Data assimilation with carbon cycle models DRAFT

Ewan Pinnington

January 13, 2017

1 Data assimilation methods

Data assimilation provides techniques for combining observations and prior knowledge of a system in an optimal way to find a consistent solution referred to as the analysis. The prior knowledge of a system often takes the form of a numerical model and an initial guess of the model state/parameters. Many statistical methods have been developed for data assimilation. These methods can largely be categorised as either sequential or variational. Sequential algorithms solve the system of equations needed to find an optimal solution explicitly at each observation time. Variational methods solve the equations needed for an optimal solution implicitly by minimising a cost function for all available observations over some time window. This thesis is mainly concerned with the variational technique of four-dimensional variational data assimilation (4D-Var). In numerical weather prediction data assimilation has been predominately used for state estimation whilst keeping parameters fixed. This is because numerical weather prediction is mainly dependent on the initial state with model physics being well understood. Ecosystem carbon cycle models are more dependent on finding the correct set of parameters to describe the ecosystem of interest [Luo et al., 2015]. We therefore discuss data assimilation for joint state and parameter estimation. In the next sections we give a general introduction to data assimilation, then expand this to 4D-Var and finally we briefly discuss other data assimilation methods not directly used in this thesis but applicable to subsequent discussion.

1.1 Introduction to data assimilation

We consider a system that can be described by a numerical model with a true model state $\mathbf{z}^t \in \mathbb{R}^n$ and true parameters $\mathbf{p}^t \in \mathbb{R}^q$. We then define the true augmented state as

$$\mathbf{x}^t = \begin{pmatrix} \mathbf{p}^t \\ \mathbf{z}^t \end{pmatrix} \in \mathbb{R}^{q+n}. \tag{1}$$

The initial guess to this model augmented state $\mathbf{x}^b \in \mathbb{R}^{q+n}$ (often referred to as the prior or background) and observations of the system $\mathbf{y} \in \mathbb{R}^m$ will only be approximations to the true system state, such that

$$\mathbf{x}^b = \mathbf{x}^t + \boldsymbol{\epsilon}^b, \tag{2}$$

$$\mathbf{y} = h(\mathbf{x}^t) + \boldsymbol{\epsilon}^o, \tag{3}$$

where $\boldsymbol{\epsilon}^b$ and $\boldsymbol{\epsilon}^o$ are the prior and observation errors respectively, and $h : \mathbb{R}^{q+n} \rightarrow \mathbb{R}^m$ is the observation operator (can be linear or non-linear) mapping the augmented state to the observations. The errors in the prior and observations are assumed to be unbiased and mutually independent with known covariance matrices $\mathbf{B} = \mathbb{E}[\boldsymbol{\epsilon}^b(\boldsymbol{\epsilon}^b)^T]$ and $\mathbf{R} = \mathbb{E}[\boldsymbol{\epsilon}^o(\boldsymbol{\epsilon}^o)^T]$.

The best estimate to $\mathbf{x}^t$ satisfying both equation (2) and (3) is often called the analysis or the maximum a posteriori estimate, here denoted $\mathbf{x}^a$. It is possible to derive this analysis by applying Bayesian methods to probability density functions. Bayes' theorem, first discussed in Bayes and Price [1763] but formalised by Laplace [1781], states that the posterior probability of event A given that event B occurs, is proportional to the prior probability of A multiplied by the probability of event B given that event A occurs, this can be expressed mathematically as

$$\mathbb{P}(A|B) \propto \mathbb{P}(A)\mathbb{P}(B|A). \tag{4}$$

For data assimilation event A represents the augmented state of the system $\mathbf{x}$ and event B the observations $\mathbf{y}$. Maximising the probability $\mathbb{P}(A|B)$ is then equivalent to finding the augmented state that best represents the observations.

If we make the assumption of Gaussian probability density functions (pdf) with

$$\mathbb{P}^b(\mathbf{x}) = \frac{1}{\sqrt{|2\pi\mathbf{B}|}}\exp\big(-\frac{1}{2}(\mathbf{x}-\mathbf{x}^b)^T\mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}^b)\big) \tag{5}$$

and

$$\mathbb{P}^o(\mathbf{y}|\mathbf{x}) = \frac{1}{\sqrt{|2\pi\mathbf{R}|}}\exp\big(-\frac{1}{2}(\mathbf{y}-h(\mathbf{x}))^T\mathbf{R}^{-1}(\mathbf{y}-h(\mathbf{x}))\big), \tag{6}$$

where $\mathbb{P}^b(\mathbf{x})$ is the pdf for the prior and $\mathbb{P}^o(\mathbf{y}|\mathbf{x})$ the pdf of the observations given the augmented state. Then from Bayes' theorem (equation (4)) the posterior pdf for the augmented state

$$\begin{aligned}\mathbb{P}^a(\mathbf{x}|\mathbf{y}) &\propto \frac{1}{\sqrt{|2\pi\mathbf{B}|}\sqrt{|2\pi\mathbf{R}|}}\exp\big(-\frac{1}{2}(\mathbf{x}-\mathbf{x}^b)^T\mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}^b)-\frac{1}{2}(\mathbf{y}-h(\mathbf{x}))^T\mathbf{R}^{-1}(\mathbf{y}-h(\mathbf{x}))\big)\\ &\propto \exp\big(-\frac{1}{2}(\mathbf{x}-\mathbf{x}^b)^T\mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}^b)-\frac{1}{2}(\mathbf{y}-h(\mathbf{x}))^T\mathbf{R}^{-1}(\mathbf{y}-h(\mathbf{x}))\big),\end{aligned} \tag{7}$$

here we can ignore the constant multiplying the exponential function as it is independent of $\mathbf{x}$. We want to maximise the probability of the augmented state $\mathbf{x}$ given the observations $\mathbf{y}$, from equation (7) we can see that to maximise $\mathbb{P}^a(\mathbf{x}|\mathbf{y})$ we must maximise the terms in the exponent, this is equivalent to minimising the quadratic cost function

$$J(\mathbf{x}) = \frac{1}{2}(\mathbf{x}-\mathbf{x}^b)^T\mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}^b)+\frac{1}{2}(\mathbf{y}-h(\mathbf{x}))^T\mathbf{R}^{-1}(\mathbf{y}-h(\mathbf{x})). \tag{8}$$

This is the cost function minimised in three-dimensional variational data assimilation (3D-Var), where the minimum is found using a descent algorithm evaluating equation (8) and its gradient [Courtier et al., 1998]. We can approximate the minimum of (8) by finding its gradient and setting it to zero to obtain the best linear unbiased estimate (BLUE) [Talagrand, 1997] where

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{K}(\mathbf{y} - h(\mathbf{x}^b)), \tag{9}$$

$$\mathbf{K} = \mathbf{B}\mathbf{H}^T(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1}, \tag{10}$$

where $\mathbf{K}$ is the Kalman gain matrix specifying the weight of the analysis increment and $\mathbf{H} = \frac{\partial h(\mathbf{x})}{\partial \mathbf{x}}$ is the linearised observation operator. We can also approximate the analysis error covariance matrix as

$$\mathbf{A} = (\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} + \mathbf{B}^{-1})^{-1}, \tag{11}$$

if h is linear then (9) and (11) are exact solutions.

1.2 4D-Var

4D-Var extends 3D-Var to allow for the assimilation of observations distributed throughout some time interval t_0 to t_N. Sasaki [1970] proposed a method for combining a time series of observations with a numerical model, which was then further developed for use in numerical weather prediction [Dimet and Talagrand, 1986]. In 4D-Var we minimise the cost function,

$$J(\mathbf{x}_0) = \frac{1}{2}(\mathbf{x}_0 - \mathbf{x}^b)^T\mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}^b) + \frac{1}{2}\sum_{i=0}^{N}(\mathbf{y}_i - \mathbf{h}_i(\mathbf{x}_i))^T\mathbf{R}_i^{-1}(\mathbf{y}_i - \mathbf{h}_i(\mathbf{x}_i)), \tag{12}$$

to obtain the analysis $\mathbf{x}_0^a$, valid at the initial time t_0, subject to the strong constraint that the model states $(\mathbf{x}_0, \ldots, \mathbf{x}_N)$ must satisfy the model equations,

$$\mathbf{x}_i = \mathbf{m}_{i-1\rightarrow i}(\mathbf{x}_{i-1}), \tag{13}$$

where $\mathbf{x}_i$ is the model augmented state at time t_i, $\mathbf{m}_{i-1\rightarrow i}$ is the possibly nonlinear augmented system model evolving $\mathbf{x}_{i-1}$ from time t_{i-1} to time t_i, $\mathbf{y}_i$ is the vector of observations at time t_i, and h_i is the observation operator at time t_i. We can rewrite equation (12) to avoid the sum notation as

$$J(\mathbf{x}_0) = \frac{1}{2}(\mathbf{x}_0 - \mathbf{x}^b)^T\mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}^b) + \frac{1}{2}(\hat{\mathbf{y}} - \hat{\mathbf{h}}(\mathbf{x}_0))^T\hat{\mathbf{R}}^{-1}(\hat{\mathbf{y}} - \hat{\mathbf{h}}(\mathbf{x}_0)) \tag{14}$$

where,

$$\hat{\mathbf{y}} = \begin{pmatrix} \mathbf{y}_0 \\ \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_N \end{pmatrix}, \ \hat{\mathbf{h}}(\mathbf{x}_0) = \begin{pmatrix} \mathbf{h}_0(\mathbf{x}_0) \\ \mathbf{h}_1(\mathbf{m}_{0\to 1}(\mathbf{x}_0)) \\ \vdots \\ \mathbf{h}_N(\mathbf{m}_{0\to N}(\mathbf{x}_0)) \end{pmatrix}, \text{ and } \ \hat{\mathbf{R}} = \begin{pmatrix} \mathbf{R}_{0,0} & \mathbf{R}_{0,1} & \dots & \mathbf{R}_{0,N} \\ \mathbf{R}_{1,0} & \mathbf{R}_{1,1} & \dots & \mathbf{R}_{1,N} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{R}_{N,0} & \mathbf{R}_{N,1} & \dots & \mathbf{R}_{N,N} \end{pmatrix}. \tag{15}$$

For 4D-Var we approximate the analysis error covariance matrix as

$$\mathbf{A} = (\hat{\mathbf{H}}^T\hat{\mathbf{R}}^{-1}\hat{\mathbf{H}} + \mathbf{B}^{-1})^{-1}, \tag{16}$$

where $\hat{\mathbf{H}}$ is that observability matrix given by

$$\hat{\mathbf{H}} = \begin{pmatrix} \mathbf{H}_0 \\ \mathbf{H}_1\mathbf{M}_0 \\ \vdots \\ \mathbf{H}_N\mathbf{M}_{N,0} \end{pmatrix} \tag{17}$$

with $\mathbf{H}_i = \frac{\partial \mathbf{h}_i(\mathbf{x}_i)}{\partial \mathbf{x}_i}$ the linearised observation operator and $\mathbf{M}_{i,0} = \mathbf{M}_{i-1}\mathbf{M}_{i-2}\cdots\mathbf{M}_0$ the tangent linear model with $\mathbf{M}_i = \frac{\partial \mathbf{m}_{i-1\to i}(\mathbf{x}_i)}{\partial \mathbf{x}_i}$. The tangent linear model can be difficult to implement, however using techniques such as automatic differentiation [Renaud, 1997] can reduce the time taken to implement the derivative of a model.

1.3 Sequential and Markov chain Monte Carlo approaches

Markov chain Monte Carlo (MCMC) methods refer to a suite of related algorithms (Metropolis-Hastings, simulated annealing and Gibbs sampling), with the first MCMC method being the Metropolis algorithm [Metropolis et al., 1953]. These methods sample a cost function measuring the model-data miss-match, usually similar to the negation of the second term in the 4D-Var cost function shown in equation (14). As these methods use the negation of the cost function in equation (14) they seek to find a global optimum for this cost function rather than a minimum. This is achieved by iteratively sampling the cost function, with each iteration of the parameter and state values being uniquely determined by the previously sampled parameter and state values. The output of the MCMC methods is a set of accepted parameter and state values from which analysis or posterior error covariances can be calculated. These methods are easy to implement and do not require the derivative of the model code. However, they come with high computational cost as they often require in the order of 10^6 model evaluations [Zobitz et al., 2011], meaning these methods become infeasible for global implementations of more complex models.

Whereas variational and MCMC techniques assimilate all available observations over some time window at once, sequential algorithms update the model trajectory at each observation time. These algorithms approximate the BLUE formula in equation (9) to update the model parameter and state values whenever an observation is available. This means that parameter values can change over time and state and parameter analysis trajectories will become discontinuous (unless using a sequential 'smoother' method). The first sequential method was the Kalman Filter (KF) [Kalman, 1960]. The KF method requires the evolution of the error covariance matrix $\mathbf{B}$ through the time window as observations are assimilated. This becomes infeasible for large systems. The Ensemble Kalman Filter (EnKF) [Evensen, 2003] was therefore developed to address this problem, the error covariance matrix for the state/parameters is approximated using an ensemble of state/parameter vectors therefore the evolution of the error covariance matrix $\mathbf{B}$ is avoided. These methods are also easy to implement, however, dependent on the complexity of the model, the ensemble size can be limited by computational cost, meaning that covariances can be subject to noise and techniques have to be employed to avoid this. The ensemble can also collapse on the same value after assimilating a number of observations, this can be avoided by adding random noise to the system [REF look at Sarah's DA notes].

2 Applications to the carbon cycle

DA for NWP is considered a state estimation problem as the physics of the problem are well understood and therefore parameterisations should not change over time. For the C cycle DA is more of a joint parameter and state estimation problem with the vast majority of studies using DA to estimate both parameter and state variables for a given system. Parameters governing land surface C uptake can change over time with developing forest and disturbance events.

2.1 Site-level applications

Two of the first examples of combining site-level eddy covariance data with models of ecosystem carbon balance were using the Data Assimilation Linked Ecosystem Carbon (DALEC) and SImplified PhotosyNthesis and EvapoTranspiration (SIPNET) models in Williams et al. [2005] and Braswell et al. [2005] respectively. These are both simple process based models of ecosystem carbon dynamics. In Braswell et al. [2005] MCMC techniques (based on the Metropolis algorithm) are used to combine half-daily observations of NEE with the SIPNET model. The DA technique is used to estimate initial model parameter and state values as well as the standard deviation in NEE flux observation (found to be approximately 1 g C m^{-2}). It is shown that NEE has limited ability in constraining some model parameters as the model prediction of NEE is insensitive to these parameters at the time-scales shown in the study (10 years). Williams et al. [2005] assimilate a more diverse set of daily carbon flux and stock observations with the DALEC model at the Metolius ponderosa pine site in Oregon, USA. In this study an EnKF is nested within a quasi-Newton optimisation scheme to find the initial set of parameter and state values that require least correction by the EnKF. As the aim of this study was to estimate the initial state and parameter values of the model it would perhaps be more logical to use a variational or MCMC technique. Williams et al. [2005] found large reductions in model prediction error after assimilation, noting that rare measurements of carbon stocks have limited impact on assimilation results but suggesting that longer time-series of these stock measurements will be important to constrain carbon pool turnover rates. Assimilate modelled GPP REWORD? suggest analogous to satellite derived GPP and that this could help improve results if used in future studies.

As data assimilation became more widespread with models and observations of ecosystem carbon dynamics Trudinger et al. [2007] conducted the Optimisation InterComparison (OptIC) project to better understand the benefits and issues of different DA implementations. In this study participant researchers used a variety of distinct DA implementations to estimate the parameters of a highly simplified model of terrestrial carbon balance. No single DA method was found to perform better than others and the representation of the cost function was shown to be more important than the method. In different optimisation experiments the representation of error added to pseudo observations was varied (Gaussian, lognormal, temporally correlated distributions, etc.). It was stated that the main criterion for success was accurate specification of errors. In particular, none of the participant researchers made an effort to account for temporally correlated error and this resulted in biased results. The REgional Flux Estimation eXperiment (REFLEX) was a similar study conducted using the DALEC model by Fox et al. [2009], here 9 participants were asked to combine both synthetic and observed NEE and LAI data with the DALEC model. Again a variety of DA methods were used (although no variational methods present), with no DA technique performing consistently better than others. Across all methods parameters linked directly to GPP and TER were best constrained, with parameters linked to slower processes (allocation and turnover of fine root and wood carbon pools) being poorly constrained. Fox et al. [2009] suggest that observations of slow large carbon pools would add useful constraint to DA schemes and compliment eddy covariance data, it is also discussed that future studies should investigate the importance of prior error estimates. Currently the representation of prior and observational errors are still very basic in the majority of DA schemes for ecosystem carbon balance. [[Dietze et al., 2013] also stress the need to improve the representation of unceraincy in DA schemes. As data assimilation with ecological applications becomes more prevalent it is important that tools for information management and data assimilation are made more accessible. The Predictive Ecosystem Analyser (PEcAn) is an effort to achieve this, this system also allows for easier comparison of different implemented models [Dietze et al., 2013] improving the standard and reproducibility of experimental results is important. (Make this sit nicely in this paragraph)]

Satellite observations of reflectance have also been used with these simple models to assess their impact on modelled estimates. In Quaife et al. [2008] used earth observation data from the MODIS instrument on NASA's TERRA and AQUA satellites in an EnKF with the DALEC model at the Metolius forest in Oregon, USA. They found that, after assimilation of MODIS data, modelled LAI was over-predicted when compared to site-level estimates. Over-prediction of LAI lead to an over-estimate in both GPP and TER. Despite this modelled NEE is improved after assimilation when compared to site flux tower observations and significant reductions in modelled flux uncertainties are achieved. Satellite data has also been used with the SIPNET model, in Zobitz et al. [2014] earth observation data is assimilated on different time-scales with flux tower NEE. Through a combination of assimilation studies and use of the Bayesian information criterion [Schwarz et al., 1978] measure of information content they show that the best combination of observations is remotely sensed annually averaged fraction of absorbed photosynthetically active radiation with twice-daily observations of NEE.

The ecosystem carbon models of SIPNET and DALEC have both been used in many other experiments combining a variety of observations relevant to the carbon balance of terrestrial ecosystems [Keenan et al., 2011, Moore et al., 2008, Sacks et al., 2007, Zobitz et al., 2008]. One problem facing studies working with NEE flux observations alongside other ancillary site-level data is the overweighting of NEE flux data in the assimilation as in general other site-level measurements are made at longer time-scales so that the number of NEE flux observations in the any given assimilation can outnumber other available observations by anywhere from a factor of 10 to a factor of 1000 (dependent on the time-step of the model). In order to reduce the problem of overweighting flux observations Richardson et al. [2010] used a cost function taking the product of the observation-model miss-matches, rather than the sum, to give an absolute rather than relative measure of the model fit to observations. This study used MCMC techniques to combine a diverse set of observations from the Howland forest flux site in Maine, USA with the DALEC model. They found in particular that woody biomass accumulation increment provided an orthogonal constraint to NEE data and reduced uncertainties in parameter estimates. In Keenan et al. [2012] the problem of overweighting NEE in assimilation results is addressed by calculating the model-observation mismatch and then dividing it by the number of data points for each distinct data stream. This problem could also be treated by better specifying the observation error covariance matrix in the DA scheme. Keenan et al. [2012] work with MCMC techniques and the Forest Biomass, Assimilation, Allocation and Respiration (FöBAAR) model . This study discusses the impact of complimentary datasets in addition to NEE, with Keenan et al. [2013] further investigating the information content in observations using a set of data denial experiments at the Harvard Forest in Massachusetts, USA. They found that data relating to the turnover of carbon pools provides the most information when combined with observations of NEE. This study uses true observations, it is important to develop new twin experiments and other methods to understand the impact of possible observations and allow for more consideration when planning a measurement campaign REWORD.

As ecosystem carbon cycle DA is predominantly a parameter estimation problem equifinality is an ever-present issue, with available data often not being able to constrain all of the optimised model parameters. Wu et al. [2009] found that only 6 out of 16 model parameters were identifiable using a conventional MCMC technique to assimilate observations of NEE with a flux-based ecosystem model. In Bloom and Williams [2015] a set of ecological "common sense" dynamical constraints are implemented in a MCMC DA scheme, these are constraints on things such as carbon pool turnover rates and parameter inequalities. These additional constraints act to ensure the retrieved parameter and state values from DA are physically reasonable. Another option for reducing the problem of equifinality would be to better specify the background and observation error covariance matrices so that there is more constraint on data assimilation results. This would be particularly true for the background error covariance matrix where off-diagonal elements would act to enforce balances between different parameter/state values. It is also important that we continue to produce new distinct sets of observations in order to reduce equifinality further and better understand where model structure can be improved [Carvalhais et al., 2010].

2.2 Global implementations

At a similar time to site-level DA implementations with flux tower records, observations of atmospheric CO_2 concentration were being used with atmospheric transport models and variational DA methods to perform global inversions and estimate parameters relating to land surface carbon dynamics. An example of this is in Rayner et al. [2005] where 4D-Var is implemented with the Biosphere Energy Transport HYdrology (BETHY) model [Knorr and Heimann, 2001] in a Carbon Cycle Data Assimilation System (CCDAS) to assimilate both satellite observations and atmospheric CO_2 concentrations in a stepwise manner on a global scale. It has been shown that, if possible, it is beneficial to assimilate all data streams concurrently rather than in series [MacBean et al., 2016], but this may not be practical in some scenarios. In the CCDAS automatic differentiation is used to find the Jacobian and Hessian of the cost function, the inverse Hessian of the cost function is then used to find an estimate to posterior parameter errors [Rayner et al., 2005], it is found that uncertainty in long-term soil carbon storage is the largest contributor to uncertainty in net CO_2 flux. Scholze et al. [2007] show how this estimate to posterior parameter uncertainties from the cost function Hessian can be propagated through time for future modelled predictions. A review of the CCDAS implementation can be found in Kaminski et al. [2013].

The ORganising Carbon and Hydrology In Dynamic Ecosystems Environment (ORCHIDEE) model [Krinner et al., 2005] is a dynamic global vegetation model that has been used in many data assimilation experiments. ORCHIDEE has been used with both sequential [Demarty et al., 2007] and variational methods [Bacour et al., 2015]. The 4D-Var data assimilation routine for ORCHIDEE outlined in Kuppel et al. [2012] also uses automatic differentiation to find the adjoint of the ORCHIDEE model used in the calculation of the derivative of the cost function. An adjoint has also recently been developed for the Joint UK Land Environment Simulator (JULES) model to allow for the implementation of variational data assimilation [Raoult et al., 2016]. Variational techniques have been preferred in these large scale applications due to computational efficiency, with automatic differentiation techniques reducing the time it takes to implement the adjoint of a model. Current variational methods have made the approximation of diagonal background and observation error covariance matrices.

Although in these global implementations variational methods have been prevalent due to computational efficiency, in Bloom et al. [2016] MCMC techniques (with prior constraints from Bloom and Williams [2015]) are used to find a global a $1^o \times 1^o$ DALEC2 map. Using MCMC techniques is possible as DALEC2 is a simple model which requires little computational cost to run. In this study MODIS LAI and soil carbon observations from the harmonised world soil database are assimilated. Using the ecological dynamical constraints from Bloom and Williams [2015] in this global implementation could be an issue as not all ecosystems will adhere to these constraints (especially if subjected to severe disturbances such as fire or insect outbreak). Bloom et al. [2016] use the retrieved global DALEC2 map to gain insight into ecosystem functioning and suggest that conventional land cover maps cannot adequately describe the spatial variability of carbon states and processes. The results from this study could be used as a set of prior model estimates for variational methods which may prove more feasible in the long term.

2.3 Issues faced in carbon cycle data assimilation

There are many opportunities for further development in the field of carbon cycle data assimilation. Here we discuss three major issues (REWORD?):

- Equifinality: Many different combinations of parameters and state values able to recreate assimilated observations. As discussed, data assimilation for the carbon cycle is both a parameter and state estimation problem with available data not not allowing for all parameters to be identifiable [Luo et al., 2009]. The majority of observations in many experiments are NEE flux measurements, these measurements represent the difference between two large fluxes (GPP and TER), therefore both GPP and TER can be grossly misspecified by a model but still achieve the observed NEE, contributing to the problem of equifinality. It is important that new methods and observations are produced to reduce this issue.

- Understanding the Information content in current and potential observations: In order to reduce the problem of equifinality it is important to combine as many distinct data streams as possible, it is of great importance that we understand the information content in potential new data streams so that we can focus efforts on campaigns that will add the most information possible to DA schemes. In particular understanding what measurements best compliment eddy covariance data Rayner [2010].

- Representation of prior and observational errors: Current DA schemes take a very simple approach to defining errors and many studies reviewed here comment on the need to better characterise uncertainties. Improving the representation of prior errors in DA schemes will also help reduce the problem of equifinality by adding extra constraint and imposing balances on assimilation results. It is important that more efforts are made to fully characterise all sources of uncertainty [Keenan et al., 2011, Raupach et al., 2005]. Dietze et al. [2013] comment that tools for information management and DA need to be more accessible and reproducible, which could also lead to the improved characterisation of uncertainties.

3 Conclusion

Many efforts and much progress is being made in the field of carbon cycle DA. Currently there are areas that need addressing; the specification of errors, the information content in available and possible new data streams and continued application of DA to new problems involving the carbon cycle are all important areas for progress. In this thesis we choose to work with the 4D-Var method of data assimilation as this allows us to use measures of information content that require the derivative of the model code, allows us to specify different covariance structures in both the background and observation error covariance matrices and is also scalable to large scale DA implementations for the carbon cycle.

References

C. Bacour, P. Peylin, N. MacBean, P. J. Rayner, F. Delage, F. Chevallier, M. Weiss, J. Demarty, D. Santaren, F. Baret, D. Berveiller, E. Dufrêne, and P. Prunet. Joint assimilation of eddy-covariance flux measurements and FAPAR products over temperate forests within a process-oriented biosphere model. *Journal of Geophysical Research: Biogeosciences*, pages n/a–n/a, 2015. ISSN 21698953. doi: 10.1002/2015JG002966. URL `http://doi.wiley.com/10.1002/2015JG002966`.

T. Bayes and R. Price. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions (1683-1775)*, 53:370–418, 1763. ISSN 02607085.

A. A. Bloom and M. Williams. Constraining ecosystem carbon dynamics in a data-limited world: integrating ecological "common sense" in a modeldata fusion framework. *Biogeosciences*, 12(5):1299–1315, 2015. ISSN 1726-4189. doi: 10.5194/bg-12-1299-2015. URL `http://www.biogeosciences.net/12/1299/2015/`.

A. A. Bloom, J.-F. Exbrayat, I. R. van der Velde, L. Feng, and M. Williams. The decadal state of the terrestrial carbon cycle: Global retrievals of terrestrial carbon allocation, pools, and residence times. *Proceedings of the National Academy of Sciences*, 113(5):1285–1290, 2016.

B. H. Braswell, W. J. Sacks, E. Linder, and D. S. Schimel. Estimating diurnal to annual ecosystem parameters by synthesis of a carbon flux model with eddy covariance net ecosystem exchange observations. *Global Change Biology*, 11(2):335–355, 2005.

N. Carvalhais, M. Reichstein, P. Ciais, G. J. Collatz, M. D. Mahecha, L. Montagnani, D. Papale, S. Rambal, and J. Seixas. Identification of vegetation and soil carbon pools out of equilibrium in a process model via eddy covariance and biometric constraints. *Global Change Biology*, 16(10):2813–2829, 2010. ISSN 1365-2486. doi: 10.1111/j.1365-2486.2010.02173.x. URL `http://dx.doi.org/10.1111/j.1365-2486.2010.02173.x`.

P. Courtier, E. Andersson, W. Heckley, D. Vasiljevic, M. Hamrud, A. Hollingsworth, F. Rabier, M. Fisher, and J. Pailleux. The ecmwf implementation of three-dimensional variational assimilation (3d-var). i: Formulation. *Quarterly Journal of the Royal Meteorological Society*, 124(550):1783–1807, 1998.

J. Demarty, F. Chevallier, A. D. Friend, N. Viovy, S. Piao, and P. Ciais. Assimilation of global modis leaf area index retrievals within a terrestrial biosphere model. *Geophysical*

Research Letters, 34(15):n/a–n/a, 2007. ISSN 1944-8007. doi: 10.1029/2007GL030014. L15402.

M. C. Dietze, D. S. Lebauer, and R. Kooper. On improving the communication between models and data. *Plant, Cell & Environment*, 36(9):1575–1585, 2013. ISSN 1365-3040. doi: 10.1111/pce.12043. URL `http://dx.doi.org/10.1111/pce.12043`.

F.-X. l. Dimet and O. Talagrand. Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. *Tellus A*, 38A(2):97–110, 3 1986. ISSN 1600-0870. doi: 10.1111/j.1600-0870.1986.tb00459.x.

G. Evensen. The ensemble kalman filter: theoretical formulation and practical implementation. *Ocean Dynamics*, 53(4):343–367, 2003. ISSN 1616-7228. doi: 10.1007/s10236-003-0036-9.

A. Fox, M. Williams, A. D. Richardson, D. Cameron, J. H. Gove, T. Quaife, D. Ricciuto, M. Reichstein, E. Tomelleri, C. M. Trudinger, et al. 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(10):1597–1615, 2009.

R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 03 1960.

T. Kaminski, W. Knorr, G. Schürmann, M. Scholze, P. J. Rayner, S. Zaehle, S. Blessing, W. Dorigo, V. Gayler, R. Giering, N. Gobron, J. P. Grant, M. Heimann, a. Hooker-Stroud, S. Houweling, T. Kato, J. Kattge, D. Kelley, S. Kemp, E. N. Koffi, C. Köstler, P. P. Mathieu, B. Pinty, C. H. Reick, C. Rödenbeck, R. Schnur, K. Scipal, C. Sebald, T. Stacke, a. T. Van Scheltinga, M. Vossbeck, H. Widmann, and T. Ziehn. The BETHY/JSBACH Carbon Cycle Data Assimilation System: Experiences and challenges. *Journal of Geophysical Research: Biogeosciences*, 118(4):1414–1426, 2013. ISSN 21698961. doi: 10.1002/jgrg.20118.

T. F. Keenan, M. S. Carbone, M. Reichstein, and A. D. Richardson. The model–data fusion pitfall: assuming certainty in an uncertain world. *Oecologia*, 167(3):587, 2011. ISSN 1432-1939. doi: 10.1007/s00442-011-2106-x.

T. F. Keenan, E. Davidson, A. M. Moffat, W. Munger, and A. D. Richardson. Using model-data fusion to interpret past trends, and quantify uncertainties in future projections, of terrestrial ecosystem carbon cycling. *Global Change Biology*, 18(8):2555–2569, 2012. ISSN 1365-2486. doi: 10.1111/j.1365-2486.2012.02684.x.

T. F. Keenan, E. A. Davidson, J. W. Munger, and A. D. Richardson. Rate my data: quantifying the value of ecological data for the development of models of the terrestrial carbon cycle. *Ecological Applications*, 23(1):273–286, 2013. ISSN 1939-5582. doi: 10.1890/12-0747.1.

W. Knorr and M. Heimann. Uncertainties in global terrestrial biosphere modeling: 1. a comprehensive sensitivity analysis with a new photosynthesis and energy balance scheme. *Global Biogeochemical Cycles*, 15(1):207–225, 2001.

G. Krinner, N. Viovy, N. de Noblet-Ducoudré, J. Ogée, J. Polcher, P. Friedlingstein, P. Ciais, S. Sitch, and I. C. Prentice. A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system. *Global Biogeochemical Cycles*, 19(1):1–33, 2005. ISSN 08866236. doi: 10.1029/2003GB002199.

S. Kuppel, P. Peylin, F. Chevallier, C. Bacour, F. Maignan, and A. D. Richardson. Constraining a global ecosystem model with multi-site eddy-covariance data. *Biogeosciences*, 9:3757–3776, 2012. doi: 10.5194/bg-9-3757-2012.

P. d. Laplace. Mémoire sur les probabilités. *Mémoires de l'Académie Royale des sciences de Paris*, 1778:227–332, 1781.

Y. Luo, E. Weng, X. Wu, C. Gao, X. Zhou, and L. Zhang. Parameter identifiability, constraint, and equifinality in data assimilation with ecosystem models. *Ecological Applications*, 19(3):571–574, 2009. ISSN 1939-5582. doi: 10.1890/08-0561.1.

Y. Luo, T. F. Keenan, and M. Smith. Predictability of the terrestrial carbon cycle. *Global change biology*, 21(5):1737–1751, 2015.

N. MacBean, P. Peylin, F. Chevallier, M. Scholze, and G. Schürmann. Consistent assimilation of multiple data streams in a carbon cycle data assimilation system. *Geoscientific Model Development*, 9(10):3569, 2016.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6): 1087–1092, 1953.

D. J. Moore, J. Hu, W. J. Sacks, D. S. Schimel, and R. K. Monson. Estimating transpiration and the sensitivity of carbon uptake to water availability in a subalpine forest using a simple ecosystem process model informed by measured net {CO_2} and {H_2O} fluxes. *Agricultural and Forest Meteorology*, 148(10):1467 – 1477, 2008. ISSN 0168-1923. doi: http://dx.doi.org/10.1016/j.agrformet.2008.04.013.

T. Quaife, P. Lewis, M. De Kauwe, M. Williams, B. E. Law, M. Disney, and P. Bowyer. Assimilating canopy reflectance data into an ecosystem model with an Ensemble Kalman Filter. *Remote Sensing of Environment*, 112(4):1347–1364, 2008. ISSN 00344257. doi: 10.1016/j.rse.2007.05.020.

N. M. Raoult, T. E. Jupp, P. M. Cox, and C. M. Luke. Land-surface parameter optimisation using data assimilation techniques: the adjules system v1. 0. *Geoscientific Model Development*, 9(8):2833, 2016.

M. Raupach, P. Rayner, D. Barrett, R. DeFries, M. Heimann, D. Ojima, S. Quegan, and C. Schmullius. Model–data synthesis in terrestrial carbon observation: methods, data requirements and data uncertainty specifications. *Global Change Biology*, 11(3):378–397, 2005.

P. Rayner, M. Scholze, W. Knorr, T. Kaminski, R. Giering, and H. Widmann. Two decades of terrestrial carbon fluxes from a carbon cycle data assimilation system (ccdas). *Global Biogeochemical Cycles*, 19(2), 2005.

P. J. Rayner. The current state of carbon-cycle data assimilation. *Current Opinion in Environmental Sustainability*, 2(4):289–296, 2010.

J. Renaud. Automatic differentiation in robust optimization. *AIAA journal*, 35(6):1072–1079, 1997.

A. D. Richardson, M. Williams, D. Y. Hollinger, D. J. Moore, D. B. Dail, E. A. Davidson, N. A. Scott, R. S. Evans, H. Hughes, J. T. Lee, et al. Estimating parameters of a forest ecosystem c model with measurements of stocks and fluxes as joint constraints. *Oecologia*, 164(1):25–40, 2010.

W. J. Sacks, D. S. Schimel, and R. K. Monson. Coupling between carbon cycling and climate in a high-elevation, subalpine forest: a model-data fusion analysis. *Oecologia*, 151 (1):54–68, 2007. ISSN 1432-1939. doi: 10.1007/s00442-006-0565-2.

Y. Sasaki. Some basic formalisms in numerical variational analysis. *Monthly Weather Review*, pages 875–883, 1970.

M. Scholze, T. Kaminski, P. Rayner, W. Knorr, and R. Giering. Propagating uncertainty through prognostic carbon cycle data assimilation system simulations. *Journal of Geophysical Research: Atmospheres*, 112(D17), 2007.

G. Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2): 461–464, 1978.

O. Talagrand. Assimilation of observations, an introduction. *Journal-Meteorological Society of Japan Series 2*, 75:81–99, 1997.

C. M. Trudinger, M. R. Raupach, P. J. Rayner, J. Kattge, Q. Liu, B. Pak, M. Reichstein, L. Renzullo, A. D. Richardson, S. H. Roxburgh, et al. Optic project: An intercomparison of optimization techniques for parameter estimation in terrestrial biogeochemical models. *Journal of Geophysical Research: Biogeosciences*, 112(G2), 2007.

M. Williams, P. A. Schwarz, B. E. Law, J. Irvine, and M. R. Kurpius. An improved analysis of forest carbon dynamics using data assimilation. *Global Change Biology*, 11(1):89–105, 2005.

X. Wu, Y. Luo, E. Weng, L. White, Y. Ma, and X. Zhou. Conditional inversion to estimate parameters from eddy-flux observations. *Journal of Plant Ecology*, 2009. doi: 10.1093/jpe/rtp005.

J. Zobitz, A. Desai, D. Moore, and M. Chadwick. A primer for data assimilation with ecological models using markov chain monte carlo (mcmc). *Oecologia*, 167(3):599–611, 2011.

J. Zobitz, D. J. Moore, T. Quaife, B. H. Braswell, A. Bergeson, J. A. Anthony, and R. K. Monson. Joint data assimilation of satellite reflectance and net ecosystem exchange data constrains ecosystem carbon fluxes at a high-elevation subalpine forest. *Agricultural and Forest Meteorology*, 195:73–88, 2014.

J. M. Zobitz, D. J. P. Moore, W. J. Sacks, R. K. Monson, D. R. Bowling, and D. S. Schimel. Integration of process-based soil respiration models with whole-ecosystem co2 measurements. *Ecosystems*, 11(2):250–269, 2008. ISSN 1435-0629. doi: 10.1007/s10021-007-9120-1.