Available online at www.sciencedirect.com
ScienceDirect
Current Opinion in Environmental Sustainability

The current state of carbon-cycle data assimilation

Peter J Rayner

Model–data fusion (MDF), also called 'data assimilation', describes the approach of combining as much data as possible with explicit dynamical models to constrain the evolution of any system. The paradigm is weather prediction. MDF holds the promise of combining the plethora of data available on the terrestrial carbon cycle but to date there have been few successful applications using multiple streams of data. The two most likely reasons for this are a lack of statistical sophistication in the attempts and weaknesses in current terrestrial biosphere models; these failures do not represent a failure of the paradigm. Solutions lie in a more rigorous consideration of the statistics of the problem and the inclusion of more elements of the model within the MDF framework.

Address
LSCE/CEA-Saclay, Bat. 701, Orme des Merisiers, Gif sur Yvette 91191, France

Corresponding author: Rayner, Peter J (peter.rayner@cea.fr)

Current Opinion in Environmental Sustainability 2010, **2**:289–296

This review comes from a themed issue on Carbon and nitrogen cycles
Edited by Josep G Canadell

Received 8 February 2010; Accepted 10 May 2010
Available online 23rd June 2010

1877-3435/$ – see front matter

DOI 10.1016/j.cosust.2010.05.005

Introduction

The behavior of the terrestrial biosphere, now and in the future, represents a critical uncertainty in the global carbon cycle [1,2]. The terrestrial biosphere is more sensitive to direct human influence than the ocean as demonstrated by the large but uncertain flux of carbon to the atmosphere caused by land-use change [3]. This makes it a tempting target for interventions aimed at limiting atmospheric CO_2 concentrations. Finally the terrestrial biosphere is highly sensitive to short term climate variability [4,5] and among models this sensitivity is mirrored by a widely varying response to future climate change [6,7]. For these reasons, quantification of the current and future state of the terrestrial biosphere has attained practical as well as scientific importance.

The importance of the problem is matched by its difficulty. One obvious problem is the heterogeneity of the biosphere. One can regard the conditions of soil type, climate and land-use history as exogenous to the biosphere.[1] If one maps the overlay of these three factors one immediately sees variations on the scale of hectares. If we regard the biosphere as a simple deterministic system responding to all these inputs we see already that any local measurement represents one point in an immense space of potential drivers. Thus the sampling problem to characterize the response to all these drivers is serious indeed.

The intractability of the measurement problem allied with the need for prediction provides practical motivation to construct models. The intellectual motivation is also to construct coherent and comprehensible views of the system. We know that similar physical and chemical processes occur in many different situations so we do have some basis for building generalized models. The large-scale responses to climate anomalies [e.g. 8, 5] suggest that there are coherent responses to such drivers but we cannot guarantee *a priori* that the emergent properties of the biosphere are predictable from understanding simple processes. Ref. [9] noted the difficulty of building predictive models. Various processes are not included in terrestrial biosphere models. Thus they fail to match some observations, casting doubt on their predictions. The proposed solution was to urge the testing of such models against as wide a range of observations as possible. Here I advocate the same solution but note that the traditional approach of running the model and comparing with observations does not account for the uncertainty of parameters in the model and is not necessarily an aid to model development.

Commensurate with the complexity of biospheric behavior and range of scales on which it responds, there are a wide range of measurements probing different aspects and covering different levels of spatial and temporal integration. These were summarized in [10•]. The question of how to combine these measurements with models is the subject of this paper. Of course the problem of inferring underlying causes from their effects is widespread in science. The algorithms to do this are often termed inverse approaches since they reverse the direction of causality. The new field of carbon-cycle data assimilation inherits much of the apparatus and

[1] This is obviously a simplification since the evolution of soil is a result of biospheric function and even climate is influenced by changes in land cover.

intellectual background from other fields but the analogy is nowhere exact. It probably shares most with the assimilation used in modern numerical weather prediction with common points like a dynamical rather than static model, multiple types of measurements and an interest in forecasting. Numerical weather prediction does not usually concern itself with improving the underlying process understanding and work in the carbon cycle lacks the day-to-day testing of forecast evaluation but this is the paradigm I use.

After introducing the abstract framework I will briefly summarize the state of the art at the time of writing. This will highlight two significant problems, namely the statistical treatment of the observations and the elucidation of flaws in the models themselves. The last section of the paper will suggest approaches for addressing these.

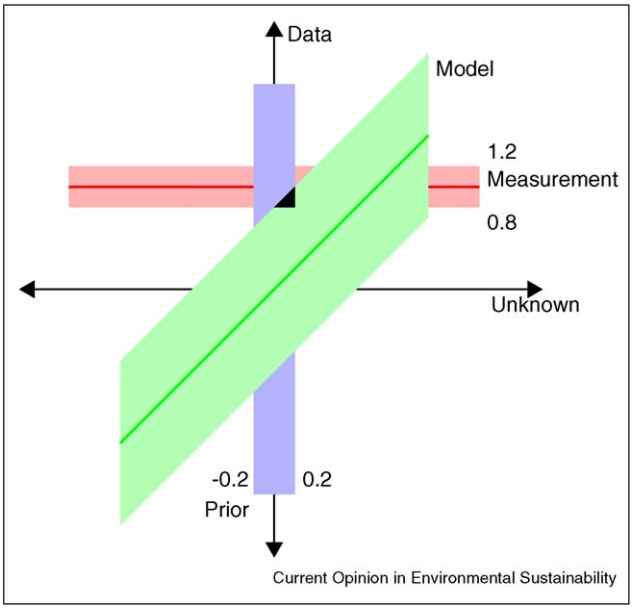
A simple theory of model–data fusion

The usual description of the task of MDF is to constrain the behavior of a model by requiring it to match certain observations. I will show later that it is equally important to quantify the consistency between the model and the observations. This problem has been routinely treated differently within Applied Mathematics and Statistics. Ref. [11] showed the equivalence of these different descriptions. The statistical description [e.g. 12••] has come to dominate the field and we follow this approach.

Although it is possible to develop the theory using the most general description of a model I will simplify the definition here to illustrate the approach. I define a model as a numerical calculation which takes a series of inputs and calculates a series of outputs. The inputs can be divided into external forcing, initial conditions (both boundary conditions in the language of differential equations) and a series of parameters. We group these as the ''unknowns'' of the problem.[2] The usual case is a model which starts from an initial state (e.g. the current state of the atmosphere) and marches forward in time. At each timestep the model calculates a new state described by a set of state variables. The outputs from the model comprise both things we observe or things we would like to know. Often some of the intermediate state variables are observed but just as often the observables are themselves the product of complex calculation from state variables.

The MDF problem consists of using the observables to improve our knowledge of the unknowns and hence of the desired outputs. The problem is complicated by imperfections in both the model and the data while it is usually ameliorated by the existence of prior knowledge (also imperfect) of the unknowns. Mathematically the task is to express then combine these three sources of information. We achieve this by expressing the information in the form of probability density functions (PDFs) then combining them by the laws of probability [see 12••, for a full development].

Figure 1



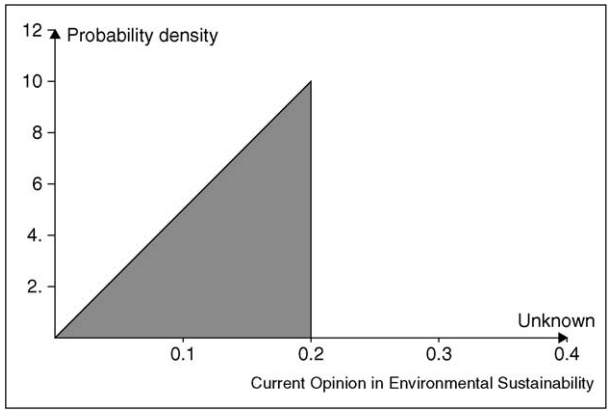
Schematic describing MDF with one unknown and one observable. The meaning of the colored regions is described in the text.

Figure 1 illustrates the approach for the simplest case of one unknown and one observable. The unknown is plotted on the X-axis and the observable on the Y-axis. The prior PDF of the unknown is represented by the vertically aligned, light blue rectangle between $x = -0.2$ and $x = 0.2$.[3] The horizontally aligned, light red rectangle between $y = 0.8$ and $y = 1.2$ represents the PDF of the true value of the observable. This is likely to be based on a measurement indicated by the heavy red line. In this case the measurement corresponds to the expectation of the true value, that is, the measurement is unbiased. This is often not the case. Although we usually regard such measurement bias as an experimental problem it can equally be regarded (and often treated) as a statistical problem of the measurement PDF. The deterministic model is shown by the heavy green line. It describes what value the observable should have for a given value of the unknown. The model is imperfect however so the true

[2] In numerical weather prediction the unknowns are usually the initial state of the atmosphere.

[3] We do not show a heavy blue line indicating the central estimate to emphasize that the prior PDF can be a range from a literature survey or a range of physically meaningful values.

Figure 2



Solution of the inverse problem by projecting the solution shown in Figure 1 onto the unknown axis and normalizing it.

value of the observable lies within the green region. The solution is the region of the unknown-observable space consistent with all three inputs, the black triangle.

The 'solution' of the inverse problem is generated by projecting the black triangle shown in Figure 1 onto the unknown axis and normalizing it to give a total probability of 1. This is shown in Figure 2.

Many interesting features of data assimilation are evident in Figures 1 and 2. Figure 2 shows that only values of the unknown between 0.0 and 0.2 are possible after the measurement, that is, we have added information. The most likely value of the unknown is 0.2 while its mean is ≈ 0.13333. Figure 1 shows that the perfect model (green line) is nowhere consistent with the prior unknown and measurement PDFs. Had the light green region been much narrower there would have been no overlap. This would constitute a falsification of the assumptions since the probability that our three pieces of information (prior knowledge, measurement and model) were simultaneously true would be zero. This would cast serious doubt on the model.

Real-world data assimilation problems are vastly larger than the one shown in Figure 1 but this is a computational not intellectual matter. More important is the usual form chosen to represent probabilities. We usually choose a Gaussian of the form

$$P(x) = K\exp\ -\frac{(x-\mu)^2}{2\sigma^2} \qquad (1)$$

where μ is the mean, σ the standard deviation and K a normalization constant. σ describes the width of the distribution while μ describes its position. One important consequence of forms like Eqn (1) is that probabilities never equal 0 but it remains true that some combinations of model, prior information and measurement are highly unlikely. Secondly the parameters μ and σ are sufficient to describe the PDF completely. Thirdly if the model is linear we can show that if the three input PDFs are Gaussian the combined PDF will also be Gaussian [12••, Chapter 1]. In this case it suffices to find the mean and standard deviation of the posterior PDF to describe it completely.

Various other approximations and simplifications are necessary to turn the above theory into a practical assimilation system and there is a vast literature on the computational machinery for describing the posterior PDF or maximizing it but that is not critical to this discussion. Readers are referred to texts such as [12••,13] for the general theory, [14] for its application to biogeochemistry and [15] for aspects of numerical optimization.

Applying assimilation to the carbon cycle

The theory illustrated in 'A simple theory of model–data fusion' is exceedingly general. Within carbon-cycle research it has been applied in two forms. In flux inversions [e.g. 4,5,16–21] it is used to estimate surface carbon fluxes without the use of an explicit dynamical model.[4]

In assimilation studies we add an underlying dynamical constraint such as a carbon-cycle model.[5] This imposes relationships among model outputs. Consequently observations can carry more weight since their influence is propagated by model dynamics. We see immediately that the quality of the inferences made depends critically on that of the dynamical model. It also depends on the quality of the mapping between quantities simulated by the model and those observed. This mapping is achieved by a submodel usually termed an observation operator. These can be as simple as an identity if the simulated quantity is directly observed and as complex as a radiative transfer model or atmospheric transport model.

A discussion of the relative value of inversion and assimilation is beyond my scope and is discussed in [22•,23]. We can summarize some of the important points as:

- Only assimilation allows prediction.
- Inversion does not filter atmospheric data through imperfect carbon-cycle models.

[4] However the outputs of dynamical models often inform the prior PDFs of fluxes.

[5] Since assimilation as described here is technically more difficult it is often regarded as a desirable word in a title. The value judgment is incorrect and the field would be well served if the distinction was respected.

- Assimilation greatly reduces the underdeterminacy of the inverse problem [24].
- Assimilation can ingest a much wider variety of data.

For our purposes the last point is key. The ability of data assimilation systems to integrate the range of observations made of the carbon cycle makes them an attractive tool and an impressive intellectual structure for large measurement programs. This ability, however, imposes its own demands. For example, we never have just one measurement and the measurements are usually not of the same quantity. As we add new measurements the posterior PDF of the unknowns becomes more tightly focused, that is, their uncertainty is reduced. It is possible that the PDF will be reduced to zero, that is we cannot find a solution consistent with our prior information, model and measurements. We cannot know without considering simultaneously the uncertainties in all three components.

The question then is whether we can generate a quantitative description of the carbon cycle consistent with all available observations. In the next section I shall detail some recent steps toward this and draw some lessons from their common difficulties. I will break the problem down into some technical aspects, then describe some more fundamental problems.

The technical challenge

Before starting we need to know whether we could identify our target model even if we know it exists. Concretely this means describing the PDF of the unknowns consistent with the PDFs of model dynamics and measurements. [25•] showed that many of the techniques currently used for such calculations worked well for a simple biosphere model at one site. Despite the simplicity of the model it demonstrated sufficiently complex dynamics [26] to be a reasonable test of assimilation methods. [25•] found that the choice of cost function was much more important in the success of an assimilation than the choice of assimilation methods. A similar finding was noted by [27], who found that the experience of the practitioners was more important than the method chosen. Most of the methods trialed in [25•] were successful where success could be reasonably expected. Thus it seems the barriers to a carbon-cycle data assimilation system are not technical.

Experience to date

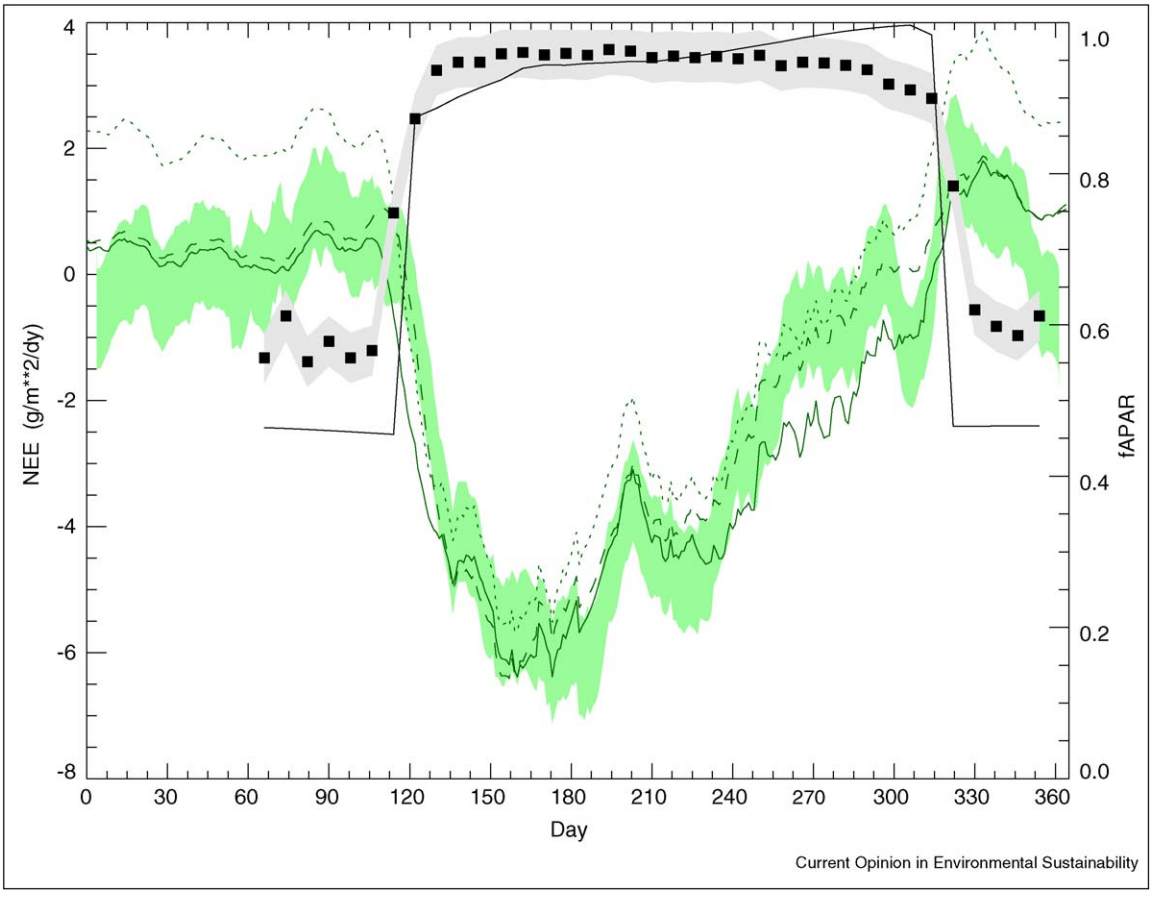
There have been a considerable number of studies using formal parameter estimation methods to constrain carbon-cycle models, for example [22•,23,28–37,38•]. Rather than give a comment on each of these here I have annotated them in the bibliography. Most of these studies have taken a similar form: perform an assimilation, check the quality of fit to the observations then make inferences on the carbon cycle using the constrained model.

There are two other important commonalities. Few of these studies use more than one type of data, for example, vegetation indices, eddy covariance measurements, and soil carbon measurements. One reason for this is structural. Experience from numerical weather prediction and the experience from CCDAS suggest that the ingestion of a new data type into a CCDAS requires two experts, one who understands the vaguaries of the data (biases, levels of dependence among different measurements, etc.) and another capable of making the inevitable changes in the model needed to interface to the new data. The large consortia necessary to handle multiple streams of data generally do not exist within the carbon-cycle community.

The other reason is more profound and is demonstrated in Figure 3 taken from (Bacour, personal communication, Joint assimilation of eddy-covariance flux measurements and satellite observations within a process-oriented biosphere model, Glob Biogeochem Cycle 24, unpublished data) [39]. Here we attempt to fit simultaneously *in situ* measurements of the fraction of absorbed photosynthetically active radiation (FAPAR) and fluxes of CO_2 and water vapor. We optimize a series of parameters including those that control the timing of leaf onset and senescence and relationships between CO_2 and latent heat fluxes. We assimilate the observations either separately by type or together. The figure shows observations of FAPAR and CO_2 flux surrounded by their 1σ uncertainties. We show the prior simulation of CO_2 flux plus the simulation assimilating either FAPAR or FAPAR and flux. Comparing the dashed and dotted green lines we see that assimilating FAPAR slightly degrades the fit to the CO_2 flux. Assimilating both quantities (solid green line) significantly improves the fit to the CO_2 flux data. Comparing the FAPAR simulation (solid black line) to its observations however shows that we cannot simultaneously fit both sets of observations. This suggests the model does not find the CO_2 and FAPAR observations consistent. So rather than optimize parameters of the model we have falsified some element of its structure. Far from seeing this as a disappointment I would argue it is an exemplary application of data assimilation. Note that if we had not carried out the parameter optimization we could never have distinguished between parametric and structural errors in the model.

The second common feature is rather a common lack. Only one of the cited studies [38•] attempts to test their parameter set against truly independent data. Even that study imposed the phenology both at the site for the assimilation and at the validation site. My comment is not a critique of these studies since it was not their aim to produce a general model of the terrestrial carbon cycle. However it is only a test against independent data which can give confidence in the generality and predictive power of an optimized model.

Figure 3



Comparison between flux and FAPAR observation data and the corresponding ORCHIDEE model simulations before assimilation (dotted), after assimilating FAPAR data (dashed) and after assimilating both FAPAR and *in situ* flux data (solid) at Fontainebleau France. Green lines represent NEE while the black line represents FAPAR. The light green band shows the measured $\pm 1\sigma$ confidence interval for the observed NEE while the gray region shows that for the FAPAR. Black squares represent FAPAR observations used in the assimilation.

It appears then that current terrestrial carbon-cycle models fail the test of generality, that is it is difficult to transfer understanding gained at one site or from one type of measurement to other places and other quantities. This is only to be expected. Conditions differ from site to site and we should not expect our models to function in the absence of complete knowledge. The methodology of data assimilation accounts for this since site-specific unknowns can be optimized at each site while still testing whether parameters optimized at one site are consistent with measurements at a second. Returning to Figure 1 we can ask where the prior information on the unknown came from. In reality it came from other measurements or experiments and the width of the distribution reflects the uncertainty in the unknown *after* these previous measurements. The uncertainty in the model simulation based on this prior PDF is a proper reflection of how much we expect to be able to transfer our knowledge. It is the compatibility of this distribution with the PDF of the measurement which tells us whether our model has succeeded; a simple simulation of the mean model with the mean value of the prior would tell us almost nothing.

Some reasons for failure

If we accept that data assimilation has not delivered a general model of the terrestrial carbon cycle we can suggest some reasons for the problem.

Scaling. The different measurements of the terrestrial carbon-cycle appertain to different temporal and spatial scales. Ref. [40] pointed out that parameters that hold at one scale may not be valid for another. Thus if we infer the maximum photosynthetic rate ($_{MAX}$) from a global assimilation using a model run at $2^\circ \times 2^\circ$ we should not expect it to apply to measurements from a point source like a flux tower. This is a computational not fundamental problem. Ref. [9] proposed the use of structural ecosystem models as one solution.

Information content of measurements. When assimilating any set of measurements we must assess the information which can be extracted from them. This is entirely model dependent as can be seen from Figure 1 where the model uncertainty and measurement uncertainty appear in the same dimension. The simplest measure of information content is the uncertainty we attach to the measurement. Statistical consistency requires that this uncertainty describes the distribution of the difference between the simulation of the optimized model and the measurements. For Gaussian distributions there are algorithms [41•] which adjust these uncertainties to achieve this. This, however, is the easy part of the problem. Much more difficult is the correlation between simulations of different measurements. In a periodic timeseries, such as hourly measurements of flux, a model is very likely to make the same mistake each day. This may well be true even after optimization. Thus there are consistent signals in the data which cannot be extracted by the model. Treating these measurements as independent pieces of data will both overestimate the failure of the model and overestimate the information content available from the measurements.

Approaches to this problem have usually involved either ''thinning'' the data or compositing it in some way. Ref. [31] chose a series of snapshots from their timeseries and optimized only these. Ref. [35] composited their timeseries into mean diurnal cycles and daily means. This is approximately subsetting the data by frequency rather than by time. More generally, in the context of their flux inversion, [16] projected measurements and model into frequency space. This took advantage of the fact that for stationary timeseries with multivariate Gaussian statistics, the errors in frequency space are independent.

Observation operators. Most terrestrial models exist as part of climate models and so their major focus is climate simulation. As noted in 'Applying assimilation to the carbon cycle', they must be coupled to observations via observation operators before use in assimilation mode. These introduce their own errors and often need extra variables in the assimilation that describe the observation process. Thus a detailed understanding of the process of making and simulating each observation is required for a meaningful assimilation.

Model deficiencies. The problems mentioned so far concern the conduct of MDF. Once these are addressed we will still see that our models do not meet the test of generality. Developers must then pay close attention to the core model. This process will be iterative, each time the model improves we will be able to extract more information from the measurements and our uncertainty specification must be revised.

Recommendations and conclusions

At the close of a rather negative appraisal of the status of carbon-cycle data assimilation, I will make some recommendations on improving the situation.

- MDF remains an essential tool for carbon-cycle research.
- The algorithmic details of MDF are not critical to its success.
- The task of data assimilation should move from the province of applications of a model into the development phase. As we have seen a common outcome of assimilation is a falsification of the model structure. This seems important information for model developers. The choice of parameterizations in a model is also amenable to assimilation via the addition of a discreet variable representing the choices. Applications to probabilistic climate forecasting are described in [42••] and references therein.
- Assimilation needs to respect the inherent scale in each measurement. This is computationally infeasible globally but it was shown in [43•] that flux and concentration measurements were mutually consistent with a high-resolution transport model. Thus this problem can be investigated in a regional assimilation.
- Much more attention should be paid to the information content retrievable from measurements. In particular assessments of the independence of residuals should form part of the assessment of quality of fit.
- More effort should be paid to the development of observation operators and their associated errors.

Acknowledgement

Part of the work presented in this paper was performed with the support of the European Space Agency Grant Number AO/1-4937/05/NL/HE.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Denman K, Brasseur G, Chidthaisong A, Ciais P, Cox PM, Dickinson RE, Hauglustaine D, Heinze C, Holland E, Jacob D, *et al.*: **Couplings between changes in the climate system and biogeochemistry**, In *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Edited by Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL. C.U.P., 2007.

2. Raupach MR, Canadell JG, Ciais P, Friedlingstein P, Rayner PJ **Quantifying and ranking vulnerabilities in the carbon-climate-human system**, *Tellus* 62B, in press.
Provides a 90% confidence interval on the additional warming from a range of processes. Its greatest contribution is this reduction to a common basis.

3. McGuire AD, Sitch S, Clein J, Dargaville R, Esser G, Foley J, Heimann M, Joos F, Kaplan J, Kicklighter DW *et al.*: **Carbon balance of the terrestrial biosphere in the twentieth century: analyses of CO_2, climate and land use effects with four process-based ecosystem models**. *Global Biogeochem Cycle* 2001, **15**:183-206.

Factorial experiments reveal the relative impact of various forcings on the terrestrial carbon-cycle

4. Bousquet P, Peylin P, Ciais P, Le Quéré C, Friedlingstein P, Tans PP: **Regional changes in carbon dioxide fluxes of land and oceans since**. *Science* 2000, **290**:1342-1346.
Previous inverse studies of interannual variability of the carbon-cycle suggested as much variability in the ocean as land, contradicting *in situ* measurements. The main contribution of this study was not its estimates themselves since the greater variability of land fluxs was forced by larger prior uncertainties on those fluxes. A detailed analysis of the atmospheric data itself, however, showed that sites with more influence from terrestrial sources were more variable.

5. Rayner PJ, Law RM, Allison CE, Francey RJ, Pickett-Heaps C: **The interannual variability of the global carbon cycle (1992–2005) inferred by inversion of atmospheric CO_2 and $\delta^{13}CO_2$ measurements**. *Global Biogeochem Cycle* 2008, **22**:GB3008 doi: 10.1029/2007GB003068.
An update of a previous inversion to more recent periods. Most controversially it suggested that the anomalous CO_2 growth-rate of 2003 was forced more by large fires in Siberia than the European heat-wave.

6. Friedlingstein P, Cox P, Betts R, Bopp L, von Bloh W, Brovkin V, Doney S, Eby M, Fung I, Govindasamy B *et al.*: **Climate-carbon cycle feedback analysis, results from the C4MIP model intercomparison**. *J Clim* 2006, **19**:3737-3753.
Presented a synthesis of results from the C4MIP project comparing the carbon-cycle climate feedback from a range of models.

7. Thornton PE, Doney SC, Lindsay K, Moore JK, Mahowald N, Randerson JT, Fung I, Lamarque JF, Feddema JJ, Lee YH: **Carbon–nitrogen interactions regulate climate-carbon cycle feedbacks: results from an atmosphere–ocean general circulation model**. *BioGeosciences* 2009, **6**:2099-2120.
This work demonstrates the importance of nitrogen feedbacks for carbon-cycle climate interactions. Although it is presented as showing nitrogen effects to suppress the carbon-cycle climate feedback, it is more noteworthy that Nitrogen limitation reduces future terrestrial uptake with or without climate change.

8. Peylin P, Bousquet P, Le Quéré C, Sitch S, Friedlingstein P, McKinley G, Gruber N, Rayner P, Ciais P: **Multiple constraints on regional CO_2 flux variations over land and oceans**. *Global Biogeochem Cycle* 2005, **19**:GB1011 doi: 10.1029/2003GB002214.
Investigates the drivers of interannual variability in the global carbon-cycle by considering a range of inverse (top-down) and mechanistic (bottom-up) studies.

9. Moorcroft PR: **How close are we to a predictive science of the biosphere?** *Trends Ecol Evol* 2006, **21**:400-407.
This study describes the state of the art in terrestrial biosphere modelling. Advocates the testing of the models against as much current data as possible.

10. • Raupach MR, Rayner PJ, Barrett DJ, DeFries R, Heimann M, Ojima DS, Quegan S, Schmullius C: **Model–data synthesis in terrestrial carbon observation: methods, data requirements and data uncertainty specifications**. *Global Change Biol* 2005, **11**:378-397.
This work provides an overview of model–data fusion for terrestrial models with particular guidance on specifying the observation error. Its primary importance is a concise summary of the different kinds of problems that can affect data of various types.

11. Evans SN, Stark P: **Inverse problems as statistics**. *Inverse Problems* 2002, **18**:55-97.
A highly mathematical treatment of the general inverse problem. Its main contribution is to unify the hitherto separate treatments within Applied Mathematics and Statistics

12. •• Tarantola A: *Inverse Problem Theory and Methods for Model Parameter Estimation*. SIAM; 2004.
(ISBN 0-89871-572-5). A rigorous and insightful introduction to inverse problems and their solution. It is relatively free from attachment to a particular field. A great textbook should change one's way of thinking as well as teach one the tools of the discipline. I find if one needs to go beyond the standard methods in any direction this book is my preferred starting-point.

13. Rodgers CD: *Inverse Methods for Atmospheric Sounding: Theory and Practice*World Scientific Publishing Co. Pte. Ltd.; 2000.
More readable than Tarantola's book it is also less elegant, less general and more targeted to applications in remote sensing.

14. Enting IG: *Inverse Problems in Atmospheric Constituent Transport*. Cambridge University Press; 2002.
A treatment of inverse problems with particular application to atmospheric chemistry and surface fluxes.

15. Bonnans JF, Gilbert JC, Lemaréchal JF, Sagastizébal CA: *Numerical Optimization, Theoretical and Practical Aspects*. New York: Springer-Verlag; 2003.
Text and reference on numerical optimisation.

16. Enting IG, Trudinger CM, Francey RJ: **A synthesis inversion of the concentration and δ^{13} C of atmospheric CO_2**. *Tellus* 1995, **47B**:35-52.
The paper that introduced Bayesian inverse methods to carbon-cycle studies.

17. Rayner PJ, Enting IG, Francey RJ, Langenfelds RL: **Reconstructing the recent carbon cycle from atmospheric CO_2, δ^{13} C and O_2/N_2 observations**. *Tellus* 1999, **51B**:213-232.
The first application of Bayesian methods to produce a space-time distribution of surface CO_2 fluxes.

18. Gurney KR, Law RM, Denning AS, Rayner PJ, Baker D, Bousquet P, Bruhwiler L, Chen Y-H, Ciais P, Fan S *et al.*: **Towards robust regional estimates of CO_2 sources and sinks using atmospheric transport models**. *Nature* 2002, **415**:626-630.
A rigorous intercomparison of inversions using different transport models in a simplified framework. It showed that while transport uncertainty was the most important contributor to uncertainty in the extratropics, tropical uncertainty is driven largely by lack of concentration observations.

19. Baker DF, Law RM, Gurney KR, Rayner P, Peylin P, Denning AS, Bousquet P, Bruhwiler L, Chen Y-H, Ciais P *et al.*: **Transcom 3 inversion intercomparison: Impact of transport model errors on the interannual variability of regional CO_2 fluxes, 1988–2003**. *Global Biogeochem Cycle* 2006, **20**:GB1002 doi: 10.1029/2004GB002439.
Extended the study of Gurney et al. to interannual variability. It showed that interannual variations suggested by the different inverse models were more similar than the long-term mean fluxes.

20. Peters W, Miller JB, Whitaker J, Denning AS, Hirsch A, Krol MC, Zupanski D, Bruhwiler L, Tans PP: **An ensemble data assimilation system to estimate CO_2 surface fluxes from atmospheric trace gas observations**. *J Geophys Res* 2005, **110**:24304 doi: 10.1029/2005JD006157.
Introduced an ensemble Kalman Filter technique to estimate surface fluxes.

21. Chevallier F, Ciais P, Conway TJ, Aalto T, Anderson BE, Bousquet P, Brunke EG, Ciattaglia L, Esaki Y, Frhlich M *et al:* **CO_2 surface uxes at grid point scale estimated from a global 21-year reanalysis1 of atmospheric measurements**, *J. Geophys*. res. 115, in press.
An application of an iterative inversion using as much in situ CO_2 data as possible

22. • Kaminski T, Knorr W, Rayner P, Heimann M: **Assimilating atmospheric data into a terrestrial biosphere model: A case study of the seasonal cycle**. *Global Biogeochem Cycle* 2002, **16**:1066 doi: 10.1029/2001GB001463.
Used a very simple model to demonstrate the assimilation of atmospheric concentration data into a biosphere model. Its most important lasting contribution is probably its discussion of the strengths and weaknesses of data assimilation compared to direct inversion.

23. Rayner PJ, Scholze M, Knorr W, Kaminski T, Giering R, Widmann H: **Two decades of terrestrial carbon fluxes from a carbon cycle data assimilation system (CCDAS)**. *Global Biogeochem Cycle* 2005, **19**:2026 doi: 10.1029/2004GB002254.
Assimilated concentration data over two decades into a more complete terrestrial carbon-cycle model. Found that interannual variability of net flux was driven more by changes in productivity than respiration.

24. Kaminski T, Rayner PJ, Heimann M, Enting IG: **On aggregation errors in atmospheric transport inversions**. *J Geophys Res* 2001, **106**:4703-4715.
Showed that the then common procedure of using large source regions when performing atmospheric inversions could introduce large errors.

25. • Trudinger C, Raupach M, Rayner P, Kattge J, Liu Q, Pak B, Reichstein M, Renzullo L, Richardson AD, Roxburgh S *et al.*: **The OptIC project: An intercomparison of optimisation techniques**

for parameter estimation in terrestrial biogeochemical models. *J Geophys Res* 2007, **112**:2027 doi: 10.1029/2006JG000367.
Compared a number of approaches to estimating parameters in the same forward model using the same data and uncertainties. Found no clear preference for a given method. Its methodological summary and bibliography also act as an important start to a literature review of the various methods.

26. Raupach MR: **Dynamics of resource production and utilisation in two-29 component biosphere–human and terrestrial carbon systems**. *Hydrol Earth Syst Sci* 2007, **11**:875-889.
Described in detail the dynamics of the model used in the previous paper.

27. Zimmerman DA, de Marsily G, Gotway CA, Marietta MG, Axness CL, Beauheim RL, Bras RL, Carrera J, Dagan G, Davies PB *et al.*: **A comparison of seven geostatistically based inverse approaches to estimate transmissivities for modeling advective transport by groundwater flow**. *Water Resour Res* 1998, **34**:1373-1414.
Although rather outside the scope of biogeochemistry, this paper also found that the method used for the assimilation was less important than many other details.

28. Wang YP, Leuning R, Cleugh H, Coppin PA: **Parameter estimation in surface exchange models using non-linear inversion: How many parameters can we estimate and which measurements are most useful?** *Global Change Biol* 2001, **7**:495-510.
Attempted formal parameter estimation for a biosphere model using eddy covariance measurements.

29. Barrett DJ: **Steady state turnover time of carbon in the Australian terrestrial biosphere**. *Global Biogeochem Cycle* 2002, **16**:1108 doi: 10.1029/2002GB001860.
Used a genetic algorithm to assimilate measurements of net primary production and soil carbon to estimate residence time of soil carbon in Australia

30. Braswell BH, Sacks WJ, Linder E, Schimel DS: **Estimating diurnal to annual ecosystem parameters by synthesis of a carbon flux model with eddy covariance net ecosystem exchange observations**. *Global Change Biol* 2005, **11**:335-355.
Optimized several parameters of a terrestrial carbon-cycle model against the Harvard Forest eddy covariance data.

31. Knorr W, Kattge J: **Inversion of terrestrial biosphere model parameter values against eddy covariance measurements using Monte Carlo sampling**. *Global Change Biol* 2005, **11**:1333-1351.
Assimilated half-hourly eddy covariance data from a forest and grassland site to constrain parameters in BETHY. They find they can reduce the uncertainty of uptake over a 2-year period after assimilating only 7 days of data.

32. Richardson AD, Hollinger DY: **Statistical modelling 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**. *Agric For Meteorol* 2005, **131**:191-208.
Fits simple models of respiration to half-hourly flux measurements from one site. A study of the errors in the fit suggests that the Gaussian assumption does not hold.

33. Williams M, Schwarz PA, Law BE, Irvine J, Kurpius MR: **An improved analysis of forest carbon dynamics using data assimilation**. *Global Change Biol* 2005, **11**:89-105.
Used an ensemble Kalman filter with a simple box model. It was able to assimilate several different types of data.

34. Demarty J, Chevallier F, Friend AD, Viovy N, Piao S, Ciais P: **Assimilation of global MODIS leaf area index retrievals within a terrestrial biosphere model**. *Geophys Res Lett* 2007, **34**:L15402 doi: 10.1029/2007GL030014.
Assimilated MODIS observations to correct the state or trajectory of the model rather than underlying parameters. This is an important alternative approach but less relevant to the construction of predictive models.

35. Santaren D, Peylin P, Viovy N, Ciais P: **Optimizing a process-based ecosystem model with eddy-covariance flux measurements: a pine forest in southern France**. *Global Biogeochem Cycle* 2007, **21**:GB2013 doi: 10.1029/2006GB002834.
Assimilated eddy covariance and radiation data from a forest site. They found that one year of observations was not enough to constrain parameters relevant to respiration even though they worked well for photosynthetic parameters.

36. Scholze M, Kaminski T, Rayner P, Knorr W, Geiring R: **Propagating uncertainty through prognostic CCDAS simulations**. *J Geophys Res* 2007, **112**:d17305 doi: 10.1029/2007JD008642.
This extended the atmospheric data assimilation study of [23]. Its most interesting innovation was to consider the uncertainty in predictions beyond the assimilation period

37. Wang YP, Baldocchi D, Leuning R, Falge E, Vesala T: **Estimating parameters in a land-surface model by applying nonlinear inversion to eddy covariance flux measurements from eight FLUXNET sites**. *Global Change Biol* 2007, **13**:652-670.
Optimized parameters separately at several flux sites. Obtained better fits when parameters were allowed to vary seasonally (as expected with more degrees of freedom) and found they could not fit sensible heat fluxes along with CO_2 and latent heat fluxes.

38. • Medvigy D, Wofsy SC, Munger JW, Hollinger DY, Moorcroft PR: **Mechanistic scaling of ecosystem function and dynamics in space and time: Ecosystem demography model version 2**. *J Geophys Res* 2009, **114**:G01002 doi: 10.1029/2008JG000812.
Fitted parameters using measurements from the Harvard Forest flux site then tested these parameters at two other sites. Its use of independent testing of the optimization is an important advance.

39. Assimilated FAPAR and flux measurements separately and jointly at several sites. Generally found it hard to reconcile the different measurements and model dynamics.

40. Raupach MR, Finnigan JJ: **Scale issues in boundary-layer meteorology: surface energy balances in heterogeneous terrain**. *Hydrol Process* 1995, **9**:589-612.
An important review of the problem of scaling in biospheric modelling.

41. • Michalak AM, Hirsch A, Bruhwiler L, Gurney KR, Peters W, Tans PP: **Maximum likelihood estimation of covariance parameters for Bayesian atmospheric trace gas surface flux inversions**. *J Geophys Res* 2005, **110**:D24107 doi: 10.1029/2005JD005970.
Provides an algorithm for optimizing variance parameters as well as best estimates in an inversion. The method is an important generalization of the classical inversion framework and points the way to other extensions.

42. •• Meinshausen M, Meinshausen N, Hare W, Raper SCB, Frieler K, Knutti R, Frame DJ, Allen MR: **Greenhouse gas emission targets for limiting global warming to 2°C**. *Nature* 2009, **458**:1158-1162 doi: 10.1038/nature00817.
Describes, especially in the supplementary material, some of the technicalities of assigning probabilities to climate forecasts based on current observations. Although developed for the particular case of climate forecasts the approach they use is equally applicable to any other data assimilation problem using ensembles of models. When we must deal explicitly with model imperfections this is an important problem to solve and this one possible approach.

43. • Lauvaux T, Gioli B, Rayner PJ, Chevallier F, Miglietta F, Ceschia E, Elbers JA, Hutjes R, Legain D, Uliasz M: **Bridging the gap between atmospheric concentrations and local ecosystem measurements**. *Geophys Res Lett* 2009, **36**:L19809 doi: 10.1029/2009GL039574.
Describes the testing of a regional atmospheric inversion by the use of aircraft measurements of flux. They find that the inversion improves the fit to the independent flux measurements relative to the prior. To my knowledge this is the first independent test of a direct CO_2 flux inversion and the first successful test of a regional inversion for this gas.