. The future of distributed models: model calibration and uncertainty prediction. Hydrological Processes 6, 279–298.

Bousquet, P., Peylin, P., Ciais, P., Le Quéré, C., Friedlingstein, P., Tans, P.P., 2000. Regional changes in carbon dioxide fluxes of land and oceans since 1980. Science 290, 1342–1346.

Braswell, B.H., Sacks, W.J., Linder, E., Schimel, D.S., 2005. Estimating diurnal to annual ecosystem parameters by synthesis of a carbon flux model with eddy covariance net ecosystem exchange observations. Global Change Biology 11, 335–355.

Bruhwiler, L.M.P., Michalak, A.M., Peters, W., Baker, D.F., Tans, P., 2005. An improved Kalman Smoother for atmospheric inversions. Atmospheric Chemistry and Physics 5, 2691–2702.

Carvalhais, N., Reichstein, M., Seixas, J., Collatz, G.J., Pereira, J.S., Berbigier, P., Carrara, A, Granier, A., Montagnani, L., Papale, D., Rambal, S., Sanz, M.J., Valentini, R., 2008. Implications of the carbon cycle steady state assumption for biogeochemical modeling performance and inverse parameter retrieval. Global Biogeochemical Cycles 22, GB2007, doi:10.1029/2007GB003033.

Chatfield, C., 1995. Model uncertainty, data mining and statistical inference. Royal Society Statistical Society Series A 158, 419–466.

Ciais, P., Tans, P., White, J., Trolier, M., Francey, R., Berry, J., Randall, D., Sellers, P., Collatz, J., Schimel, D., 1995. Partitioning of ocean and land uptake of CO_2 as inferred by $\delta^{13}C$ measurements from the NOAA Climate Monitoring and Diagnostics Laboratory Global Air Sampling Network. Journal of Geophysical Research 100 (D3), 5051–5070.

Ciais, P., Reichstein, M., Viovy, N., Granier, A., Ogée, J., Allard, V., Aubinet, M., Buchmann, N., Bernhofer, C., Carrara, A., Chevallier, F., De Noblet, N., Friend, A.D., Friedlingstein, P., Grünwald, T., Heinesch, B., Keronen, P., Knohl, A., Krinner, G., Loustau, D., Manca, G., Matteucci, G., Miglietta, F., Ourcival, J.M., Papale, D., Pilegaard, K., Rambal, S., Seufert, G., Soussana, J.F., Sanz, M.J., Schulze, E.D., Vesala, T., Valentini, R., 2005. Europe-wide reduction in primary productivity caused by the heat and drought in 2003. Nature 437, 529–533.

Comins, H.N., McMurtrie, R.E., 1993. Long-term response of nutrient-limited forests to CO_2-enrichment; equilibrium behaviour of plant-soil models. Ecological Applications 3, 666–681.

Cramer, W., Kicklighter, D.W., Bondeau, A., Moore III, B., Churkina, G., Nemry, B., Ruimy, A., Schloss, A.L., and the Participants of the POTSDAM NPP Model Intercomparison, 1999. Comparing global models of terrestrial net primary productivity (NPP): overview and key results. Global Change Biology 5, 1–15.

Denning, A.S., Holzer, M., Gurney, K.R., Heimann, M., Law, R.M., Rayner, P.J., Fung, I.Y., Fan, S.M., Taguchi, S., Friedlingstein, P., Balkanski, Y., Taylor, J., Maiss, M., Levin, I., 1999. Three-dimensional transport and concentration of SF_6 a model intercomparison study (TransCom 2). Tellus 51B, 266–297.

Enting, I.G., Trudinger, C.M., Francey, R.J., 1995. A synthesis inversion of the concentration and $\delta^{13}C$ of atmospheric CO_2. Tellus 47B, 35–52.

Enting, I.G., 2000. Green's function methods of tracer inversion. Inverse methods in global biogeochemical cycles. American Geophysical Monograph, vol. 114. American Geophysical Union, Washington DC, pp. 19–31.

Enting, I.G., 2002. Inverse Problems in Atmospheric Constituent Transport. Cambridge Univ. Press, New York, p. 392.

Enting, I.G., 2008. Assessing the information content in environmental modelling: a carbon cycle perspective. Entropy 10, 556–575.

Evensen, G., 2003. The ensemble Kalman filter: theoretical formulation and practical implementation. Ocean Dynamics 53, 343–367.

Evensen, G., 2007. Data Assimilation: The Ensemble Kalman Filter. Springer-Verlag, Berlin, p. 280.

Finnigan, J.J., Raupach, M.R., 1988. Single-layer evapotranspiration models are incorrect but useful, whereas multilayer models are correct but useless: discuss. Australian Journal of Plant Physiology 15, 705–716.

Fox, A., Williams, M., Richardson, A.D., Cameron, D., Gove, J.H., Quaife, T., Ricciuto, D., Reichstein, M., Tomelleri, E., Trudinger, C.M., Van Wijk, M.T., 2009. The REFLEX project: Comparing different algorithms and implementations for the inversion of a terrestrial ecosystem model against eddy covariance data. Agricultural and Forest Meteorology 149, 1597–1615, doi:10.1016/j.agrformet.2009.05.002.

Franks, S.W., Beven, K.J., Quinn, P.F., Wright, I.R., 1997. On the sensitivity of soil-vegetation-atmosphere transfer (SVAT) schemes: equifinality and the problem of robust calibration. Agricultural and Forest Meteorology 86, 63–75.

Friedlingstein, P., Cox, P., Betts, R., Bopp, L., Von Bloh, W., Brovkin, V., Cadule, P., Doney, S., Eby, M., Fung, I., Bala, G., John, J., Jones, C., Joos, F., Kato, T., Kawamiya, M., Knorr, W., Lindsay, K., Matthews, H.D., Raddatz, T., Rayner, P., Reick, C., Roeckner, E., Schnitzler, K.G., Schnur, R., Strassmann, K., Weaver, A.J., Yoshikawa, C., Zeng, N., 2006. Climate–carbon cycle feedback analysis: results from the C^{4}MIP Model Intercomparison. Journal of Climate 19, 3337–3353, doi:10.1175/JCLI3800.1.

Garthwaite, P.H., Jolliffe, I.T., Jones, B., 2002. Statistical Inference, Second edition. Oxford University Press, Oxford, p. 328.

Gelb, A., 1974. Applied Optimal Estimation. MIT Press, Cambridge, MA, p. 374.

Gove, J.H., Hollinger, D.Y., 2006. Application of a dual unscented Kalman filter for simultaneous state and parameter estimation in problems of surface-atmosphere exchange. Journal of Geophysical Research 11 , doi:10.1029/2005JD00621.

Gurney, K.R., Law, R.M., Denning, A.S., Rayner, P.J., Baker, D., Bousquet, P., Bruhwilerk, L., Chen, Y.H., Ciais, P., Fan, S.M., Fung, I.Y., Gloor, M., Heimann, M., Higuchi, K., John, J., Maki, T., Maksyutov, S., Masarie, K., Peylin, P., Prather, M., Pak, B.C., Randerson, J., Sarmiento, J., Taguchi, S., Takahashi, T., Yuen, C.W., 2002. Towards robust regional estimates of CO_2 sources and sinks using atmospheric transport models. Nature 415, 626–630.

Gurney, K.R., Law, R.M., Denning, A.S., Rayner, P.J., Baker, D., Bousquet, P., Bruhwilerk, L., Chen, Y.H., Ciais, P., Fan, S.M., Fung, I.Y., Gloor, M., Heimann, M., Higuchi, K., John, J., Kowalczyk, E., Maki, T., Maksyutov, S., Peylin, P., Prather, M., Pak, B.C., Sarmiento, J., Taguchi, S., Takahashi, T., Yuen, C.W., 2003. TransCom3 CO_2 inversion intercomparison: 1. Annual mean control results and sensitivity to transport and prior flux information. Tellus 55B, 555–579.

Harley, P.C., Thomas, R.B., Reynolds, J.F., Strain, B.R., 1992. Modelling photosynthesis of cotton growth in elevated CO_2. Plant, Cell and Environment 15, 271–282.

Heimann, M., Maier-Reimer, E., 1995. On the relations between oceanic budgets of CO_2 and its isotopes. Global Biogeochemical Cycles 10, 89–110.

Hoeting, J., Madigan, A.D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: a tutorial. Statistical Sciences 14 (4), 382–417.

Hsu, K., Gupta, H.V., Gao, X., Sorooshian, S., Imam, B., 2002. Self-Organizing Linear Output map (SOLO): an artificial neural network suitable for hydrological modeling and analysis. Water Resources Research 38 (12), 1302, doi:10.1029/2001WR000795.

Kalman, R.E., 1960. A new approach to linear filtering and prediction problems. Journal of Basic Engineering (ASME) 82D, 35–45.

Kalman, R.E., Bucy, R., 1961. New results in linear filtering and prediction theory. Journal of Basic Engineering (ASME) 83D, 98–108.

Kaminski, T., Knorr, W., Rayner, P., Heimann, M., 2002. Assimilating atmospheric data into a terrestrial biosphere model: a case study of the seasonal cycle. Global Biogeochemical Cycles 16, 1066, doi:10.1029/2001GB001463.

Katul, G., Lai, C.T., Schäfer, K., Vidakovic, B., Albertson, J., Ellsworth, D., Oren, R., 2001. Multiscale analysis of vegetation surface fluxes: from seconds to years. Advances in Water Resources 24, 1119–1132.

Keeling, C.D., Piper, S.C., Heimann, M., 1989. A three-dimensional model of atmospheric CO_2 transport based on observed winds: 4. Mean annual gradients and interannual variations. Aspects of Climate Variability in the Pacific and the Western Americas, Geophysical Monograph Series, vol. 55. American Geophysical Union, Washington, DC, pp. 305–363.

Law, R.M., Rayner, P.J., Wang, Y.P., 2004. Inversion of diurnally varying synthetic CO_2: network optimization for an Australian test case. Global Biogeochemical Cycles 18, GB1044, doi:10.1029/2003GB002136.

Liu, Y.Q., Gupta, H.V., 2007. Uncertainty in hydrological modeling: toward an integrated data assimilation framework. Water Resource Research 43, W07401, doi:10.1029/2006WR005756.

Luo, Y., Reynolds, J.F., 1999. Validity of extrapolating field CO_2 experiments to predict carbon sequestration in natural ecosystems. Ecology 80, 1568–1583.

Luo, Y., White, L., Canadell, J., DeLucia, E., Ellsworth, D., Finzi, A., Lichter, J., Shclesinger, W., 2003. Sustainability of terrestrial carbon sequestration: a case study in Duke Forest with inversion approach. Global Biogeochemical Cycles 17 (1), 1021, doi:10.1029/2002GB001923.

Mathieu, P.-P., O'Neill, A., 2008. Data assimilation: from photon counts to Earth System forecasts. Remote Sensing of Environment 112 (4), 1258–1267.

Meldyn, B.E., Robinson, A.P., Clement, R., McMurtrie, R.E., 2005. On the validation of models of forest CO_2 exchange using eddy covariance data: some perils and pitfalls. Tree Physiology 25, 839–857.

Medlyn, B.E., McMurtrie, R.E., 2005. Effects of CO_2 on plants at different timescales. In: A History of Atmospheric CO_2 and its Impacts on Plants, Animals, and Ecosystems, Springer-Verlag, pp. 441–467.

Menke, W., 1989.. Geophysical data analysis: discrete inverse theory. International Geophysics Series, Second edition, vol. 45. Academic Press, San Diedgo, p. 289.

Ménard, R., Chang, L.P., 2000. Assimilation of stratospheric chemical tracer observations using a Kalman filter. Part II: χ^2-validated results and analysis of variance and correlation dynamics. Monthly Weather Review 128, 2672–2686.

Michalak, A.M., Hirsch, A., Bruhwiler, L., Gurney, K.R., Peters, W., Tans, P.P., 2005. Maximum likelihood estimation of covariance parameters for Bayesian atmospheric trace gas surface flux inversions. Journal of Geophysical Research – Atmospheres 110, D24107, doi:10.1029/2005JD005970.

Mo, X.G., Beven, K., 2004. Multi-objective parameter conditioning of a three-source wheat canopy model. Agricultural and Forest Meteorology 122, 39–63.

Palmer, T.N., Brankovic, C., Richardson, D.S., 2000. A probability and decision-model analysis of PROVOST seasonal multi-model ensemble integrations. The Quarterly Journal of the Royal Meteorological Society 126 (567), 2013–2033.

Peters, W., Miller, J.B., Whitaker, J., Denning, A.S., Hirsch, A., Krol, M.C., Zupanski, D., Bruhwiler, L., Tans, P.P., 2005. An ensemble data assimilation system to estimate CO_2 surface fluxes from atmospheric trace gas observations. Journal of Geophysical Research – Atmospheres 110 (D24), D24304.

Peters, W., Jacobson, A.R., Sweeney, C., Andrews, A.E., Conway, T.J., Masarie, K., Miller, J.B., Bruhwiler, L.M.P., Pétron, G., Hirsch, A.I., Worthy, D.E.J., van der Werf, G.R., Randerson, J.T., Wennberg, P.O., Krol, M.C., Tans, P.P., 2007. An atmospheric perspective on North American carbon dioxide exchange: CarbonTracker. Proceedings of the National Academy of Sciences of the United States of America 104, 18925–18930.

Peylin, P., Bousquet, P., Le Quéré, C., Peylin, P., Bousquet, P., Le Quéré, C., Sitch, S., Friedlingstein, P., McKinley, G., Gruber, N., Rayner, P., Ciais, P., 2005. Multiple constraints on regional CO_2 flux variations over land and oceans. Global Biogeochemical Cycles 19, GB1011, doi:10.1029/2003GB002214.

Piao, S.L., Ciais, P., Friedlingstein, P., Peylin, P., Reichstein, M., Luyssaert, S., Margolis, H., Fang, J.Y., Barr, A., Chen, A.P., Grelle, A., Hollinger, D.Y., Laurila, T., Lindroth, A., Richardson, A.D., Vesala, T., 2008. Net carbon dioxide losses of northern ecosystems in response to autumn warming. Nature 451, 49–53.

Press, W.H., Teukolsky, S.A., Vetterling, W., Flannery, B.P., 2001. Numerical Recipes in Fortran 77: The Art of Scientific Computing, Second edition. Cambridge University Press, Cambridge.

Prihodko, L., Denning, A.S., Hanan, N.P., Baker, I., Davis, K., 2008. Sensitivity, uncertainty and time dependence of parameters in a complex land surface model. Agricultural and Forest Meteorology 148, 268–287.

Quay, P.D., Tilbrook, B., Wong, C.S., 1992. Oceanic uptake of fossil fuel CO_2: carbon-13 evidence. Science 256, 74–79.

Rayner, P.J., Enting, I.G., Trudinger, C.M., 1996. Optimizing the CO_2 observing network for constraining sources and sinks. Tellus 48B (4), 433–444.

Rayner, P.J., Enting, I.G., Francey, R.J., Langenfelds, R.L., 1999. Reconstructing the recent carbon cycle from atmospheric CO_2, $\delta^{13}C$ and O_2/N_2 observations. Tellus 51B, 213–232.

Rayner, P.J., O'Brien, D.M., 2001. The utility of remotely sensed CO_2 concentration data in surface source inversions. Geophysical Research Letters 28 (1), 175–178.

Rayner, P., Scholze, M., Knorr, W., Kaminski, T., Giering, R., Widmann, H., 2005. Two decades of terrestrial carbon fluxes from a Carbon Cycle Data Assimilation System (CCDAS). Global Biogeochemical Cycles 19 , doi:10.1029/2004GB002254.

Rayner, P.J., Law, R.M., Allison, C.E., Francey, R.J., Trudinger, C.M., Pickett-Heaps, C.A., 2008. Interannual variability of the global carbon cycle (1992–2005) inferred by inversion of atmospheric CO_2 and $\delta^{13}CO_2$ measurements. Global Biogeochemical Cycles 22, doi:10.1029/2007GB003068.

Raupach, M.R., Rayner, P.J., Barrett, D.J., DeFries, R., Heimann, M., Ojima, D.S., Quegan, S., Schmullius, C., 2005. Model-data synthesis in terrestrial carbon observation: methods, data requirements and data uncertainty specifications. Global Change Biology 11, 378–397, doi:10.1111/j.1365-2486.2005.00917.x.

Reichstein, M., Tenhunen, J., Roupsard, O., Ourcival, J.-M., Rambal, S., Miglietta, F., Peressotti, A., Pecchiari, M., Tirone, G., Valentini, R., 2003. Inverse modelling of seasonal drought effects on canopy CO_2/H_2O exchange in three Mediterranean ecosystems. Journal of Geophysical Research – Atmospheres 108 (D23), doi:10.1029/2003JD003430.

Richardson, A.D., Hollinger, D.Y., 2005. Statistical modeling of ecosystem respiration using eddy covariance data: maximum likelihood parameter estimation, and Monte Carlo simulation of model and parameter uncertainty, applied to three simple models. Agricultural and Forest Meteorology 131, 191–208.

Richardson, A.D., Mahecha, M.D., Falge, E., Kattge, J., Moffat, A.M., Papale, D., Reichstein, M., Stauch, V.J., Braswell, B.H., Churkina, G., Kruijt, B., Hollinger, D.Y., 2008. Statistical properties of random CO_2 flux measurement uncertainty inferred from model residuals. Agricultural and Forest Meteorology 148, 38–50.

Rödenbeck, C., Houweling, S., Gloor, M., Heimann, M., 2003. CO_2 flux history 1982-2001 inferred from atmospheric data using a global inversion of atmospheric transport. Atmospheric Chemistry and Physics 3, 1919–1964.

Sacks, W.J., Schimel, D.S., Monson, R.K., 2007. Coupling between carbon cycling and climate in a high-elevation, subalpine forest: a model-data fusion analysis. Oecologia 151, 54–68.

Saleske, S.R., Miller, S.D., Matross, D.M., Goulden, M.L., Wofsy, S.C., da Rocha, H.R., de Camargo, P.B., Crill, P., Daube, B.C., de Freitas, H.C., Hutyra, L., Keller, M., Kirchhoff, V., Menton, M., Munger, J.M., Pyle, E.H., Rice, A.H., Silva, H., 2003. Carbon in Amazon forests: unexpected seasonal fluxes and disturbance-induced losses. Science 302, 1554–1557.

Saltelli, A., Tarantola, S., Campolongo, F., Ratto, M., 2004. Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models. Wiley, Chichester.

Santaren, D., Peylin, P., Viovy, N., Ciais, P., 2007. Optimizing a process-based ecosystem model with eddy-covariance flux measurements: a pine forest in southern France. Global Biogeochemical Cycles 21, GB2013, doi:10.1029/2006GB002834.

Sarmiento, J.L., Sundquist, E., 1992. Revised budget for the oceanic uptake of anthropogenic carbon dioxide. Nature 356, 589–593.
Schimel, D., Melillo, J., Tian, H.Q., McGuire, A.D., Kicklighter, D., Kittel, T., Rosenbloom, N., Running, S., Thornton, P., Ojima, D., Parton, W., Kelly, R., Sykes, M., Neilson, R., Rizzo, B., 2000. Contribution of increasing CO_2 and climate to carbon storage by ecosystems in the United States. Science 287, 2004–2006.
Schimel, D.S., House, J.I., Hibbard, K.A., Bousquet, P., Ciais, P., Peylin, P., Braswell, B.H., Apps, M.J., Baker, D., Bondeau, A., Canadell, J., Churkina, G., Cramer, W., Denning, A.S., Field, C.B., Friedlingstein, P., Goodale, C., Heimann, M., Houghton, R.A., Melillo, J.M., Moore III, B., Murdiyarso, D., Noble, I., Pacala, S.W., Prentice, I.C., Raupach, M.R., Rayner, P.J., Scholes, R.J., Steffen, W.L., Wirth, C., 2001. Recent patterns and mechanisms of carbon exchange by terrestrial ecosystems. Nature 414, 169–172.
Schimel, D.S., Churkina, G., Braswell, B.H., Trembath, J., 2005. Remembrance of weather past: ecosystem response to climate variability. In: A History of Atmospheric CO_2 and Its Effects on Plants, Animals and Ecosystems, Springer-Verlag, pp. 350–368.
Scholze, M., Kaminski, T., Rayner, P., Knorr, W., Giering, R., 2007. Propagating uncertainty through prognostic carbon cycle data assimilation system simulations. Journal of Geophysical Research – Atmospheres 112, D17305, doi:10.1029/2007JD008642.
Shaby, B.A., Field, C.B., 2006. Regression tools for CO_2 inversions: application of a shrinkage estimator to process inversion. Tellus 58B, 279–292.
Stoy, P.C., Katul, G.G., Siqueira, M.B.S., Juang, J.Y., McCarthy, H.R., Kim, H.S., Oishi, A.C., Oren, R., 2005. Variability in net ecosystem exchange from hourly to inter-annual time scales at adjacent pine and hardwood forests: a wavelet analysis. Tree Physiology 25, 887–902.
Stöckli, R., Lawrence, D.M., Niu, G.Y., Oleson, K.W., Thornton, P.E., Yang, Z.-L., Bonan, G.B., Denning, A.S., Running, S.W., 2008. Use of FLUXNET in the community land model development. Journal of Geophysical Research – Biogeosciences 113, G01025, doi:10.1029/2007JG0000562.
Tang, J., Zhuang, Q., 2008. Equifinality in parameterization of process-based biogeochemistry models: a significant uncertainty source to the estimation of regional carbon dynamics. Journal of Geophysical Research – Biogeosciences 113, G04010, doi:10.1029/2008JG000757.
Tans, P.P., Fung, I.Y., Takahashi, T., 1990. Observational constraints on the global atmospheric CO_2 budget. Science 247, 1431–1438.
Tans, P.P., Berry, J.A., Keeling, R.F., 1993. Oceanic $^{13}C/^{12}C$ observations: a new window on ocean CO_2 uptake. Global Biogeochemical Cycles 7, 353–368.
Tarantola, A., 1987. Inverse Problem Theory: Methods for Data Analysis and Model Parameter Estimation. Elsevier, Amsterdam.
Tarantola, A., 2005. Inverse Problem Theory and Methods for Model Parameter Estimation. SIAM, Philadelphia, p. 342.
Todling, R., 2000. Estimation theory and atmospheric data assimilation. Inverse methods in global biogeochemical cycles. American Geophysical Monograph, vol. 114. American Geophysical Union, Washington DC, pp. 49–65.
Trudinger, C.M., Enting, I.G., Rayner, P.J., 2002. Kalman filter analysis of ice core data, 1. Method development and testing the statistics. Journal of Geophysical Research – Atmospheres 107 (D20), 4422, doi:10.1029/2001JD001111.
Trudinger, C.M., Raupach, M.R., Rayner, P.J., Kattge, J., Liu, Q., Pak, B., Reichstein, M., Renzullo, L., Richardson, A.D., Roxburgh, S.H., Styles, J., Wang, Y.P., Briggs, P., Barrett, D., Nikolova, S., 2007. OptIC project: an intercomparison of optimization techniques for parameter estimation in terrestrial biogeochemical models. Journal of Geophysical Research – Biogeosciences 112, G02027, doi:10.1029/2006JG000367.
Trudinger, C.M., Raupach, M.R., Rayner, P.J., Enting, I.G., 2008. Using the Kalman filter for parameter estimation in biogeochemical models. Environmetrics, doi:10.1002/env.910.
Wang, Y.P., Leuning, R., Cleugh, H.A., Coppin, P.A., 2001. Parameter estimation in surface exchange models using non-linear inversion: how many parameters can we estimate and which measurements are most useful? Global Change Biology 7, 495–510.
Wang, Y.P., Barrett, D.J., 2003. Estimating regional terrestrial carbon fluxes for the Australian continent using a multiple-constraint approach I. Using remotely sensed data and ecological observations of net primary production. Tellus 55, 270–289.
Wang, Y.P., McGregor, J.L., 2003. Estimating regional terrestrial carbon fluxes for the Australian continent using a multiple-constraint approach II. The atmospheric constraint. Tellus 55, 290–304.
Wang, Y.P., McMurtrie, R.E., Medlyn, B.E., Pepper, D., 2006. Modelling plant ecosystem responses to elevated CO_2 at decadal to century timescales. In: Plant Growth and Climate Change, Blackwell Publishing, pp. 165–186.
Wang, Y.P., Baldocchi, D., Leuning, R., Falge, E., Vesala, T., 2007. Estimating parameters in a land surface model by applying nonlinear inversion to eddy covariance flux measurements from eight FLUXNET sites. Global Change Biology 13, 652–670, doi:10.1111/j.1365-2486.2006.01225x.
White, L.W., Luo, Y.Q., 2002. Estimation of carbon transfer coefficients using Duke Forest free-air CO_2 enrichment data. Applied Mathematics and Computation 130, 101–120.
White, L.W., Luo, Y., Xu, T., 2005. Carbon sequestration: inversion of FACE data and prediction. Applied Mathematics and Computation 163, 783–800.
Williams, M., Schwarz, P.A., Law, B.E., Irvine, J., Karpius, M.R., 2005. An improved analysis of forest carbon dynamics using data assimilation. Global Change Biology 11, 89–105.
Williams, M., Richardson, A.D., Reichstein, M., Stoy, P.C., Peylin, P., Verbeeck, H., Carvalhais, N., Jung, M., Hollinger, D.Y., Kattge, J., Leuning, R., Luo, Y., Tomelleri, E., Trudinger, C.M., Wang, Y.-P., 2009. Improving land surface models with FLUXNET data. Biogeosciences 6, 1341–1359.
Xu, T., White, L., Hui, D., Luo, Y., 2006. Probabilistic inversion of a terrestrial ecosystem model: analysis of uncertainty in parameter estimation and model prediction. Global Biogeochemical Cycles 20, GB2007, doi:10.1029/2005GB002468.
Yuan, W.P., Liu, S.G., Zhou, G.S., et al., 2007. Deriving a light use efficiency model from eddy covariance flux data for predicting daily gross primary production across biomes. Agricultural and Forest Meteorology 143, 189–207.
Zobitz, J.M., Moore, D., Sacks, W.J., Monson, R.K., Bowling, D.R., Schimel, D.S., 2008. Integration of process-based soil respiration models with whole-ecosystem CO_2 measurements. Ecosystems 11, 250–269.