University of Reading

Understanding the information content in diverse observation of forest carbon stocks and fluxes for data assimilation and ecological modelling

PhD in Atmosphere, Oceans and Climate

Department of Meteorology

Ewan Mark Pinnington

February 2017

Declaration

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

- Ewan Mark Pinnington

Abstract

Forest are important here is some stuff to better understand them!

Acknowledgements

Acknowledgements go here.

Contents

1 Introduction **2**

1.1 The global carbon cycle 2

1.2 Observations of terrestrial carbon balance 6

1.3 The role of models 7

1.4 Data assimilation 9

2 Thesis aims and outline **10**

3 Current state of data assimilation for the carbon cycle **11**

3.1 Data assimilation methods 11

3.1.1 Introduction to data assimilation 12

3.1.2 4D-Var 14

3.1.3 Sequential and Markov chain Monte Carlo approaches 15

3.2 Applications to the carbon cycle 16

3.2.1 Site-level applications 17

3.2.1.1 Early efforts 17

3.2.1.2 Data assimilation comparison projects 17

3.2.1.3 Use of Earth observation data 18

3.2.1.4 Current challenges 19

3.2.2 Global implementations . . . 20
3.2.3 Issues faced in carbon cycle data assimilation . . . 22
3.3 Summary . . . 23

4 Model and data **24**

4.1 Introduction . . . 24
4.2 Alice Holt research site . . . 24
4.3 Establishment of sampling points . . . 26
4.4 Leaf area index observations . . . 26
4.4.1 Ceptometer . . . 27
4.4.2 Hemispherical photographs . . . 29
4.4.3 Litter traps . . . 29
4.4.4 Comparison of methods . . . 30
4.5 Point-centred quater observations . . . 31
4.6 Flux tower observations and data processing . . . 33

5 Information content in observations relevant to forest carbon balance **36**

5.1 Introduction . . . 36
5.2 Background material . . . 37
5.2.1 Metolius forest site . . . 37
5.2.2 Observability . . . 38
5.2.3 Information content measures . . . 39
5.2.3.1 Sensitivity of analysis to observations . . . 40
5.2.3.2 Degrees of freedom for signal . . . 41
5.2.3.3 Shannon information content . . . 42
5.3 Results . . . 43

5.3.1 DALEC1 observability . . . 43
5.3.2 DALEC2 observability . . . 46
5.3.2.1 Observability for observations randomly distributed through time . 48
5.3.3 DALEC1 information content . . . 50
5.3.3.1 Information content for a single observation . . . 50
5.3.3.2 Information content for observations at a single time . . . 55
5.3.3.3 Information content in successive observations . . . 57
5.3.3.4 Effect of time correlations between observation errors on information content . . . 59
5.3.4 DALEC2 information content . . . 60
5.3.4.1 Information content in observations for DALEC2 . . . 60
5.3.4.2 Effect of time correlations on observation information content . . . 63
5.4 Conclusions . . . 64

6 Investigating the role of prior and observation error correlations 66

6.1 Abstract . . . 66
6.2 Introduction . . . 67
6.3 Model and Data Assimilation Methods . . . 71
6.3.1 Alice Holt research forest . . . 71
6.3.2 The DALEC2 model . . . 71
6.3.3 4D-Var . . . 73
6.3.4 Implementation and testing of 4D-Var system . . . 75
6.3.4.1 Test of tangent linear model . . . 76
6.3.4.2 Test of adjoint model . . . 78
6.3.4.3 Gradient test . . . 79

6.3.5 Including correlations in the background error covariance matrix 80
6.3.6 Specifying serial correlations in the observation error covariance matrix . . . 81
6.4 Results 83
6.4.1 Experiments 83
6.4.2 Experiment A 83
6.4.3 Experiment B 85
6.4.4 Experiment C 86
6.4.5 Experiment D 87
6.4.6 Summary 87
6.5 Discussion 92
6.6 Conclusion 97
6.7 Acknowledgements 97
6.8 Appendix 98

7 Using data assimilation to understand the effect of disturbance of the carbon dynamics of the Alice Holt forest **100**

8 Conclusion **101**

Bibliography **102**

Chapter 1

Introduction

1.1 The global carbon cycle

Carbon is one of the most abundant elements, making up around half of all living dry mass on Earth. The global carbon cycle describes the movement of carbon through the Earth system. In the Earth system large amounts of carbon are present in the oceans, atmosphere, land surface and crust. These stores of carbon are referred to as reservoirs or pools. The amount of carbon in this system can be considered constant, given that nuclear transmutation is not common under terrestrial conditions. Therefore terrestrial processes involving carbon can only transfer it between the global carbon pools. This is referred to as a flux. In pre-industrial times, the fluxes of carbon between different pools has only varied over long time scales ($\sim$100000 years) (Lüthi et al., 2008).

The greenhouse effect describes the process by which gases (CO_2, water vapour, ozone, etc.) in the Earth's atmosphere contribute to the warming of the planet by absorbing long-wave radiation emitted from the Earth's surface and reradiating this absorbed energy in all directions, causing more warming below (Mitchell, 1989). The natural greenhouse gas effect raises the global mean surface temperature by 30K, making the Earth habitable for its many lifeforms. The increase in atmospheric greenhouse gases due to anthropogenic activities since the industrial revolution, has amplified the greenhouse effect and resulted in increased global warming. CO_2 has been found to be the most important human-contributed compound to this warming (Falkowski et al., 2000). In figure 1.1 we show a simplified schematic of the global carbon cycle taken from the fifth Intergovernmental Panel on Climate Change (IPCC) report. In this schematic we can see the large rise in atmospheric CO_2 since the industrial revolution up to 2011, with an increase of 240 Pg C.

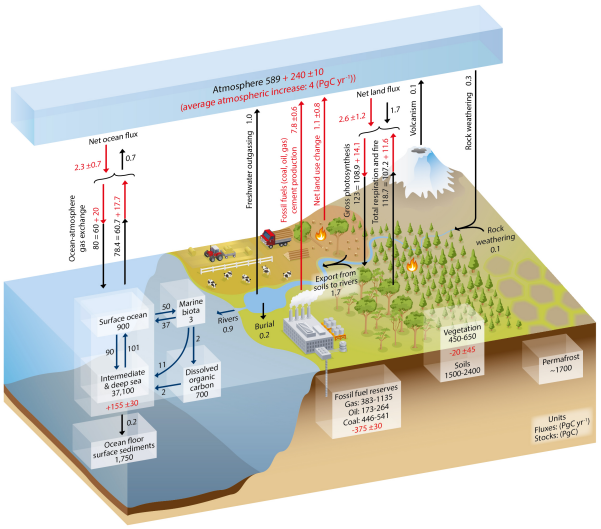


Figure 1.1: Global carbon cycle simplified schematic (Ciais et al., 2014). Black numbers and arrows represent reservoir mass and exchange fluxes estimated for the time prior to the industrial era ($\sim$ 1750). Red numbers and arrows represent annual fluxes averaged over the 2000-2009 time period. Red numbers in the reservoirs indicate the cumulative change of carbon over the industrial period (1750-2011).

As atmospheric CO_2 levels have risen, natural sinks of CO_2 (fluxes out of the atmosphere) have intensified with both the land surface and oceans absorbing more CO_2 from the atmosphere than in pre-industrial times. This can be see in figure 1.1, with the the net ocean flux of CO_2 to the atmosphere decreasing from an estimated +0.7 Pg C yr^{-1} to -2.3 Pg C yr^{-1}, and the land surface flux of CO_2 to the atmosphere decreasing from -1.7 Pg C yr^{-1} to -2.6 Pg C yr^{-1}. More recent estimates from Le Quéré et al. (2015) indicate these sinks have further intensified with the ocean sink estimated to be 2.9 $\pm$0.5 Pg C yr^{-1} and the land surface sink 4.1 $\pm$0.9 Pg C yr^{-1} for the year 2014. The intensification of the land carbon sink is thought to be partly due to a combination of forest regrowth as well as rising CO_2 and increased nitrogen deposition having a fertilisation effect (Ciais et al., 2014). It has also been shown that the land surface sink has been enhanced by an increase in diffuse photosynthetically active radiation as a result of increased cloud cover associated with increased anthropogenic emissions (Mercado et al., 2009).

The partitioning of global carbon fluxes between emissions and sinks is important to better model the carbon cycle. However, current estimates are subject to high levels of uncertainty, which are reflected by the errors shown in Figure 1.1. Current best estimates of global CO_2 emissions and their partitioning between atmospheric growth rate and sinks are shown in Figure 1.2. It is vitally important to understand the future response of sinks of CO_2 (land surface and oceans) to climate change. If either the oceans or land surface were to stop absorbing the same percentage of CO_2,

we would see even more dramatic increases in atmospheric CO_2 levels and thus a much greater rate of global warming. There is a high level of confidence that ocean carbon uptake will continue under all future emission scenarios (Ciais et al., 2014). There is much less confidence for the land surface and Booth et al. (2012) have shown that global warming is particularly sensitive to land surface carbon cycle processes, highlighting the need to improve understanding of land surface carbon uptake. Some estimates show the land surface changing from a sink of CO_2 to a source of CO_2 under certain future emission scenarios (Sitch et al., 2008; Cox et al., 2000; Scholze et al., 2006). In the latest IPCC report land surface carbon uptake is still considered the least understood process in the global carbon cycle (Ciais et al., 2014).

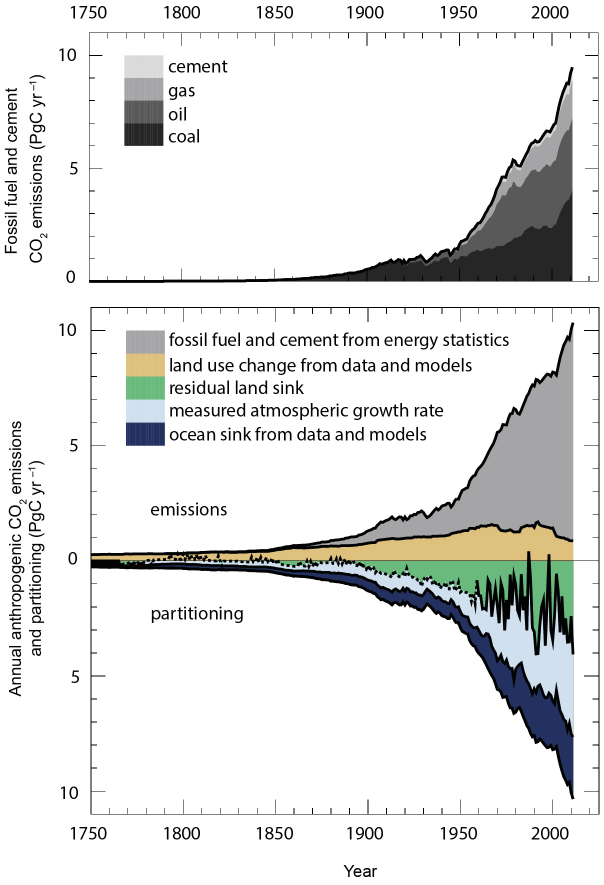


Figure 1.2: Annual anthropogenic CO_2 emissions and their partitioning among the atmosphere, land and ocean from 1750 to 2011 (Ciais et al., 2014).

Currently land surface carbon uptake is estimated by taking the residual of all other calculated sources and sinks of carbon, so that

$$S_{LAND} = E_{FF} + E_{LUC} - (G_{ATM} + S_{OCEAN}) \tag{1.1}$$

where S_{LAND} is the global residual land sink of CO_2, E_{FF} is the CO_2 emissions from fossil fuels,

E_{LUC} is the CO_2 emissions from land use change (mainly deforestation), G_{ATM} is the atmospheric CO_2 growth rate and S_{OCEAN} is the mean ocean CO_2 sink (Le Quéré et al., 2015). Figure 1.2 shows the growth in the estimated residual land sink as emissions increase. The high variability shown in this sink is largely due to year to year variations in precipitation, surface temperature, radiation and volcanic eruptions. Figure 1.2 shows that in 1986 and 1997 the land sink drops to zero, both of these years were among the strongest El Niño's in recent history. In 1997 tropical droughts, often associated with El Niño, were particularly severe leading to wildfires that released vast amounts of stored carbon (Schimel, 2013).

Terrestrial ecosystems are made up of autotrophs (organisms capable of photosynthesis) and heterotrophs (organisms that feed on organic carbon). The Gross Primary Productivity (GPP) of an ecosystem is the total amount of carbon removed from the atmosphere by photosynthesis. The total ecosystem respiration (RT) is made up of autotrophic respiration (e.g. from plants) and heterotrophic respiration (e.g. from soil and litter organisms). The total carbon uptake is then equal to -GPP+RT. A representation of these fluxes for a forest ecosystem are shown in Figure 1.3.

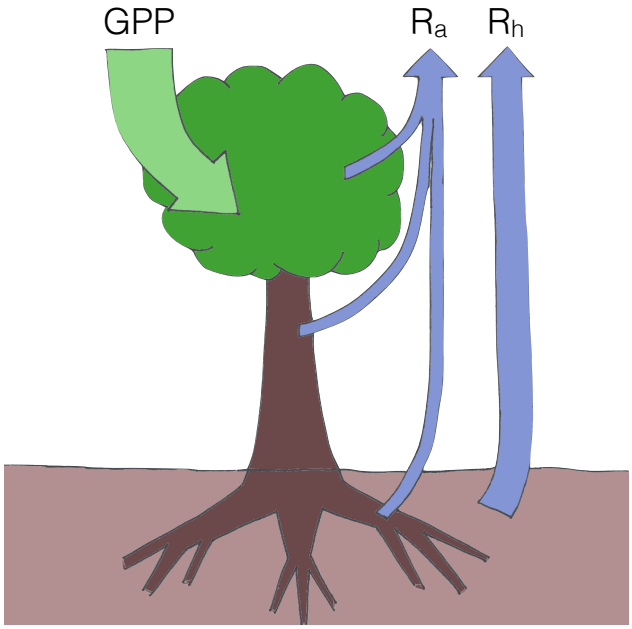


Figure 1.3: Fluxes of carbon through a forest ecosystem. Gross Primary Productivity (GPP) represents total photosynthesis, R_a is autotrophic respiration from foliage, wood and roots, R_h is heterotrophic respiration from soil and litter. Total ecosystem respiration of carbon to the atmosphere (RT) is equal to R_a + R_h. The net ecosystem exchange of CO_2 is equal to -GPP + RT.

Disturbance of terrestrial ecosystems from fire, felling and insect outbreak can have significant impacts on carbon dynamics. Land use change is the second largest anthropogenic source of CO_2. However, It is not well understood how much CO_2 is removed from the atmosphere by regrowth

of previously disturbed ecosystems (either by felling or fire), although it is thought that regrowth of forests in partcular could be stronger carbon sinks than their predecessors, due to more rapid biomass accumulation under succession (Pan et al., 2011). Better understanding the response of the land surface to disturbance will help constrain future carbon budgets.

1.2 Observations of terrestrial carbon balance

There are an increasing number of available observations relevant to understanding the carbon balance of forests and the terrestrial biosphere. These observations include a range of variables, perhaps two of the most common are the Net Ecosystem Exchange (NEE) of CO_2 and Leaf Area Index (LAI), which is the area of leaves per unit area ground. These variables can be directly measured at site level and can also be estimated from satellite remote sensing.

At site level, flux towers measuring ecosystem-atmosphere fluxes of CO_2, water and energy using the micrometeorological technique of eddy covariance provide one of the most valuable sources of information. Direct observations of ecosystem CO_2 uptake are made at a fine temporal resolution, with observations every half-hour. A global flux network (FLUXNET), was established in 1997 (Baldocchi et al., 2001), to consolidate the information from a growing number of flux tower sites. Currently there are 517 active FLUXNET sites which are shown in Figure 1.4, as can be seen these sites are not uniformly distributed so it is not possible to use FLUXNET sites alone to produce global estimates of terrestrial CO_2 balance. However, these sites do provide an invaluable resource for model and satellite calibration. In turn this can be used to produce estimates on a global scale. At many flux tower sites and forest stands other diverse observations relevant to terrestrial carbon budgets are also being made. These include observations of soil and litter respiration, woody biomass and LAI. However, because they are labour intensive these observations are made much less frequently.

The Moderate Resolution Imaging Spectroradiometer (MODIS) on the TERRA and AQUA satellites produces global estimates of LAI and Gross Primary Productivity (GPP) for terrestrial ecosystems (Running et al., 2004). However, MODIS actually measures reflected sunlight, this is then converted to vegetation indices, such as the Normalised Difference Vegetation Index (NDVI). These indices are correlated with the fraction of absorbed visible sunlight to estimate LAI or used in simple algorithms to estimate GPP (Yuan et al., 2007). It is therefore important to understand

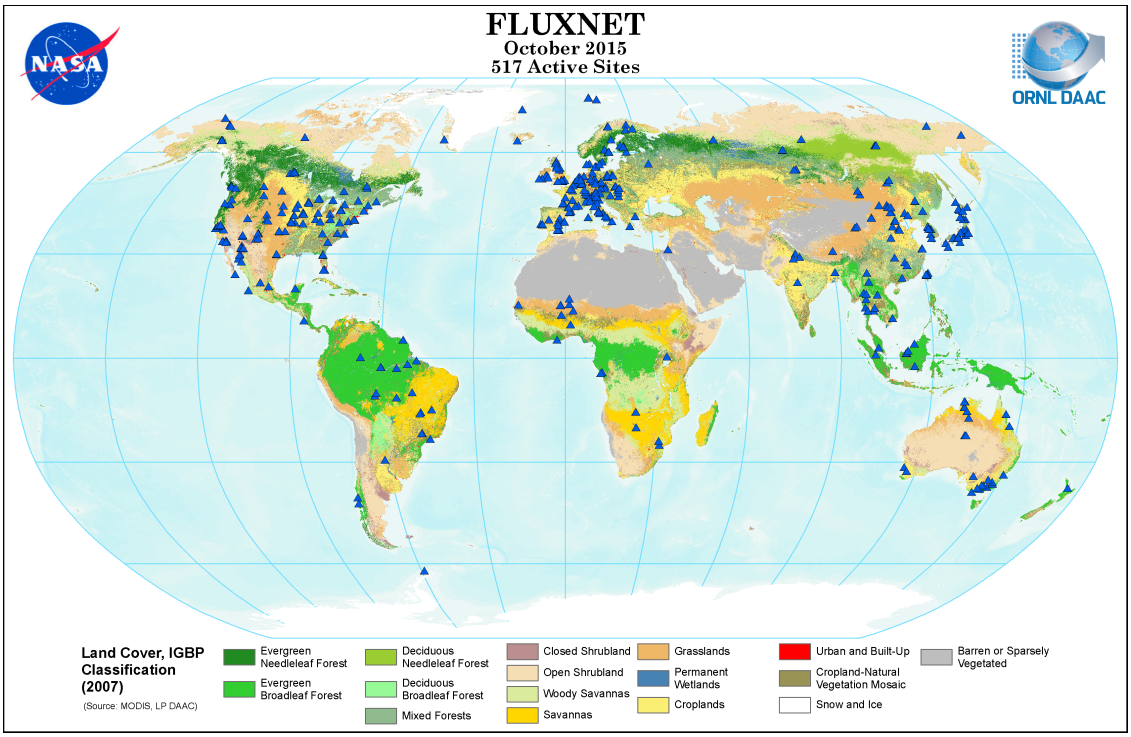


Figure 1.4: FLUXNET sites and land cover (MODIS IGBP classification) (Oak Ridge National Laboratory Distributed Active Archive Center ORNL DAAC, 2013).

the limitations when interpreting satellite products as they do not represent direct observations. For LAI it has been shown that remotely sensed estimates saturate when measuring ecosystems with a LAI above 3 (Myneni et al., 2002). Terrestrial fluxes of carbon estimated from satellite measurements are subject to large errors in representativity, as satellites view a scene almost instantaneously and then derive daily mean fluxes (Baldocchi, 2008).

1.3 The role of models

Observations can only tell us about the current and past state of a system. In order to produce future predictions and better understand current terrestrial carbon dynamics we must use mathematical models. Figure 1.5 show a comparison of the residual land sink (described in section 1.1) with the global terrestrial CO_2 sink estimated from different process based global carbon cycle models. We see that although there is a high variability between modelled estimates there is good agreement between the multi-model mean and the residual land sink.

Representative Concentration Pathways (RCPs) of CO_2 concentrations and emissions have been developed (Moss et al., 2010) to drive climate models to produce future predictions. Under these pathways land surface carbon uptake is highly uncertain with little agreement between dif-

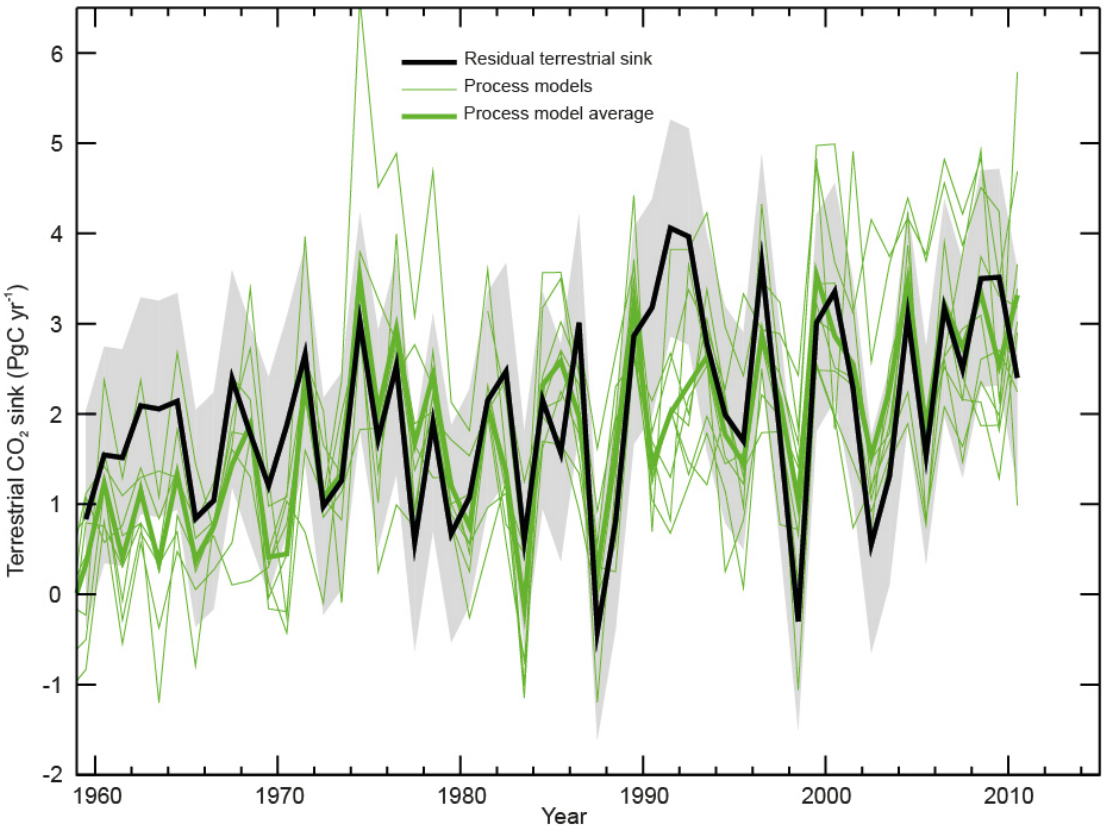


Figure 1.5: Comparison of the residual land sink (black line) with the global terrestrial CO_2 sink estimated from different process based global carbon cycle models (Ciais et al., 2014). Grey shading represents uncertainty in residual land sink.

ferent process based models. Some predicting the land surface to become a source of CO_2 and others predicting a further intensification of the residual land sink (Jones et al., 2013). This large uncertainty for land surface models is partly due to poor model parameterisations and missing processes within models. One of the main processes many current global models do not account for is the effect of disturbance on terrestrial ecosystem carbon dynamics.

It has been shown that many terrestrial carbon cycle models simulating the seasonal cycle of land-atmosphere CO_2 exchange perform poorly when compared to FLUXNET sites in North America (Schwalm et al., 2010). Here a difference between observations and model predictions of 10 times the observational uncertainty was found, highlighting the need for continued model development. In order to improve global models of terrestrial carbon balance it is important to use site-level-research to hone the processes and parameterisations of the models where we have diverse sets of direct observations with which to judge modified-model performance.

1.4 Data assimilation

As discussed above, the level of uncertainty in terrestrial carbon balance predictions arise from significant gaps in the direct observations available and from a lack of clarity and authoritative parameterisation of the constituent processes in current models. The technique of data assimilation provides a method for combining and comparing the output of predictive models with incomplete observations to find the best estimate for the state and parameters of a system. Data assimilation has had many successful applications. Perhaps the most important application has been in numerical weather prediction where data assimilation has contributed to forecast accuracy being increased at longer lead times, with the result that the four day forecast in 2014 now has the same level of accuracy as the one day forecast in 1979 (Bauer et al., 2015). Obviously, this improved forecasting is not solely due to data assimilation but also increased quality and resolution of observations along with improvements in model structure, however the introduction and evolution of data assimilation has been a key part of the improvement (Dee et al., 2011).

More recently data assimilation has been used to improve our knowledge of ecological systems. For the carbon balance of forests it has been used to combine many different observations with functional ecology models (Zobitz et al., 2011; Fox et al., 2009; Richardson et al., 2010; Quaife et al., 2008; Zobitz et al., 2014b; Niu et al., 2014). Global land surface models have also been implemented with data assimilation, mainly using data from satellite and atmospheric CO_2 observations (Kaminski et al., 2013; Scholze et al., 2007). In a few cases site level data has also been assimilated (Verbeeck et al., 2011; Bacour et al., 2015). In comparison with numerical weather prediction, the use of data assimilation in these areas is relatively new and underdeveloped. The further application of data assimilation to models of ecosystem carbon balance will help to improve model parameterisations and future predictions. The development of improved data assimilation techniques will also help to identify missing model processes and changes in model parameters and behaviour over time. In particular, understanding the change in model parameters over time will be of use in improving models predictions of the effect of disturbance in terrestrial ecosystems.

Chapter 2

Thesis aims and outline

Chapter 2: Thesis aims and outline

We aims to do this and that

- UNderstanding this!
- investigating that!
- Using something else!

Chapter 3

Current state of data assimilation for the carbon cycle

3.1 Data assimilation methods

Data assimilation provides techniques for combining observations and prior knowledge of a system in an optimal way to find an improved estimate of the system. 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 (3.1.1 to 3.1.3) 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.

3.1.1 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

$$\mathbf{x}^t = \begin{pmatrix} \mathbf{p}^t \\ \mathbf{z}^t \end{pmatrix} \in \mathbb{R}^{q+n}. \tag{3.1}$$

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

$$\mathbf{x}^b = \mathbf{x}^t + \boldsymbol{\epsilon}^b, \tag{3.2}$$

$$\mathbf{y} = h(\mathbf{x}^t) + \boldsymbol{\epsilon}^o, \tag{3.3}$$

where $\boldsymbol{\epsilon}^b$ and $\boldsymbol{\epsilon}^o$ are the prior and observation errors respectively, and $h : \mathbb{R}^{q+n} \to \mathbb{R}^m$ 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 (3.2) and (3.3) is often called the analysis or the posterior 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

$$\mathbb{P}(A|B) \propto \mathbb{P}(A)\mathbb{P}(B|A). \tag{3.4}$$

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 with

$$\mathbb{P}^b(\mathbf{x}) = \frac{1}{\sqrt{|2\pi\mathbf{B}|}}\exp\Big(-\frac{1}{2}(\mathbf{x}-\mathbf{x}^b)^T\mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}^b)\Big) \tag{3.5}$$

and

$$\mathbb{P}^o(\mathbf{y}|\mathbf{x}) = \frac{1}{\sqrt{|2\pi\mathbf{R}|}}\exp\Big(-\frac{1}{2}(\mathbf{y}-h(\mathbf{x}))^T\mathbf{R}^{-1}(\mathbf{y}-h(\mathbf{x}))\Big), \tag{3.6}$$

where $\mathbb{P}^b(\mathbf{x})$ is the probability density function for the prior and $\mathbb{P}^o(\mathbf{y}|\mathbf{x})$ the probability density function of the observations given the augmented state. Then from Bayes' theorem (equation (3.4)) the posterior probability density function for the augmented state

$$\begin{aligned}\mathbb{P}^a(\mathbf{x}|\mathbf{y}) &\propto \frac{1}{\sqrt{|2\pi\mathbf{B}|}\sqrt{|2\pi\mathbf{R}|}}\exp\Big(-\frac{1}{2}(\mathbf{x}-\mathbf{x}^b)^T\mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}^b)-\frac{1}{2}(\mathbf{y}-h(\mathbf{x}))^T\mathbf{R}^{-1}(\mathbf{y}-h(\mathbf{x}))\Big)\\ &\propto \exp\Big(-\frac{1}{2}(\mathbf{x}-\mathbf{x}^b)^T\mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}^b)-\frac{1}{2}(\mathbf{y}-h(\mathbf{x}))^T\mathbf{R}^{-1}(\mathbf{y}-h(\mathbf{x}))\Big),\end{aligned} \tag{3.7}$$

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 (3.7) 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

$$J(\mathbf{x}) = \frac{1}{2}(\mathbf{x}-\mathbf{x}^b)^T\mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}^b)+\frac{1}{2}(\mathbf{y}-h(\mathbf{x}))^T\mathbf{R}^{-1}(\mathbf{y}-h(\mathbf{x})). \tag{3.8}$$

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 (3.8) and its gradient (Courtier et al., 1998). We can approximate the minimum of (3.8) by finding its gradient and setting it to zero to obtain the Best Linear Unbiased Estimate (BLUE) (Talagrand, 1997) where

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{K}(\mathbf{y}-h(\mathbf{x}^b)), \tag{3.9}$$

$$\mathbf{K} = \mathbf{B}\mathbf{H}^T(\mathbf{H}\mathbf{B}\mathbf{H}^T+\mathbf{R})^{-1}, \tag{3.10}$$

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

$$\mathbf{A} = (\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}+\mathbf{B}^{-1})^{-1}, \tag{3.11}$$

if h is linear then (3.9) and (3.11) are exact solutions.

3.1.2 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,

$$J(\mathbf{x}_0) = \frac{1}{2}(\mathbf{x}_0 - \mathbf{x}^b)^T \mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}^b) + \frac{1}{2}\sum_{i=0}^{N}(\mathbf{y}_i - \mathbf{h}_i(\mathbf{x}_i))^T \mathbf{R}_i^{-1}(\mathbf{y}_i - \mathbf{h}_i(\mathbf{x}_i)), \tag{3.12}$$

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,

$$\mathbf{x}_i = \mathbf{m}_{i-1\to i}(\mathbf{x}_{i-1}), \tag{3.13}$$

where $\mathbf{x}_i$ is the model augmented state at time t_i, $\mathbf{m}_{i-1\to i}$ 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 (3.12) to avoid the sum notation as

$$J(\mathbf{x}_0) = \frac{1}{2}(\mathbf{x}_0 - \mathbf{x}^b)^T \mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}^b) + \frac{1}{2}(\hat{\mathbf{y}} - \hat{\mathbf{h}}(\mathbf{x}_0))^T \hat{\mathbf{R}}^{-1}(\hat{\mathbf{y}} - \hat{\mathbf{h}}(\mathbf{x}_0)) \tag{3.14}$$

where,

$$\hat{\mathbf{y}} = \begin{pmatrix} \mathbf{y}_0 \\ \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_N \end{pmatrix}, \ \hat{\mathbf{h}}(\mathbf{x}_0) = \begin{pmatrix} \mathbf{h}_0(\mathbf{x}_0) \\ \mathbf{h}_1(\mathbf{m}_{0\to 1}(\mathbf{x}_0)) \\ \vdots \\ \mathbf{h}_N(\mathbf{m}_{0\to N}(\mathbf{x}_0)) \end{pmatrix}, \text{ and } \hat{\mathbf{R}} = \begin{pmatrix} \mathbf{R}_{0,0} & \mathbf{R}_{0,1} & \ldots & \mathbf{R}_{0,N} \\ \mathbf{R}_{1,0} & \mathbf{R}_{1,1} & \ldots & \mathbf{R}_{1,N} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{R}_{N,0} & \mathbf{R}_{N,1} & \ldots & \mathbf{R}_{N,N} \end{pmatrix}. \tag{3.15}$$

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

$$\mathbf{A} = (\hat{\mathbf{H}}^T \hat{\mathbf{R}}^{-1} \hat{\mathbf{H}} + \mathbf{B}^{-1})^{-1}, \tag{3.16}$$

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

$$\hat{\mathbf{H}} = \begin{pmatrix} \mathbf{H}_0 \\ \mathbf{H}_1 \mathbf{M}_0 \\ \vdots \\ \mathbf{H}_N \mathbf{M}_{N,0} \end{pmatrix} \tag{3.17}$$

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.

3.1.3 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 one of the first MCMC methods 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 (6.16). As these methods use the negation of the cost function in equation (6.16) 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 (3.9) 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 **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 **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. Techniques have been employed to avoid this (Hamill et al., 2001). The ensemble can also collapse on the same state/parameter values after assimilating a number of observations, this can be avoided by covariance inflation (Anderson and Anderson, 1999).

3.2 Applications to the carbon cycle

For numerical weather prediction the physics of the atmosphere are well understood and, as such, parameterisations should not change over time. Therefore DA is used predominantly for state estimation. However this is not true for land surface carbon balance models where parameters are much less well understood. Indeed these parameters can change over time within a developing ecosystem or when an ecosystem is subject to a disturbance event. Therefore, the vast majority of current studies use DA to estimate both parameter and state variables.

The use of DA for the estimation of parameter and state variables of ecosystem carbon models has either been at site-level, with flux tower observations and other ancillary data relevant to ecosystem carbon balance, or for global implementations, where often the implied effect of the land surface on atmospheric CO_2 observations has been considered. It is important that we improve DA techniques both at site-level and for global implementations.

3.2.1 Site-level applications

3.2.1.1 Early efforts

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 from the Metolius ponderosa pine site (Oregon, USA) with the DALEC model. 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. The use of a variational or MCMC technique may have been more logical as the aim of the study was to estimate the initial state and parameter values of the model. Williams et al. (2005) found large reductions in model prediction error after assimilation. They noted that rare measurements of carbon stocks had limited impact on assimilation results but suggested that longer time-series of these stock measurements will be important to constrain carbon pool turnover rates. They also assimilated modelled GPP from the more complex aggregated canopy model (ACM) (Williams et al., 1997) and claimed that this was analogous to satellite derived GPP, as this more complex model was already calibrated for the Metolius forest. They suggested that, based on their results assimilating ACM modelled GPP, in future studies using satellite GPP products would be beneficial.

3.2.1.2 Data assimilation comparison projects

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 partic-

ipant 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, which resulted in biased results. Williams et al. (2009) comment that temporal error correlations between flux measurements on the scale of a day and less are likely to be severe. They suggested that these could be included in the observation error covariance matrix, although they comment that this would be a difficult task.

The REgional Flux Estimation eXperiment (REFLEX) was a similar study conducted using the DALEC model by Fox et al. (2009). In this study 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), with no DA technique performing consistently better than others. Across all methods, the parameters linked directly to GPP and TER were best constrained, while those linked to slower processes (allocation and turnover of fine root and wood carbon pools) were 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. The representation of prior and observational errors are still very basic in the majority of current DA schemes for ecosystem carbon balance. (Dietze et al., 2013) also stress the need to improve the representation of uncertainty 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. PEcAn also allows for easier comparison of different implemented models (Dietze et al., 2013) with the aim of improving the standard and reproducibility of experimental results.

3.2.1.3 Use of Earth observation data

Satellite observations of reflectance have also been used with these simple models to assess their impact on modelled estimates. 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 (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 led to an over-estimate in both GPP and TER. Despite this, the modelled NEE was improved after assimilation when compared to site flux tower observations and significant reductions in modelled flux uncertainties were achieved.

Satellite data has also been used with the SIPNET model, Zobitz et al. (2014a) assimilated earth observation data with flux tower NEE on different timescales. 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.

3.2.1.4 Current challenges

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 (Zobitz et al., 2008; Moore et al., 2008; Sacks et al., 2007; Keenan et al., 2011). One problem facing studies working with NEE flux observations alongside other ancillary site-level data is the over-weighting 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 missmatches, 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 novel methods to better understand the impact that new observations could have on carbon cycle DA results. This will also allow for a more considered approach when planning measurement campaigns. It has also been suggested that effort should be made to define improved observation operators and the specification of their errors Rayner (2010); Williams et al. (2009).

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).

3.2.2 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 with BETHY 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, Bloom et al. (2016) implemented an MCMC technique (with prior constraints from Bloom and Williams (2015)) to find a global a $1^o \times 1^o$ DALEC2 map. Using MCMC techniques in this global implementation 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 were 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) used the retrieved global DALEC2 map to gain insight

into ecosystem functioning and suggested 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.

3.2.3 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 challenges:

- Equifinality: Many different combinations of parameters and state values are 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; Williams et al., 2009).

- Representation of prior and observational errors: Current DA schemes take a very simple approach to defining errors and many of the 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 aid the improved characterisation of uncertainties.

3.3 Summary

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 data assimilation method 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, although this PhD is more concerned with site-level implementations, is also applicable to larger scale DA implementations for the carbon cycle.

Chapter 4

Model and data

4.1 Introduction

As part of this PhD an extended period of time has been spent at the Alice Holt Research Station (Hampshire, UK) working with Forest Research (The research arm of the UK Forestry Commission). After initially completing one year of an ongoing field campaign to measure stem respiration using an infra-red gas analyser, a measurement campaign was designed to produce a set of observations for use in this PhD project. This involved the establishment and sampling of three transects throughout the Straits Inclosure (part of the Alice Holt forest). The establishment of these transect and measurements are outlined in this chapter.

4.2 Alice Holt research site

The Alice Holt Forest is a research forest area managed by the UK Forestry Commission located in Hampshire, SE England. Forest Research have been operating a CO_2 flux measurement tower in a portion of the forest, the Straits Inclosure, continuously since 1998. The Straits Inclosure is a 90 ha area of deciduous broadleaved plantation woodland located on a surface water gley soil and was initially planted with oak in the 1820s (Schlich and Perrée, 1905) and then replanted in the 1930s. The majority of the canopy trees are oak (*Quercus robur* L.), with an understory of hazel (*Corylus avellana* L.) and hawthorn (*Crataegus monogyna* Jacq.) (Pitman and Broadmeadow, 2001), but there is a small area of conifers (*Pinus nigra* ssp. *laricio* (Maire) and *P. sylvestris* L.) within the tower measurement footprint area depending on wind direction. An aerial photograph of the

site is shown in Figure 4.1. The Straits Inclosure is a flat area at an altitude of approximately 80m, surrounded by mixed lowland woods and both arable and pasture agricultural land. In Wilkinson et al. (2012) an analysis of stand-scale 30 minute average net CO_2 fluxes (NEE) from 1998-2011 for the Straits Inclosure found a mean annual NEE of -486 g C m^{-2} yr^{-1} and demonstrated the forest was a substantial sink of carbon. This study also includes further details about the research site.

As part of the management regime, the Straits Inclosure is subject to thinning, whereby a proportion of trees are removed from the canopy in order to reduce competition and improve the quality of the final tree crop. At the Straits an intermediate thinning method is used with a portion of both subdominant and dominant trees being removed from the stand (Kerr and Haufe, 2011). The whole of the stand was thinned in 1995. Subsequently the eastern side of the Straits was thinned in 2007 and then the western side in 2014. The flux tower at the site is situated on the boundary between these two sides. This allows for the use of a footprint model to split the flux record and thus analyse the effect of this disturbance on carbon fluxes at the site. In Wilkinson et al. (2015) a statistical analysis of the eddy covariance flux record found that there was no significant effect on the net carbon uptake of the eastern side after thinning in 2007. In this thesis we focus on the effect of disturbance on the western side after thinning in 2014 in chapter REF. We therefore refer to the western side as "thinned" forest and the eastern side as "unthinned" forest.

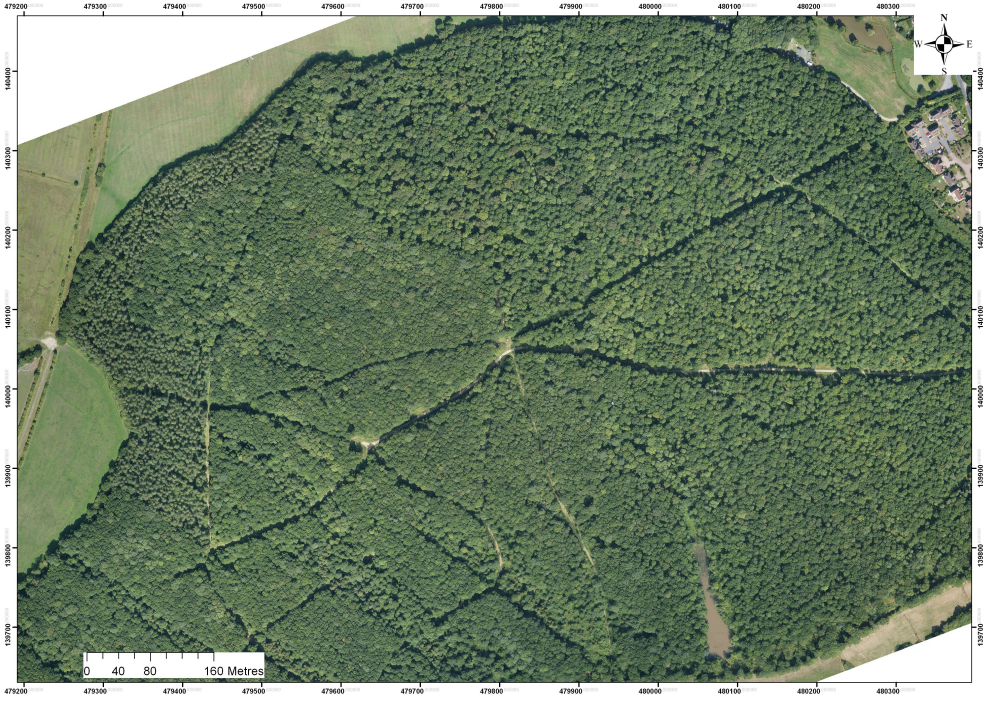


Figure 4.1: The Straits Inclosure research site in 2013.

4.3 Establishment of sampling points

For this fieldwork transects were designed to join up existing mensuration plots where measurements of woody biomass are made by Forest Research. This allowed for comparison with historic observations. In total 435 sampling points were marked at 10m intervals, these are shown in Figure 4.2. Python was used to calculated the exact latitude and longitude of each sampling point for the 3 transects, these locations were then entered into a GPS unit. When establishing the transects fluorescent spray paint was used to mark trees closest to each sampling point as shown on the GPS (see Figure 4.3). As parts of the forest site were extremely dense with vegetation a pair of loppers were used to clear a path in some areas to allow for the construction of relatively straight transects. Having all transect points numbered and corresponding to a latitude and longitude value allowed for comparison between methods and the splitting of observations between different distinct sections of the forest site.

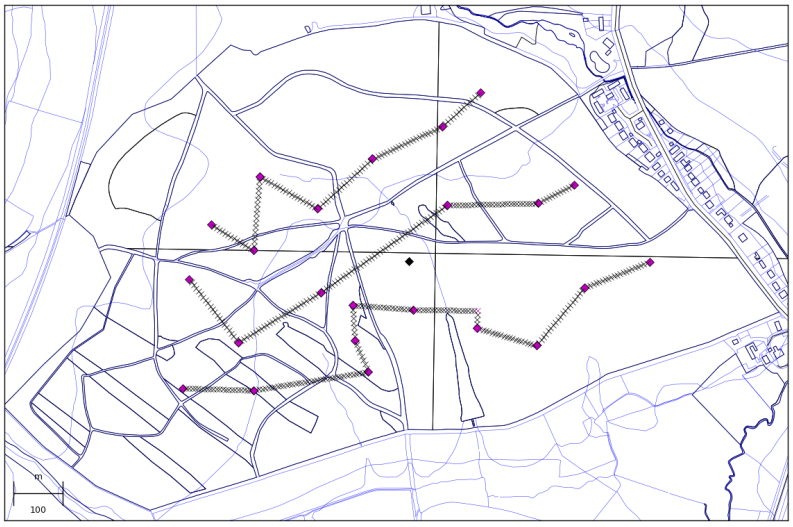


Figure 4.2: Sampling transects. Black crosses: sampling points at 10m intervals, pink diamonds: Forest Research mensuration plots, black diamond: Forest Research flux tower.

4.4 Leaf area index observations

Leaf Area Index (LAI) is an important variable in relation to the amount of CO_2 an ecosystem can remove from the atmosphere through photosynthesis. LAI is defined as the area of leaves per unit area of ground. Three different methods were used to find estimates for peak LAI (July -

Figure 4.3: Sampling point 291, showing fluorescent spray paint used to mark sampling points.

September) for the year 2015 along the three transects at different sampling intervals.

4.4.1 Ceptometer

A Decagon LP-80 ceptometer and an additional Photosynthetically Active Radiation (PAR) sensor were used to measure LAI. Here we measure below canopy PAR using the ceptometer while logging above canopy PAR using a data logger and PAR sensor positioned outside the canopy. We can then calculate LAI using the above and below canopy readings and a set of equations relying on some assumptions (Fassnacht et al., 1994). The ceptometer represents the quickest method for estimating LAI, we therefore took readings with the ceptometer at every sampling point over two walks of the transects, giving us 870 observations in total.

In order to be sure that the PAR readings from the ceptometer and external PAR sensor were consistent we had to calibrate the PAR sensor against the ceptometer. This was done by leaving both the PAR sensor and ceptometer out logging next to each other every 10 seconds for a day in the Alice Holt Research Station met square. We can then calibrate he output of the PAR sensor with that of the ceptometer as shown in Figure 4.4.

Once the PAR sensor was calibrate measurements could be made along the transects. The

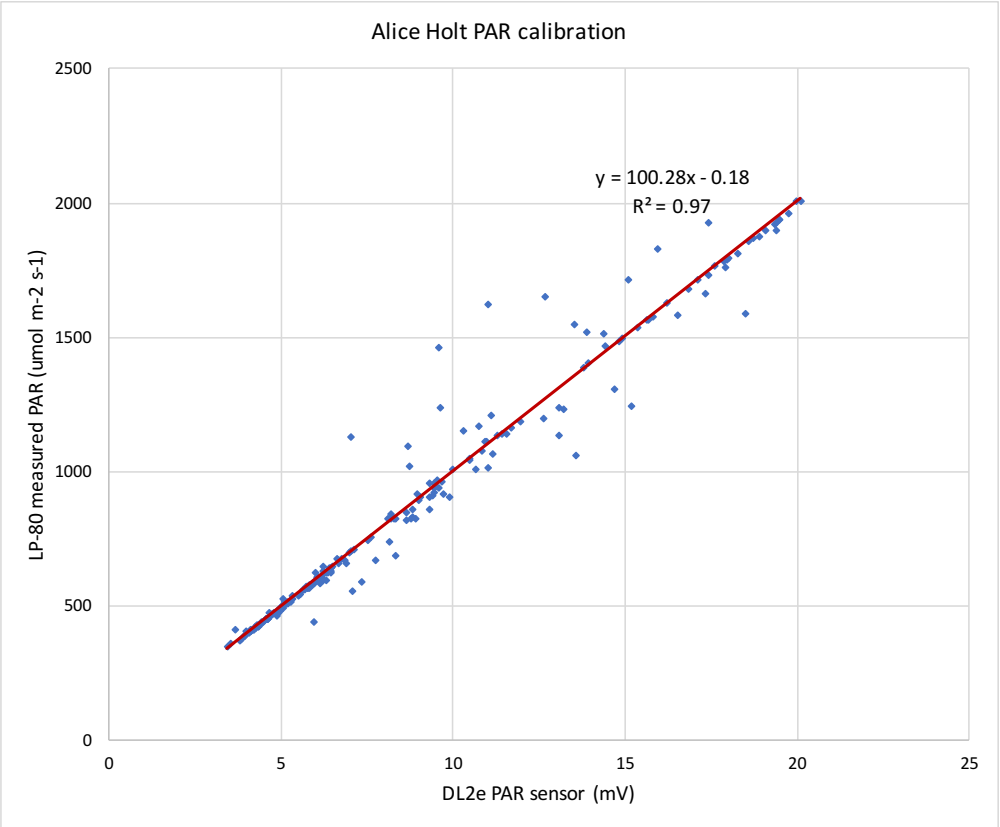


Figure 4.4: Calibration of above canopy Photosynthetically Active Radiation (PAR) sensor (measuring in mV) with LP-80 ceptometer measured PAR (μmol m^{-2} s^{-1}).

PAR sensor positioned outside of the canopy was logged every 5 seconds using a Delta-T DL2e data logger, at the start of every set of measurements the clock on the data logger and ceptometer were synchronised to ensure comparison of measurements made at the same time. After sampling the transects we had a set of above canopy and below canopy PAR readings corresponding to each sampling point for both walks of the transects. We use the same calculation for LAI as given in the Decagon LP-80 manual. This is using a simple model of radiation transmission and scattering tested against the more complex model of Norman and Jarvis (1975). The equation used to calculate LAI from the above and below canopy PAR readings is,

$$LAI = \frac{((1-\frac{1}{2K})f_b - 1)\ln\tau}{A(1-0.47f_b)}, \tag{4.1}$$

where K is the extinction coefficient, f_b is the beam fraction, $\tau = \frac{\text{below canopy PAR}}{\text{above canopy PAR}}$ and $A = 0.283 + 0.785a - 0.159a^2$ (where a is the leaf absorptivity, assumed to be 0.9 by Decagon). We assume a spherical leaf angle distribution parameter, $\chi = 1$, this means the extinction coefficient simplifies to $K = \frac{1}{2\cos\theta}$, where θ is the solar zenith angle. We took the mean of the two LAI observations at each point to give as an estimate to the peak LAI for the year 2015. We can see the LAI estimate

for the Straits Inclosure in Figure 4.5.

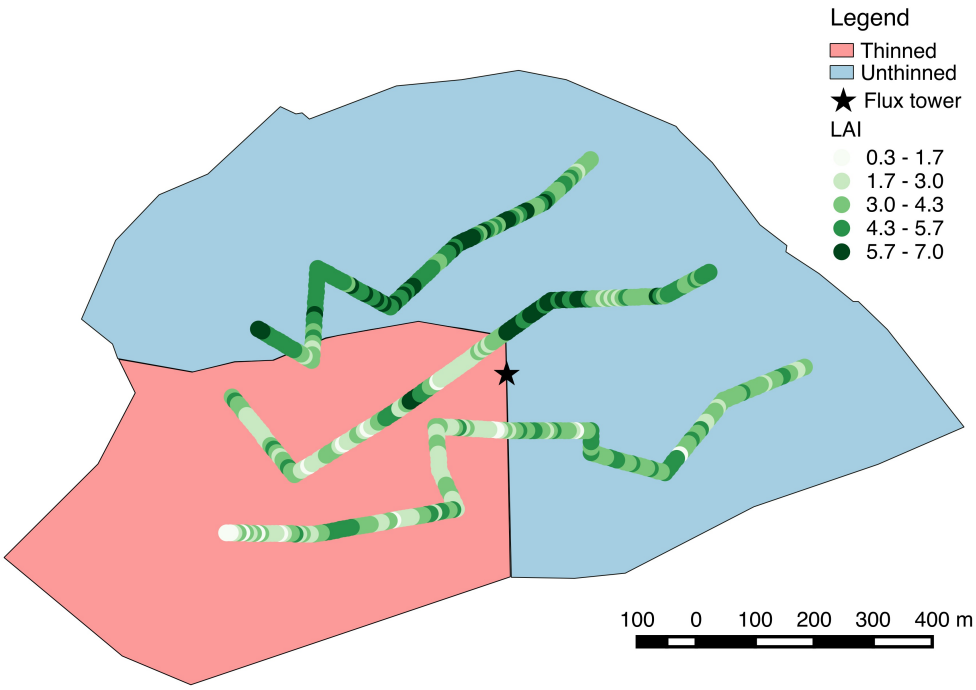


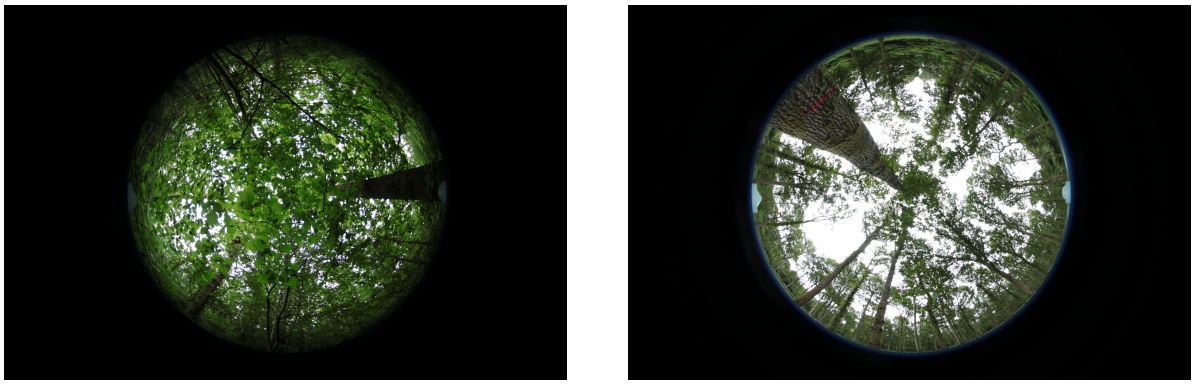
Figure 4.5: Ceptometer derived LAI for Alice Holt.

4.4.2 Hemispherical photographs

The next method used to measure LAI was hemispherical photography. Hemispherical photographs show a complete view of the sky in all directions, from these images we use the HemiView software (Rich et al., 1999) which calculates the proportion of visible sky as a function of sky direction (gap fraction) which it then uses to calculate LAI (Jonckheere et al., 2004). Hemispherical photographs were taken every 50m along the transects, giving a total of 89 images. It is important to that hemispherical photographs are taken in overcast conditions so that the sun does not obscure areas of leaf area. It is important to note that we did not remove tree trunks and branches from our calculation of LAI with HemiView so that we are actually calculating plant area index. The impacts of this assumption are discussed in section 4.4.4. In Figure 4.6 we show an example of two hemispherical photographs taken in different areas of the Straits Inclosure.

4.4.3 Litter traps

Finally litter traps were used to find estimates to LAI and leaf mass per area. Here we placed litter traps under the canopy to catch leaf litter as it falls into a bag attached to the bottom of the trap.



(a) Unthinned forest (b) Thinned forest

Figure 4.6: Hemispherical photographs from the Alice Holt flux site showing the difference between the thinned and unthinned sides of the forest.

The bags were changed every week during the litter fall period and the litter sorted into species. Every week the litter was dried in an oven at 70°C and weighed. This gave us the dry-weight of the leaf litter for the 2015 season. Towards the end of the season we scanned a subsample for each species of 100 leaves to find an area, we then dried and weighed each subsample, a relationship between dry-weight and leaf area was then be built (leaf mass per area) and used to infer the total LAI for each trap once the whole seasons litter has been collected. This method of LAI calculation is the most time consuming.

A total of six litter traps were established at points along the transects (positions shown in Figure 4.7) allowing for comparison with the other methods. The 6 litter traps are not enough to describe the LAI for the research site (Kimmins, 1973). We use these litter traps as a point of comparison and validation for the ceptometer and hemispherical photograph estimates of LAI made at the same locations and also for estimates to leaf mass per area. From our litter trap observations we find a leaf mass per area of 29 g C m^{-2} free soluble carbohydrates for both sides of the forest.

4.4.4 Comparison of methods

In Figure 4.8 and 4.9 we show a comparison of the different methods of estimating LAI for the unthinned and thinned forest respectively. We can see that in all cases LAI derived from the litter traps is always greater than LAI estimated from optical methods, this is expected from previous comparisons (Bréda, 2003).

Although the ceptometer is the fastest method for measuring LAI it is also the most variable,

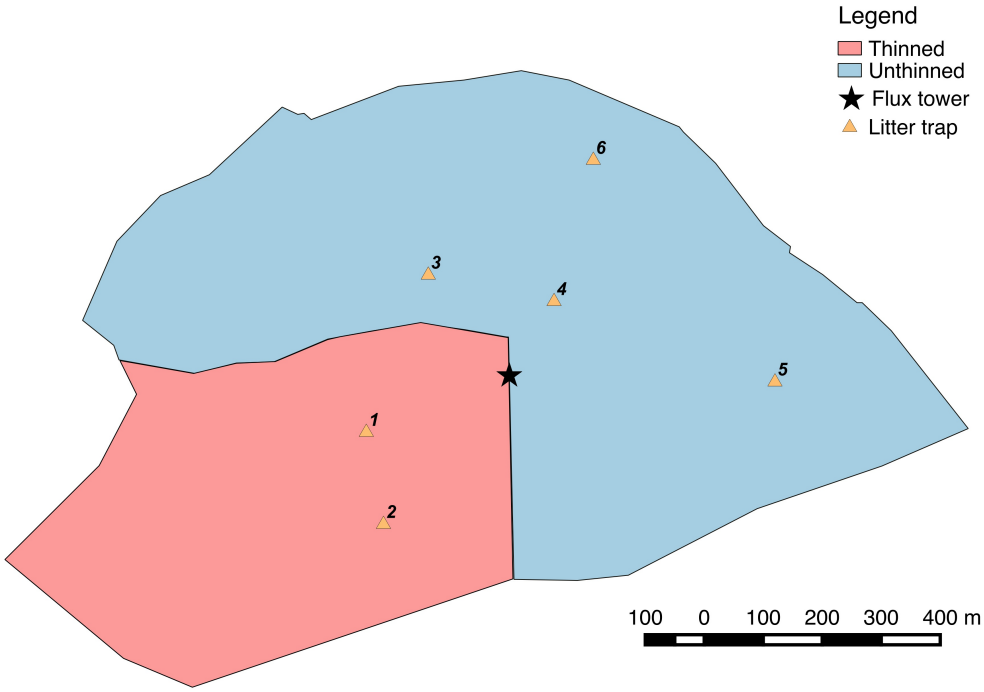


Figure 4.7: Litter trap locations for Alice Holt.

being extremely sensitive to the solar zenith angle and clear sky conditions. If the sun is low in the sky the radiation will pass through much more photosynthetically active material than if the sun is directly above head, causing spikes in the LAI value. We can see that the LAI estimates from the hemispherical photographs are much less variable than the ceptometer. As discussed in section 4.4.2 the hemispherical estimate is actually of plant area index, as we have not removed trunks and branches from the gap fraction calculation. However, this does not appear to have a great impact on results as hemispherical photograph derived LAI is still the lowest estimate of all three.

4.5 Point-centred quater observations

We used the method of Point-Centred Quarters (PCQ) (Dahdouh-Guebas and Koedam, 2006) to determine an estimate of the woody biomass for both unthinned and thinned forest in the Straits Inclosure. The PCQ method is conducted at each sampling point as follows:

- Using a compass, map 4 regions from the central sampling point
- Measure the distance from the central sampling point to the nearest tree in each quarter
- Measure the Diameter at Breast Height (DBH) for each tree (shown in Figure 4.10) and record

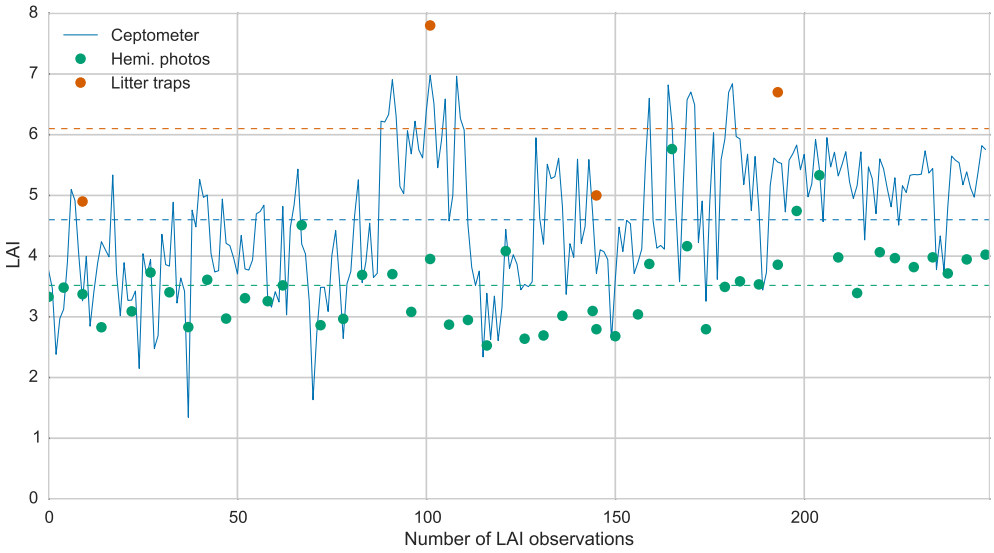


Figure 4.8: LAI comparison for unthinned forest. Dots and solid line represent observations made at different points along transects, dotted lines represent the mean of the observations.

the species

There were 114 points samples along the three transects, from these measurements we derived estimates to tree density and mean DBH for both thinned and unthinned sides of the forest. We then used allometric relationships between DBH and total above ground biomass and coarse root biomass, found in work carried out by Forest Research and in McKay et al. (2003). These relationships were as follows,

$$\text{above ground dry-mass} = 0.0678 \times \overline{\text{DBH}}^{2.619} \tag{4.2}$$

and

$$\text{below ground coarse root dry-mass} = 0.149 \times \overline{\text{DBH}}^{2.12}. \tag{4.3}$$

This gave us an estimate to the dry-mass in kilograms for the average tree in our sampling area. Assuming that half of all dry-mass is carbon we can find an estimate of total woody and coarse root carbon in g C m^{-2} using the equation,

$$\begin{aligned}&\text{total woody and coarse root carbon} = \\ &1000 \times 0.5 \times (\text{above ground dry-mass} + \text{below ground coarse root dry-mass}) \times \text{tree density}.\end{aligned} \tag{4.4}$$

Forest Research have carried out their own mensuration studies at the site, these have been conducted at the mensuration points shown in Figure 4.2. As these plots are included in our

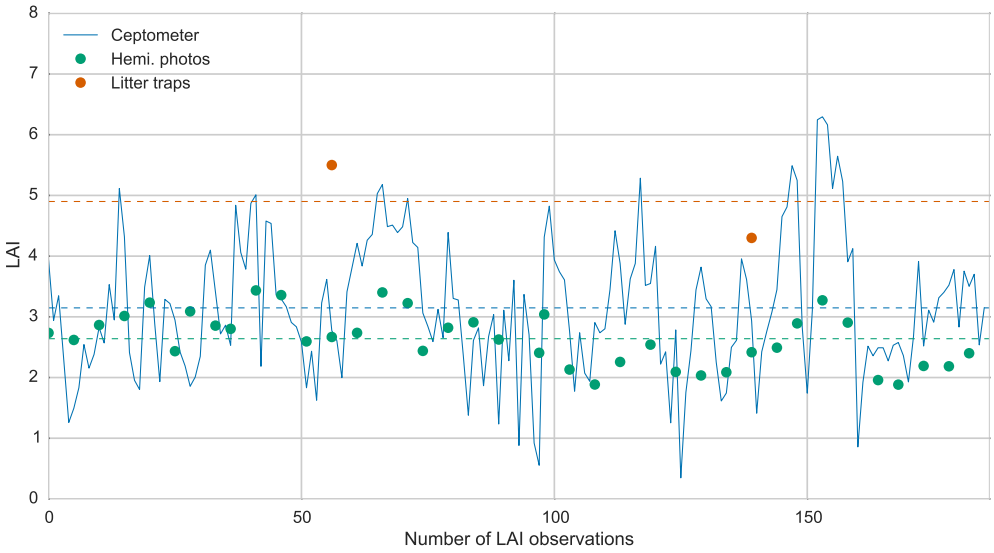


Figure 4.9: LAI comparison for thinned forest. Dots and solid line represent observations made at different points along transects, dotted lines represent the mean of the observations.

transects this means that hopefully our measurements will be comparable with those from Forest Research.

4.6 Flux tower observations and data processing

Forest Research provided half-hourly raw flux tower data for the Straits Inclosure from January 1999 to December 2015. These consist of the NEE fluxes and meteorological driving data of temperature, irradiance and atmospheric CO_2 concentration for use in the DALEC model. The view from the top of the flux tower in the Straits Inclosure can be seen in Figure 4.11. Forest Research provided this data in the form of multiple excel spreadsheets corresponding to the flux tower measurement record for each year. To prepare this data for use with data assimilation we first had to convert these 16 excel files to one Python readable data file (here we chose NetCDF), this was the further processed. To process the NEE data we first performed u^* filtering, where any half-hourly flux observation corresponding to a friction velocity of 0.2 m s^{-1} (this value represents the point at which flux measurements become unreliable and was found by Forest Research) or less were removed from the data set. We then subjected the observations of NEE to quality control procedures similar to those described by Papale et al. (2006a). For each year of the NEE dataset this procedure involved calculating the standard deviation of both the positive and negative half-hourly observations and then removing any values that were ± 3 standard deviations away from the yearly positive/negative mean. This was also repeated on a month by month basis. Gap-filling

Figure 4.10: Taking diameter at breast height measurements at Alice Holt.

procedures were not applied to the half-hourly NEE dataset so that only true observations were considered for assimilation. To match the time-step of the DALEC model we computed daily NEE observations by taking the mean over the 48 measurements made each day, selecting only days where there was no missing data.

Figure 4.11: At the top of the Alice Holt flux tower.

Chapter 5

Information content in observations relevant to forest carbon balance

5.1 Introduction

In data assimilation it is important to understand if the observations available to us provide us with enough information to find a meaningful description of our studied system. Measurements of forest carbon balance are now routinely made in forests across the world using micrometeorological techniques, with many other relevant observations such as leaf area index and standing biomass also available (Baldocchi, 2008). Many efforts have been made to combine this data with models of forest carbon balance using data assimilation techniques in order to improve modelled estimates (Zobitz et al., 2011; Fox et al., 2009; Richardson et al., 2010; Quaife et al., 2008; Zobitz et al., 2014b; Niu et al., 2014). Currently, however, the relative levels of information from different data types is not well understood.

In numerical weather prediction many measures of observation information content have been defined (Cardinali et al., 2004; Rodgers et al., 2000; Fisher, 2003). These measures can be used to identify how information content might vary both temporally and spatially, when observations are made at different times or in different locations. It is not necessary to have made a physical observation in order to estimate its information content. It is enough to have an accurate estimate to the observation error at the specified time and location. It is therefore possible to use these measures to define target observations or design new observing systems (Palmer et al., 1998;

Eyre, 1990). Often we are required to know the derivative of our model in order to implement these measures. This can prove difficult to implement. However, techniques such as automatic-differentiation (Renaud, 1997) can reduce the time taken to find the derivative of a model.

In this chapter we aim to analyse the information content in the observations used for assimilation with the DALEC1 and DALEC2 models of ecosystem carbon balance. We begin by considering the observability of our system given a set of observations. Observability is a mathematical concept from control theory. Applied to data assimilation a system is defined as observable if for a given set of observations we can uniquely define the initial state for our model. This allows us to determine if from current observations used in carbon balance data assimilation, we have enough information to find a unique model state. In practice we include a background term in our assimilation (ref. DA section) to ensure we can always find a locally unique solution. However, it is informative to show that observations alone provide us with enough information to find a unique solution.

We then consider different information content measures applied to our system in order to show how the information content varies for the different observation types available to us for DALEC1. We then extend these results to the DALEC2 model and investigate how the same set of observations can have a different level of information depending on the type of ecosystem being observed. Using these measures also allows us to consider the effect of including error correlations in our data assimilation algorithm, as explored in section (ref. 1st results chapter), on the information content in the observations.

5.2 Background material

5.2.1 Metolius forest site

In this chapter we use meteorological driving data taken from the Metolious forest site, a temperate coniferous forest in Oregon, Northwestern US. The site has been studied extensively (Law et al., 2001a), and has also been the subject of data assimilation studies (Williams et al., 2005; Quaife et al., 2008). The site has a semi-arid climate with a dominant canopy of ponderosa pine (*Pinus ponderosa*) and an understory of bitterbush (*Purshia tridentata*) and manzanita (*Arctostaphylos patula*) (Law et al., 2001b). The forest stand was felled in 1978, having previously been a mature forest. It was then allowed to regrow naturally, with some areas of older growth forest still being

left post-felling (Williams et al., 2005).

5.2.2 Observability

Observability is a mathematical concept from control theory. A system is said to be observable if it is possible to determine the state by measuring only the output. The following definition is taken from Barnett and Cameron (1985): for the linear time varying system defined as,

$$\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t) \tag{5.1}$$

$$\mathbf{y} = \mathbf{C}(t)\mathbf{x}(t) \tag{5.2}$$

where $\mathbf{A}$ is $n \times n$, $\mathbf{B}$ is $n \times m$ and $\mathbf{C}$ is $r \times n$ is *completely observable* if for any t_0 and any initial state $\mathbf{x}(t_0) = \mathbf{x}_0$ there exists a finite time $t_i > t_0$ such that knowledge of $\mathbf{u}(t)$ and $\mathbf{y}(t)$ for $t_0 \leq t \leq t_i$ suffices to uniquely determine $\mathbf{x}_0$. There is no loss of generality in assuming $\mathbf{u}(t)$ is identically zero throughout the whole interval; this is the case for data assimilation. So that in data assimilation our system is *completely observable* if knowledge of the observations $\mathbf{y}(t)$ allows us to uniquely determine the initial state $\mathbf{x}_0$.

Theorem 1. *When $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are time-invariant the system is completely observable if and only if the $nr \times n$ observability matrix*

$$\mathbf{V} = \begin{pmatrix} \mathbf{C} \\ \mathbf{CA} \\ \mathbf{CA}^2 \\ \vdots \\ \mathbf{CA}^{n-1} \end{pmatrix} \tag{5.3}$$

has rank n.

This result can be applied to the data assimilation problem (Johnson et al., 2005), where for 4D-Var the observability matrix corresponds to

$$\hat{\mathbf{H}} = \begin{pmatrix} \mathbf{H}_0 \\ \mathbf{H}_1\mathbf{M}_0 \\ \vdots \\ \mathbf{H}_N\mathbf{M}_{N,0} \end{pmatrix} \tag{5.4}$$

as defined in section (ref. DA section). In Appendix B of Zou et al. (1992) it is shown that for the linear data assimilation problem it is possible to obtain a unique analysis state over a specific assimilation window with no background term if the rank of $\hat{\mathbf{H}}$ is equal to n, the size of $\mathbf{x}_0$. For the non-linear data assimilation problem the rank of $\hat{\mathbf{H}}$ being equal to n ensures a locally unique analysis can be found without including a background term. In practice the cost function for 4D-Var data assimilation typically contains a background term which regularises the problem and means that we always have a unique solution.

5.2.3 Information content measures

In data assimilation we combine prior estimates with observations to improve our knowledge of the state of a system. In this process some observations will have a greater impact on our results than others. Many measures exist for understanding observation impact on the assimilation results.

Information content measures have been used to quantify the different levels of information provided by observations in the development of satellite instruments (Stewart et al., 2008; Engelen and Stephens, 2004) and in operational data assimilation schemes (Fisher, 2003; Singh et al., 2013). In Fowler and Van Leeuwen (2013) it is discussed that in these operational schemes information content measures have been used for

- Removing observations with a lesser impact in order to improve the efficiency of the assimilation process (Rabier et al., 2002; Singh et al., 2013; Rodgers, 1998).
- Diagnosing erroneous observations and assumed statistics (Desroziers et al., 2009).
- Improving data assimilation results by adding observations which theoretically have a high impact. This can mean defining target observations (Palmer et al., 1998) or even designing new observing systems (Wahba, 1985; Eyre, 1990).

For the following measures the data assimilation problem is assumed to be Gaussian with a linear function mapping the state to observation space ($\mathbf{H}$), following the derivation from Kalnay (2003) we have,

$$\mathbf{x}_a = \mathbf{x}_b + \mathbf{K}(\mathbf{y} - \mathbf{H}\mathbf{x}_b), \tag{5.5}$$

where $\mathbf{K}$ is the Kalman gain matrix,

$$\mathbf{K} = (\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} + \mathbf{B}^{-1})^{-1}\mathbf{H}^T\mathbf{R}^{-1}. \tag{5.6}$$

In order to consider observations over a 4D-Var time window we rewrite equation (5.5) as,

$$\mathbf{x}_a = \mathbf{x}_b + \hat{\mathbf{K}}(\hat{\mathbf{y}} - \hat{\mathbf{H}}\mathbf{x}_b), \tag{5.7}$$

using the defined matrices in section (ref DA section), with $\hat{\mathbf{K}} = (\hat{\mathbf{H}}^T\hat{\mathbf{R}}^{-1}\hat{\mathbf{H}} + \mathbf{B}^{-1})^{-1}\hat{\mathbf{H}}^T\hat{\mathbf{R}}^{-1}$.

Making the assumption of a linear and Gaussian data assimilation problem is clearly a limitation. These measures are therefore limited to a period where the forecast model remains reasonably linear. The implications of assuming Gaussian error statistics are discussed in Fowler and Van Leeuwen (2013).

5.2.3.1 Sensitivity of analysis to observations

The influence matrix measures the sensitivity of the analysis in observation space to the observations (Cardinali et al., 2004) and is defined by,

$$\mathbf{S} = \frac{\partial \mathbf{H}\mathbf{x}_a}{\partial \mathbf{y}}. \tag{5.8}$$

From equation (5.5) we see that,

$$\begin{aligned} \mathbf{S} &= \frac{\partial \mathbf{H}(\mathbf{x}_b + \mathbf{K}(\mathbf{y} - \mathbf{H}\mathbf{x}_b))}{\partial \mathbf{y}} \\ &= \mathbf{K}^T\mathbf{H}^T, \end{aligned} \tag{5.9}$$

here $\mathbf{S}$ will be a $p \times p$ matrix, where p is the number of observations. The diagonal elements of $\mathbf{S}$ are $\mathbf{S}_{i,i} = \frac{\partial(\mathbf{H}\mathbf{x}_a)_i}{\partial \mathbf{y}_i}$ and represent the 'self-sensitivity' of the i^{th} modelled observation to the i^{th} observation. The off-diagonal elements of $\mathbf{S}$ represent the 'cross-sensitivity' and are given by $\mathbf{S}_{i,j} = \frac{\partial(\mathbf{H}\mathbf{x}_a)_i}{\partial \mathbf{y}_j}$. If we wish to consider the influence matrix for observations over a 4D-Var time window we can re-write equation (5.8) as,

$$\mathbf{S} = \frac{\partial \hat{\mathbf{H}}\mathbf{x}_a}{\partial \hat{\mathbf{y}}} = \hat{\mathbf{K}}^T\hat{\mathbf{H}}^T. \tag{5.10}$$

The Kalman gain matrix $\hat{\mathbf{K}}$ can be re-written as,

$$\hat{\mathbf{K}} = \mathbf{A}\hat{\mathbf{H}}^T\hat{\mathbf{R}}^{-1}, \tag{5.11}$$

where $\mathbf{A}$ is the analysis error covariance,

$$\mathbf{A} = (\hat{\mathbf{H}}^T\hat{\mathbf{R}}^{-1}\hat{\mathbf{H}} + \mathbf{B}^{-1})^{-1}. \tag{5.12}$$

Inserting equation (5.11) into (5.10) we find,

$$\mathbf{S} = \hat{\mathbf{R}}^{-1}\hat{\mathbf{H}}\mathbf{A}\hat{\mathbf{H}}^T. \tag{5.13}$$

We can therefore see the sensitivity of the analysis to observations is inversely proportional to the observation error and proportional to the analysis error. This means that the most influential observations are those with the smallest error variance providing information about regions of state space with the largest prior error (Cardinali et al., 2004). It is possible to identify the observations that have the greatest influence over the length of the window by summing the absolute values of the columns of the influence matrix.

5.2.3.2 Degrees of freedom for signal

The degrees of freedom for signal (dfs) indicates the number of elements of the state that have been measured by the observations. If we consider a state vector $\mathbf{x}$ with n elements (or n degrees of freedom) then the maximum value the dfs could obtain would be n, in this case all elements of the state would have been measured. Conversely if $dfs = 0$ then no elements of the state would have been measured by our observations (Fowler and Van Leeuwen, 2013).

For a symmetric positive definite prior and posterior error covariance matrices $\mathbf{B}$ and $\mathbf{A}$, we can define the degrees of freedom for signal by means of a transform $\mathbf{L}$ that reduces the prior error covariance matrix, $\mathbf{B}$ to the $n \times n$ identity (Fisher, 2003). Each diagonal element of the transformed matrix $\mathbf{B}$ then corresponds to a single degree of freedom with the trace being equal to n, the total degrees of freedom.

The transform $\mathbf{L}$ can also be represented by $\mathbf{Q}^T\mathbf{L}$, where $\mathbf{Q}$ is an orthogonal matrix. So that $\mathbf{Q}^T\mathbf{L}\mathbf{B}\mathbf{L}^T\mathbf{Q} = \mathbf{Q}^T\mathbf{Q} = \mathbf{I}_{n\times n}$. By defining $\mathbf{Q}$ to be the matrix of the eigenvectors of $\mathbf{L}\mathbf{A}\mathbf{L}^T$, we reduce

$\mathbf{B}$ to the identity and $\mathbf{LAL}^T$ to the diagonal matrix of its eigenvalues, $\mathbf{\Lambda}$. The eigenvalues λ_i of $\mathbf{LAL}^T$ can be interpreted as the fractional reduction in uncertainty for the n state members. If an eigenvalue is close to zero the corresponding state member has been well observed, if it is close to one the corresponding state member has not been constrained by the assimilated observations (Stewart et al., 2008). We then define the degrees of freedom for signal as,

$$\begin{aligned} dfs &= \text{trace}(\mathbf{Q}^T\mathbf{LBL}^T\mathbf{Q} - \mathbf{Q}^T\mathbf{LAL}^T\mathbf{Q}) \\ &= \text{trace}(\mathbf{I}_{n\times n} - \mathbf{\Lambda}) \\ &= n - \text{trace}(\mathbf{\Lambda}) \\ &= n - \text{trace}(\mathbf{LAL}^T) \\ &= n - \text{trace}(\mathbf{B}^{-1}\mathbf{A}). \end{aligned} \tag{5.14}$$

In Rodgers et al. (2000) it is shown that the dfs can also be calculated as the trace of the influence matrix $\mathbf{S}$ (defined in section 5.2.3.1) with,

$$dfs = \text{trace}(\mathbf{S}) = \sum_i \lambda_i, \tag{5.15}$$

where λ_i is the i^{th} eigenvalue of $\mathbf{S}$.

5.2.3.3 Shannon information content

Shannon Information Content (SIC) is a measure of the reduction in entropy (uncertainty) given a set of observations. When a measurement is made, the entropy or uncertainty in our state decreases. The SIC of an observation is a measure of the factor by which the uncertainty decreases (Cover and Thomas, 1991). We can define this using the prior, $p(\mathbf{x})$, and posterior, $p(\mathbf{x}|\mathbf{y})$, distributions. From Rodgers et al. (2000), for the Gaussian case SIC unsprisingly becomes a function of the prior and posterior error covariance matrices with,

$$\text{SIC} = \frac{1}{2}\ln\frac{\left|\mathbf{B}\right|}{\left|\mathbf{A}\right|}. \tag{5.16}$$

The SIC can also be defined in terms of the eigenvalues of the influence matrix **S** with,

$$\text{SIC} = -\frac{1}{2}\sum_i ln|1-\lambda_i| \tag{5.17}$$

where λ_i is the i^{th} eigenvalue of **S**. In Eyre (1990) using SIC is shown to be beneficial over solely measuring the change in error variances before and after assimilation as the SIC also uses information about the change in error covariances. This is also true for the dfs.

5.3 Results

5.3.1 DALEC1 observability

DALEC1 is the original version of the DALEC2 model introduced in section (ref. D2 section). At the start of the PhD project work was undertaken with DALEC1 before the DALEC2 model was released. The version of DALEC1 used was an evergreen only model; further details of the model can be found in section (ref. D1 section) and Williams et al. (2005).

We initially consider observability of the DALEC1 state estimation system. DALEC1 is a smaller model and allows us to understand the concept of observability before moving onto work with the more complicated DALEC2 joint state and parameter estimation system. DALEC1 was implemented in a 4D-Var data assimilation scheme for state estimation, with the tangent linear model being computed analytically. Using this analytic implementation of the tangent linear model we can compute the observability of the model for differing sets of observations. We have the tangent linear model,

$$\mathbf{M}_i = \frac{\partial \mathbf{m}_{i-1\rightarrow i}(\mathbf{x}_i)}{\partial \mathbf{x}_i} =$$

$$\begin{pmatrix}
(1-\theta_{fol})+f_{fol}(1-f_{auto})\zeta^i & 0 & 0 & 0 & 0 \\
f_{roo}(1-f_{fol})(1-f_{auto})\zeta^i & (1-\theta_{roo}) & 0 & 0 & 0 \\
(1-f_{roo})(1-f_{fol})(1-f_{auto})\zeta^i & 0 & (1-\theta_{woo}) & 0 & 0 \\
\theta_{fol} & \theta_{roo} & 0 & (1-(\theta_{min}+\theta_{lit})\chi^{i-1}) & 0 \\
0 & 0 & \theta_{woo} & \theta_{min}\chi^{i-1} & (1-\theta_{som}\chi^{i-1})
\end{pmatrix}, \tag{5.18}$$

where $\mathbf{x}_i = (C^i_{fol}, C^i_{roo}, C^i_{woo}, C^i_{lit}, C^i_{som})^T$, $\zeta^i = \partial GPP^i(C^{i-1}_{fol}, \Psi)/\partial C^{i-1}_{fol}$ and $\chi^{i-1} = e^{\Theta T^{i-1}}$ with the parameters and symbols having the same meaning as in section (ref. D2 section).

We can use the linearised model with the linearised observation operator $\mathbf{H}_i$ to form the matrix in equation (5.4) and compute the observability for a specific set of observations over a finite window. We will need at least 5 observations of any type for the system to be observable as the state $\mathbf{x}_0$ is of size 5 in the DALEC1 state estimation case. We first consider the observability for 5 observations of LAI. For DALEC1 LAI takes the form

$$LAI^i = \frac{C^i_{fol}}{c_{lma}}. \tag{5.19}$$

We then have the linearised observation operator

$$\mathbf{H}_i = \frac{\partial LAI^i}{\partial \mathbf{x}_i} = \begin{pmatrix} \frac{1}{c_{lma}} & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{5.20}$$

Using the linearised observation operator and the linear model from equation (5.18) we can compute $\hat{\mathbf{H}}$ for 5 observations of LAI on consecutive time steps

$$\hat{\mathbf{H}} = \begin{pmatrix} \mathbf{H}_0 \\ \mathbf{H}_1\mathbf{M}_0 \\ \vdots \\ \mathbf{H}_4\mathbf{M}_{3,0} \end{pmatrix} = \begin{pmatrix} \frac{1}{c_{lma}} & 0 & 0 & 0 & 0 \\ \frac{1}{c_{lma}}((1-\theta_{fol}) + f_{fol}(1-f_{auto})\zeta^0) & 0 & 0 & 0 & 0 \\ \frac{1}{c_{lma}}\prod_{i=0}^{1}((1-\theta_{fol}) + f_{fol}(1-f_{auto})\zeta^i) & 0 & 0 & 0 & 0 \\ \frac{1}{c_{lma}}\prod_{i=0}^{2}((1-\theta_{fol}) + f_{fol}(1-f_{auto})\zeta^i) & 0 & 0 & 0 & 0 \\ \frac{1}{c_{lma}}\prod_{i=0}^{3}((1-\theta_{fol}) + f_{fol}(1-f_{auto})\zeta^i) & 0 & 0 & 0 & 0 \end{pmatrix}, \tag{5.21}$$

so that no matter how many observations of LAI we add, our system will not be observable as the rows of $\hat{\mathbf{H}}$ are all linearly dependant, so that $\hat{\mathbf{H}}$ in this case has rank 1. We can repeat this for different observations to see for which observation types our system is observable.

From figure 5.1 we can see that our system is observable for 5 observations of the soil and organic matter carbon pool C_{som}. In figure 5.1 we have shown results for the rank of $\hat{\mathbf{H}}$ when we have 5 observations in each case; this has also been tested with increasing numbers of observations being added to the system with the results from figure 5.1 remaining unchanged.

The system being observable for observations of C_{som} physically makes sense as all the carbon in the system that is not respired to the atmosphere eventually ends up in the soil and organic matter carbon pool, C_{som}, so that by taking observations of this pool we observe all the others. In a

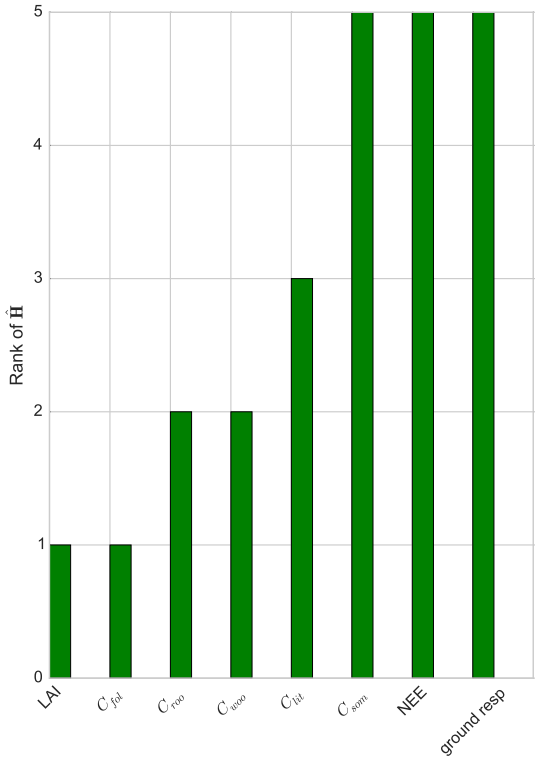


Figure 5.1: Rank of the observability matrix $\hat{\mathbf{H}}$ for 5 observations of different types. The ranks shown here are computed analytically using SymPy (Joyner et al., 2012).

similar way $\hat{\mathbf{H}}$ is also full rank for observations of NEE and ground respiration. We can see from the form of these observations in DALEC1 that they both contain indirect observations of C_{som} with NEE taking the form

$$NEE^i = -(1-f_{auto})GPP^i(C_{fol}^{i-1},\Psi) + \theta_{lit}C_{lit}e^{\Theta T^i} + \theta_{som}C_{som}e^{\Theta T^i} \tag{5.22}$$

with a corresponding linearised observation operator

$$\mathbf{H}_i = \frac{\partial NEE^i}{\partial \mathbf{x}_i} = \begin{pmatrix} -(1-f_{auto})\zeta^i & 0 & 0 & \theta_{lit}e^{\Theta T^i} & \theta_{som}e^{\Theta T^i} \end{pmatrix}, \tag{5.23}$$

and for ground respiration

$$G_{resp}^i = \frac{1}{3}f_{auto}GPP^i(C_{fol}^{i-1},\Psi) + \theta_{lit}C_{lit}e^{\Theta T^i} + \theta_{som}C_{som}e^{\Theta T^i} \tag{5.24}$$

(here we have assumed the fraction of total autotrophic respiration from below ground to be $\frac{1}{3}$)

with a corresponding linearised observation operator

$$\mathbf{H}_i = \frac{\partial G^i_{resp}}{\partial \mathbf{x}_i} = \begin{pmatrix} \frac{1}{3} f_{auto} \zeta^i & 0 & 0 & \theta_{lit} e^{\Theta T^i} & \theta_{som} e^{\Theta T^i} \end{pmatrix}. \tag{5.25}$$

At flux tower sites NEE is the most observed quantity, these results give us confidence that we can construct a unique solution when working with flux tower data. We will further explore the concept of observability for the joint parameter and state estimation case with DALEC2 in section 5.3.2.

5.3.2 DALEC2 observability

For DALEC2 we perform joint parameter and state estimation and have an augmented state of size $n = 23$. The augmented state is made up of the 6 carbon pool state members and 17 model parameters as described in section (ref. D2 section). As we are also estimating the parameters of DALEC2 the concept of observability for our system is closely linked to the concept of identifiability (Navon, 1998). A system is identifiable if given observations of the state variables and knowledge of the model dynamics it is possible to obtain a unique deterministic set of model parameter values (Ljung, 1998). If a model parameter is not observable it will not be identifiable (Jacquez and Greif, 1985). It is therefore useful to compute the observability of the DALEC2 joint parameter and state estimation system.

We compute observability in the same way as in section 5.3.1 by finding the rank of $\hat{\mathbf{H}}$ for a given set of observations. For the state and parameter estimation case we cannot compute the observability of the system analytically, it is therefore important to check that the numerical calculation of the rank of $\hat{\mathbf{H}}$ for DALEC1 is equal to the rank when calculated analytically. This will give us confidence that our implementation of the numeric rank is correct for DALEC2 when applied to a well-conditioned problem as the implementation is the same in both cases. We have tested our numeric implementation for the state estimation case with DALEC1 and find the same results for the rank of $\hat{\mathbf{H}}$ as for the analytic case, as shown in table 5.1. We calculate the rank of the $\hat{\mathbf{H}}$ matrix using a singular value decomposition (SVD) which can have issues if the condition number of $\hat{\mathbf{H}}$ is large (Paige, 1981). This is a problem we encounter in the DALEC2 case when trying to calculate the rank of $\hat{\mathbf{H}}$ directly.

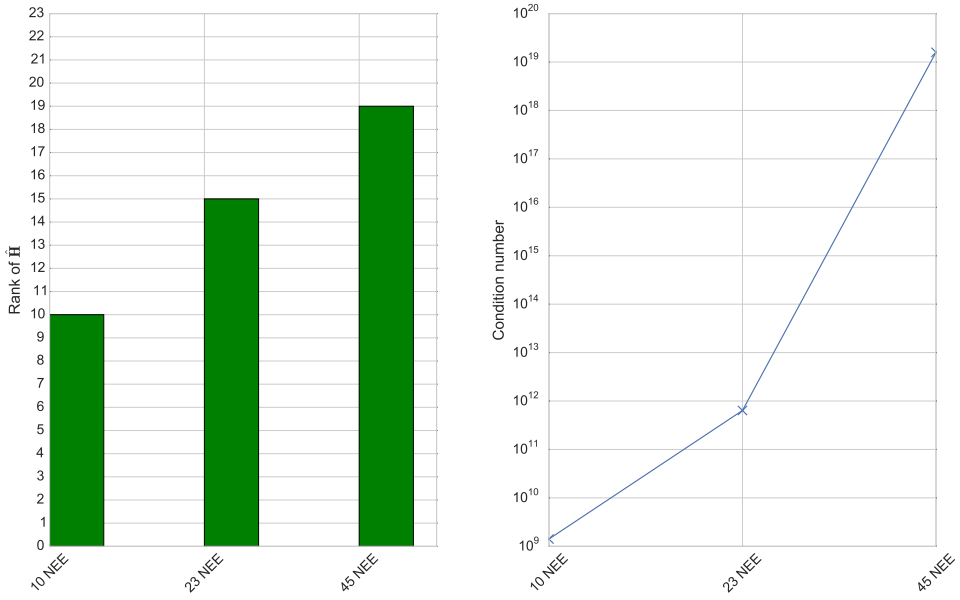
Figure 5.2 highlights the problems we have calculating the rank of the $\hat{\mathbf{H}}$ matrix for the

Observation	Rank of $\hat{\mathbf{H}}$ (numeric)	Rank of $\hat{\mathbf{H}}$ (analytic)
LAI	1	1
C_{fol}	1	1
C_{roo}	2	2
C_{woo}	2	2
C_{lit}	3	3
C_{som}	5	5
NEE	5	5
G_{resp}	5	5

Table 5.1: Rank of $\hat{\mathbf{H}}$ for 5 observations of different types for both numeric and analytic implementations with DALEC1.

DALEC2 joint parameter and state estimation case. In figure 5.2a we see that for 23 observations of NEE our system is unobservable as we have a rank deficient $\hat{\mathbf{H}}$. However, we cannot trust the rank calculation of $\hat{\mathbf{H}}$ in this case. Figure 5.2b shows that for 23 observations of NEE, $\hat{\mathbf{H}}$ has a condition number in the order of 10^{19}. The condition number of a matrix corresponds to the ratio of the largest to the smallest singular values. A condition number of this size means that we have very small singular values. In the calculation of the rank of a matrix using an SVD we define the rank to be the number of singular values greater than the threshold `tol = max(S) * max(n, m) * eps` (Press et al., 2007), where `S` is the vector of singular values, `n` and `m` are the rows and columns of the matrix whose rank we wish to calculate and `eps` is the machine accuracy for the datatype of `S` (In this case a double-precision float with `eps = 2.22e-16`). For 23 observations of NEE, $\hat{\mathbf{H}}$ is classed as being rank deficient as `tol = 1.02e-10` and the three smallest singular values of $\hat{\mathbf{H}}$ are `[1.39e-11, 7.84e-15, 1.46e-15]` but here we are working past the accuracy of the computer and so cannot have confidence that $\hat{\mathbf{H}}$ is rank deficient in this case.

In order to address the problem of ill-conditioning of the $\hat{\mathbf{H}}$ matrix we can instead calculate the rank of the control variable transform observability matrix, $\hat{\mathbf{R}}^{-1/2}\hat{\mathbf{H}}\mathbf{D}^{1/2}$, where the symbols have the same meaning as in section (ref DA CVT section, $\mathbf{D} = diag\{\mathbf{B}\}$). The rank of $\hat{\mathbf{R}}^{-1/2}\hat{\mathbf{H}}\mathbf{D}^{1/2}$ and $\hat{\mathbf{H}}$ are the same since $\hat{\mathbf{R}}$ and $\mathbf{D}$ are both full rank matrices. The results using this new better conditioned matrix are shown in Figure 5.3. From Figure 5.3b we can see this matrix is much better conditioned than $\hat{\mathbf{H}}$, and for 23 observations of NEE we now have an observable system. Although the condition numbers here are still large we can have more confidence in these results as we are working within the precision of the computer.



(a) Rank of $\hat{\mathbf{H}}$ (b) Condition number of $\hat{\mathbf{H}}$

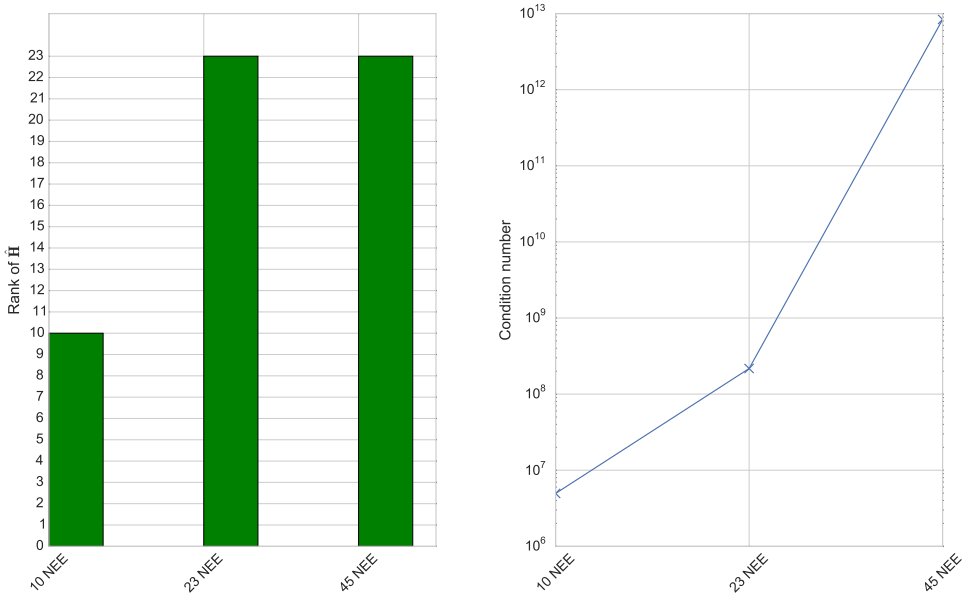
Figure 5.2: Observability of DALEC2 for $\hat{\mathbf{H}}$ with an increasing number of NEE observations displayed alongside the condition number for the $\hat{\mathbf{H}}$ matrices.

5.3.2.1 Observability for observations randomly distributed through time

In the previous experiments we have considered increasing numbers of NEE observations taken on adjacent days. It is also useful to consider the observability of the system when we have a number of observations randomly distributed throughout a time window. This is more consistent with what we expect from the real data we have to work with.

Figure 5.4 shows the observability for an increasing number of observations distributed through a 1 year assimilation window. I this case we are using the matrix $\hat{\mathbf{H}}$ and not the CVT observability matrix. In figure 5.4 we see that having the observations randomly distributed throughout a 1 year assimilation window has improved the conditioning of $\hat{\mathbf{H}}$ in comparison to figure 5.2. This is due to the observations being randomly distributed rather than adjacent. The rows of $\hat{\mathbf{H}}$ are more distinct when being evolved to different times in the year by the tangent linear model rather than evolved to adjacent days only. However, we still have a rank deficient $\hat{\mathbf{H}}$ for the 23 NEE observation case. From figure 5.4b we see that this is the case where the condition number peaks. As we add more randomly distributed observations the condition number of $\hat{\mathbf{H}}$ is reduced by an order of 10^2 and we have a full rank $\hat{\mathbf{H}}$.

In figure 5.5 we again see that using the CVT observability matrix has much improved the

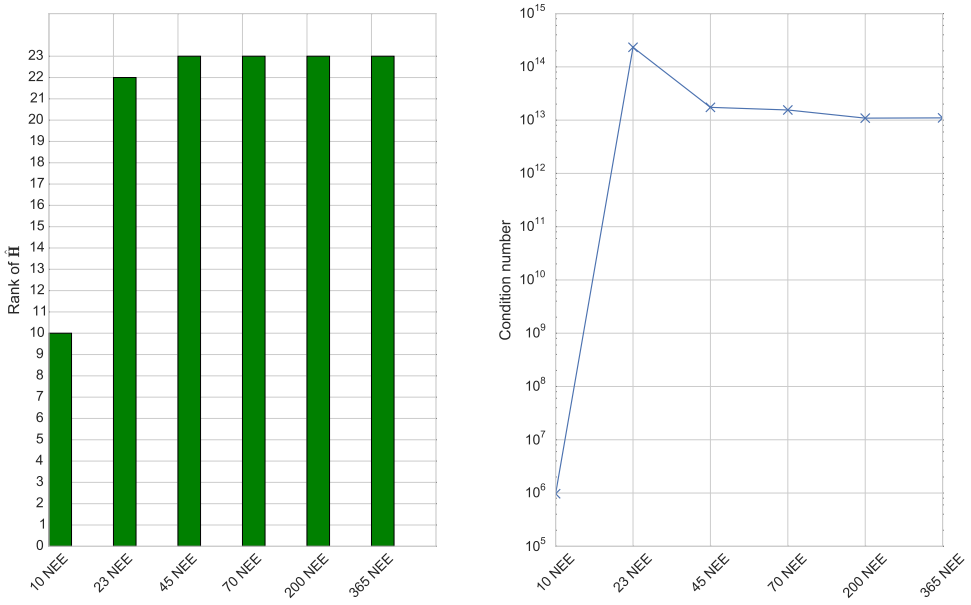


(a) Rank of $\hat{\mathbf{R}}^{-1/2}\hat{\mathbf{H}}\mathbf{D}^{1/2}$ (b) Condition number of $\hat{\mathbf{R}}^{-1/2}\hat{\mathbf{H}}\mathbf{D}^{1/2}$

Figure 5.3: Observability of the CVT DALEC2 for $\hat{\mathbf{R}}^{-1/2}\hat{\mathbf{H}}\mathbf{D}^{1/2}$ with an increasing number of NEE observations displayed alongside the condition number for the $\hat{\mathbf{R}}^{-1/2}\hat{\mathbf{H}}\mathbf{D}^{1/2}$ matrices.

conditioning of the problem in comparison to figure 5.4. We now see that the DALEC2 system is observable when we have 23 observations of NEE randomly distributed throughout the 1 year assimilation window. We have more confidence that this is the case as the condition numbers for the CVT observability matrix are almost half the values of those for $\hat{\mathbf{H}}$. We again see a similar pattern in figure 5.5 for the condition numbers with a peak for 23 NEE observations and then a reduction of order 10^2 when more observations are added.

We have tested the observability of the system for observations of NEE when we have different driving data, linearising around different states and with different distributions of observations throughout our assimilation window and in every case we have an observable system given an adequate number of NEE observations (at least 23). We can therefore have confidence that for the available data, typically 60-80 observations of daily NEE for any year's window, we can construct a unique solution with the observations alone.



(a) Rank of $\hat{\mathbf{H}}$ (b) Condition number of $\hat{\mathbf{H}}$

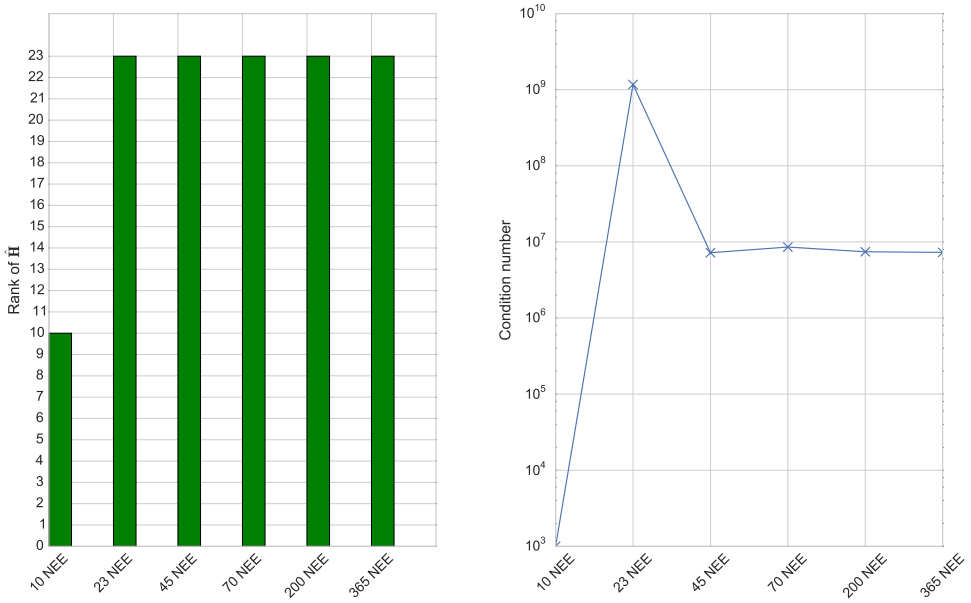
Figure 5.4: Observability of DALEC2 for a $\hat{\mathbf{H}}$ with an increasing number of NEE observations randomly distributed through a 1 year assimilation window (left). Condition number for the $\hat{\mathbf{H}}$ matrices (right).

5.3.3 DALEC1 information content

5.3.3.1 Information content for a single observation

For the DALEC1 state estimation we can calculate the analytic representation of the information content measures discussed in section 5.2.3. This will allow us to understand how the information content changes for differing numbers of observations, different observation types and the effect of including observation error correlations in the assimilation scheme, before moving onto work with the larger DALEC2 joint parameter and state estimation case. For these experiments the elements of the state vector have corresponding background standard deviations $\sigma_{cfol,b}, \sigma_{croo,b}, \sigma_{cwoo,b}, \sigma_{clit,b}, \sigma_{csom,b}$. We then have

$$\mathbf{B} = \begin{pmatrix} \sigma^2_{cfol,b} & 0 & 0 & 0 & 0 \\ 0 & \sigma^2_{croo,b} & 0 & 0 & 0 \\ 0 & 0 & \sigma^2_{cwoo,b} & 0 & 0 \\ 0 & 0 & 0 & \sigma^2_{clit,b} & 0 \\ 0 & 0 & 0 & 0 & \sigma^2_{csom,b} \end{pmatrix}. \tag{5.26}$$



(a) Rank of $\hat{\mathbf{R}}^{-1/2}\hat{\mathbf{H}}\mathbf{D}^{1/2}$ (b) Condition number of $\hat{\mathbf{R}}^{-1/2}\hat{\mathbf{H}}\mathbf{D}^{1/2}$

Figure 5.5: Observability of the CVT DALEC2 system for $\hat{\mathbf{R}}^{-1/2}\hat{\mathbf{H}}\mathbf{D}^{1/2}$ with an increasing number of NEE observations randomly distributed through a 1 year assimilation window (left). Condition number for the $\hat{\mathbf{R}}^{-1/2}\hat{\mathbf{H}}\mathbf{D}^{1/2}$ matrices (right).

We begin by considering the Shannon Information Content (SIC) and degrees of freedom for signal (dfs) for a single observation of LAI. We have the linearised observation operator

$$\mathbf{H}_i = \frac{\partial LAI^i}{\partial \mathbf{x}_i} = \frac{\partial}{\partial \mathbf{x}_i}\left(\frac{C^i_{fol}}{c_{lma}}\right) = \begin{pmatrix} \frac{1}{c_{lma}} & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{5.27}$$

As we have a single observation at one time, our observation error covariance matrix, $\mathbf{R}$, is just the variance of our observation of LAI at time t_0 ($\sigma^2_{LAI,o}$). Therefore,

$$\mathbf{R}_i = \sigma^2_{LAI,o}. \tag{5.28}$$

We then have from equation (ref. A matrix eqn),

$$\begin{aligned}
\mathbf{A} &= (\mathbf{J}'')^{-1} \\
&= (\mathbf{B}^{-1} + \hat{\mathbf{H}}^T \hat{\mathbf{R}}^{-1} \hat{\mathbf{H}})^{-1} \\
&= (\mathbf{B}^{-1} + \mathbf{H}_0^T \mathbf{R}_0^{-1} \mathbf{H}_0)^{-1} \\
&= \begin{pmatrix}
\frac{c_{lma}^2 \sigma_{LAI,o}^2 \sigma_{cfol,b}^2}{\sigma_{cfol,b}^2 + c_{lma}^2 \sigma_{LAI,o}^2} & 0 & 0 & 0 & 0 \\
0 & \sigma_{croo,b}^2 & 0 & 0 & 0 \\
0 & 0 & \sigma_{cwoo,b}^2 & 0 & 0 \\
0 & 0 & 0 & \sigma_{clit,b}^2 & 0 \\
0 & 0 & 0 & 0 & \sigma_{csom,b}^2
\end{pmatrix}.
\end{aligned} \tag{5.29}$$

We can now derive the SIC and dfs using equation (5.16) and (5.14) as,

$$\text{SIC} = \frac{1}{2} \ln \frac{|\mathbf{B}|}{|\mathbf{A}|} = \frac{1}{2} \ln \frac{(c_{lma}^2 \sigma_{LAI,o}^2 + \sigma_{cfol,b}^2)}{c_{lma}^2 \sigma_{LAI,o}^2} = \frac{1}{2} \ln \left(1 + \frac{\sigma_{cfol,b}^2}{c_{lma}^2 \sigma_{LAI,o}^2} \right) \tag{5.30}$$

and

$$dfs = n - tr(\mathbf{B}^{-1}\mathbf{A}) = 5 - \left(\frac{c_{lma}^2 \sigma_{LAI,o}^2}{(c_{lma}^2 \sigma_{LAI,o}^2 + \sigma_{cfol,b}^2)} + 4 \right) = 1 - \frac{c_{lma}^2 \sigma_{LAI,o}^2}{(c_{lma}^2 \sigma_{LAI,o}^2 + \sigma_{cfol,b}^2)}. \tag{5.31}$$

We see that in general for a direct observation of any of the carbon pools C we have

$$\text{SIC} = \frac{1}{2} \ln \left(1 + \frac{\sigma_{c,b}^2}{\sigma_{c,o}^2} \right) \tag{5.32}$$

and

$$dfs = 1 - \frac{\sigma_{c,o}^2}{(\sigma_{c,o}^2 + \sigma_{c,b}^2)}, \tag{5.33}$$

where $\sigma_{c,o}$ and $\sigma_{c,b}$ are the observation and background standard deviations respectively, corresponding to any of the 5 carbon pools. We see the SIC for a single observation of one of the carbon pools is dependent on the ratio between the observation and background variances. The carbon pool observation which will give us the highest SIC is the observation with the largest ratio $\frac{\sigma_{c,b}^2}{\sigma_{c,o}^2}$. This is also the case for dfs. Assuming a fixed background standard deviation, the carbon pool observation which will give us the highest information content is the pool which we can measure most accurately, as expected. From equations (5.30) and (5.31) for an observation of LAI the

information content is also dependent on c_{lma} the parameter describing leaf mass area.

Next we consider the information content in a single observation of NEE. We have

$$\mathbf{H}_i = \frac{\partial NEE^i}{\partial \mathbf{x}_i} = \begin{pmatrix} -(1-f_{auto})\zeta^i & 0 & 0 & \theta_{lit}e^{\Theta T^i} & \theta_{som}e^{\Theta T^i} \end{pmatrix} \tag{5.34}$$

and

$$\mathbf{R}_i = \sigma^2_{NEE,o}. \tag{5.35}$$

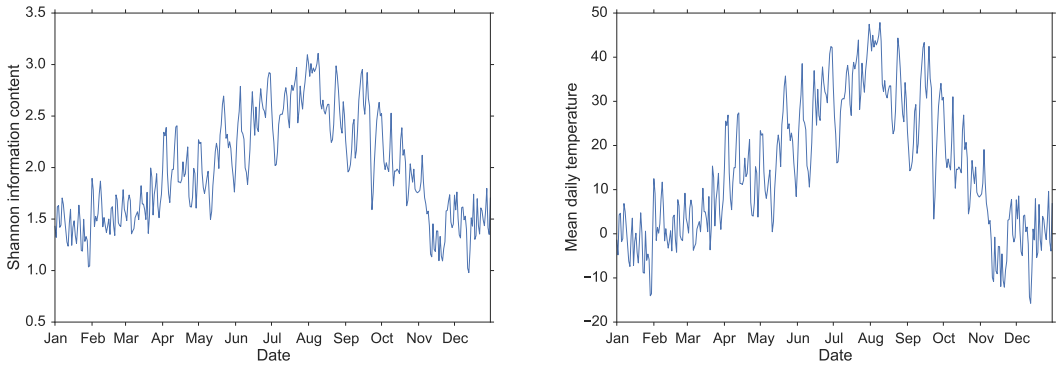
We then find

$$\text{SIC} = \frac{1}{2}\ln\left(1 + \frac{(f_{auto}-1)^2(\zeta^i)^2\sigma^2_{cfol,b} + (e^{\Theta T^i})^2(\theta^2_{som}\sigma^2_{csom,b} + \theta^2_{lit}\sigma^2_{clit,b})}{\sigma^2_{NEE,o}}\right) \tag{5.36}$$

and

$$dfs = 1 - \frac{\sigma^2_{NEE,o}}{(f_{auto}-1)^2(\zeta^i)^2\sigma^2_{cfol,b} + (e^{\Theta T^i})^2(\theta^2_{som}\sigma^2_{csom,b} + \theta^2_{lit}\sigma^2_{clit,b}) + \sigma^2_{NEE,o}}. \tag{5.37}$$

We see that Equations (5.36) and (5.37) have a similar form to Equations (5.32) and (5.33). The information content is again dependent on the ratio between the observation and background variances. The information content for the observations of NEE is also dependent on the magnitude of the first derivative of GPP with respect to C_{fol} and the magnitude of the exponential function of temperature controlling the rate of heterotrophic respiration, $e^{\Theta T^i}$. Both the first derivative of GPP and $e^{\Theta T^i}$ will be of greater magnitude when we have higher mean daily temperatures. This means that observations of NEE made at times with higher temperatures will have higher information content and more of an impact on data assimilation results.



(a) SIC for single NEE observation

(b) Mean daily temperature for year's data

Figure 5.6: SIC for a single NEE observation changing throughout a year's window using driving data from a forest of ponderosa pine in Oregon taken in 2007 (left). Mean daily temperature for the same site and period (right).

In Figure 5.6 we show how closely SIC is related to mean daily temperature for NEE observations throughout a year's window using daily driving data from a forest of ponderosa pine in Oregon (as described in section 5.2.1). Higher information content in summer observations of NEE makes physical sense. In summertime fluxes of carbon through the forest ecosystem are of greater magnitude than in winter, with more photosynthesis and respiration occurring. This gives us more information about the fluxes of carbon through our system in summertime observations of NEE. It is important to consider this result when planning for down time or routine maintenance at flux tower sites measuring NEE. The temperature dependence of information content will also hold true for other observations whose observation operators include the nonlinear temperature term controlling heterotrophic respiration. These observations include ground respiration, measured using soil respiration chambers, and total ecosystem respiration, estimated from nighttime NEE measurements.

In Figure 5.6a we have assumed constant prior and observation standard deviations. This is an accurate assumption for our prior errors. However, it has been shown that NEE errors are heteroscedastic (Richardson et al., 2008) and therefore scale with the magnitude of the flux. This would reduce the magnitude of the results shown in Figure 5.6a, as our standard deviation in observations of summer NEE would be larger, reducing the information content.

For Figure 5.6a we have used a numerical implementation in Python to calculate the SIC varying for 365 days of driving data. It is important to test our numerical implementation for correctness. In table 5.2 we show the SIC and dfs calculated both analytically and numerically. From this table we can see that both analytic and numerical implementations give us the same result to 15 or more significant figures. This gives us a degree of confidence that our implementation is also correct for DALEC2. In this table we have assumed constant prior and observation standard deviations for the carbon pools.

Obs.	SIC analytic value	SIC numeric value	dfs analytic value	dfs numeric value
NEE	0.0209343224569909	0.0209343224569913	0.0410042587324008	0.0410042587324008
C_{fol}	0.8047189562170501	0.8047189562170515	0.7999999999999998	0.7999999999999998
C_{roo}	0.1838623900626585	0.1838623900626572	0.3076923076923075	0.3076923076923083
C_{woo}	0.8047189562170501	0.8047189562170515	0.7999999999999998	0.7999999999999998
C_{lit}	0.1838623900626585	0.1838623900626572	0.3076923076923075	0.3076923076923074
C_{som}	0.1838623900626585	0.1838623900626572	0.3076923076923075	0.3076923076923074

Table 5.2: Correctness tests showing numeric and analytic values of information content calculated using 2007 driving data and parameter values from an Oregon ponderosa pine forest.

5.3.3.2 Information content for observations at a single time

We next consider the SIC when we have more than one observation at a single time. Here we will investigate the representation of information content when assimilating an observation of NEE with an observation of a carbon pool state member. We begin with a single observation of NEE and an observation of C_{fol}. We have the linearised observation operator,

$$\mathbf{H}_i = \frac{\partial}{\partial \mathbf{x}_i}\left(NEE^i, C^i_{fol}\right) = \begin{pmatrix} -(1-f_{auto})\zeta^i & 0 & 0 & \theta_{lit}e^{\Theta T^i} & \theta_{som}e^{\Theta T^i} \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{5.38}$$

and observation error covariance matrix

$$\mathbf{R}_i = \begin{pmatrix} \sigma^2_{NEE,o} & 0 \\ 0 & \sigma^2_{cfol,o} \end{pmatrix}. \tag{5.39}$$

We then find,

$$\text{SIC} = \frac{1}{2}\ln\left(1 + \frac{\sigma^2_{cfol,b}}{\sigma^2_{cfol,o}} + \frac{\xi^i}{\sigma^2_{NEE,o}} + \frac{\sigma^2_{cfol,b}(e^{\Theta T^i})^2(\theta^2_{som}\sigma^2_{csom,b} + \theta^2_{lit}\sigma^2_{clit,b})}{\sigma^2_{NEE,o}\sigma^2_{cfol,o}}\right) \tag{5.40}$$

where, $\xi^i = (f_{auto}-1)^2(\zeta^i)^2\sigma^2_{cfol,b} + (e^{\Theta T^i})^2(\theta^2_{som}\sigma^2_{csom,b} + \theta^2_{lit}\sigma^2_{clit,b})$. From equation (5.40) we can see that we have the first order terms for both NEE and C_{fol} as in equations (5.32) and (5.36). We also have a second order term for the combination of these observations. We can repeat this for the other carbon pools and find for $\mathbf{H}_i = \frac{\partial}{\partial \mathbf{x}_i}\left(NEE^i, C^i_{roo}\right)$,

$$\text{SIC} = \frac{1}{2}\ln\left(1 + \frac{\sigma^2_{croo,b}}{\sigma^2_{croo,o}} + \frac{\xi^i}{\sigma^2_{NEE,o}} + \frac{\sigma^2_{croo,b}\left((f_{auto}-1)^2(\zeta^i)^2\sigma^2_{cfol,b} + (e^{\Theta T^i})^2(\theta^2_{som}\sigma^2_{csom,b} + \theta^2_{lit}\sigma^2_{clit,b})\right)}{\sigma^2_{NEE,o}\sigma^2_{croo,o}}\right), \tag{5.41}$$

for $\mathbf{H}_i = \frac{\partial}{\partial \mathbf{x}_i}\left(NEE^i, C^i_{woo}\right)$,

$$\text{SIC} = \frac{1}{2}\ln\left(1 + \frac{\sigma^2_{cwoo,b}}{\sigma^2_{cwoo,o}} + \frac{\xi^i}{\sigma^2_{NEE,o}} + \frac{\sigma^2_{cwoo,b}\left((f_{auto}-1)^2(\zeta^i)^2\sigma^2_{cfol,b} + (e^{\Theta T^i})^2(\theta^2_{som}\sigma^2_{csom,b} + \theta^2_{lit}\sigma^2_{clit,b})\right)}{\sigma^2_{NEE,o}\sigma^2_{cwoo,o}}\right), \tag{5.42}$$

for $\mathbf{H}_i = \frac{\partial}{\partial \mathbf{x}_i}\left(NEE^i, C^i_{lit}\right)$,

$$\text{SIC} = \frac{1}{2}\ln\left(1 + \frac{\sigma^2_{clit,b}}{\sigma^2_{clit,o}} + \frac{\xi^i}{\sigma^2_{NEE,o}} + \frac{\sigma^2_{clit,b}\left((f_{auto}-1)^2(\zeta^i)^2\sigma^2_{cfol,b} + (e^{\Theta T^i})^2\theta^2_{som}\sigma^2_{csom,b}\right)}{\sigma^2_{NEE,o}\sigma^2_{clit,o}}\right) \tag{5.43}$$

and for $\mathbf{H}_i = \frac{\partial}{\partial \mathbf{x}_i}\left(NEE^i, C^i_{som}\right)$,

$$\text{SIC} = \frac{1}{2}\ln\left(1 + \frac{\sigma^2_{csom,b}}{\sigma^2_{csom,o}} + \frac{\xi^i}{\sigma^2_{NEE,o}} + \frac{\sigma^2_{csom,b}\left((f_{auto}-1)^2(\zeta^i)^2\sigma^2_{cfol,b} + (e^{\Theta T^i})^2\theta^2_{lit}\sigma^2_{clit,b}\right)}{\sigma^2_{NEE,o}\sigma^2_{csom,o}}\right). \tag{5.44}$$

Assuming constant prior and observation standard deviations across our carbon pool observations we see that the information content will be largest in equations (5.41) and (5.42). For both $\mathbf{H}_i = \frac{\partial}{\partial \mathbf{x}_i}\left(NEE^i, C^i_{roo}\right)$ and $\mathbf{H}_i = \frac{\partial}{\partial \mathbf{x}_i}\left(NEE^i, C^i_{woo}\right)$ we have an extra term in the numerator for our second order term corresponding to the combination of the two observations. If we consider the linearised observation operator for both these cases,

$$\mathbf{H}_i = \frac{\partial}{\partial \mathbf{x}_i}\left(NEE^i, C^i_{roo}\right) = \begin{pmatrix} -(1-f_{auto})\zeta^i & 0 & 0 & \theta_{lit}e^{\Theta T^i} & \theta_{som}e^{\Theta T^i} \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix} \tag{5.45}$$

and

$$\mathbf{H}_i = \frac{\partial}{\partial \mathbf{x}_i}\left(NEE^i, C^i_{woo}\right) = \begin{pmatrix} -(1-f_{auto})\zeta^i & 0 & 0 & \theta_{lit}e^{\Theta T^i} & \theta_{som}e^{\Theta T^i} \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}, \tag{5.46}$$

we can see that these observations provide an orthogonal constraint to the observation of NEE. Neither of these pools are observed with a single observation of NEE. The information content being greater when assimilating C_{roo} or C_{woo} alongside NEE is therefore expected.

In practice we cannot assume constant prior and observation errors across the different carbon pools. Root carbon is hard to measure accurately (Brown, 2002). However, woody biomass (C_{woo}) is regularly measured using mensuration (Husch et al., 2002) or point-centred quarter methods (Dahdouh-Guebas and Koedam, 2006) at forest sites. Advancements in Light Detection And Ranging (LiDAR) scanning (Lefsky et al., 1999) mean that we have increasingly more accurate observations of woody biomass. The European Space Agency BIOMASS mission (Le Toan et al., 2011) will also provide a much more abundant source of woody biomass measurements in the future. If we consider NEE to be the main observation currently used in ecosystem data assimilation, then the increasing number of available woody biomass measurements will benefit assimilation schemes greatly.

5.3.3.3 Information content in successive observations

In section 5.3.3.1 we investigate the information in observation for DALEC1 at a single time. In this section we will consider successive observations in time. It has been shown that the SIC in observations is additive with successive observations in time. The proof for this can be found in appendix A.1 of Fowler and Jan Van Leeuwen (2012). We can see this if we calculate the SIC for successive observations of foliar carbon, C_{fol}. We have the linearised observation operator and observation error covariance matrix at time t_i,

$$\mathbf{H}_i = \frac{\partial C^i_{fol}}{\partial \mathbf{x}_i} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{R}_i = \sigma^2_{cfol,o}. \tag{5.47}$$

For two successive observations of C_{fol} we have,

$$\hat{\mathbf{H}} = \begin{pmatrix} \mathbf{H}_0 \\ \mathbf{H}_1\mathbf{M}_0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ (1-\theta_{fol}) + f_{fol}(1-f_{auto})\zeta^0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{5.48}$$

and

$$\hat{\mathbf{R}} = \begin{pmatrix} \mathbf{R}_0 & 0 \\ 0 & \mathbf{R}_1 \end{pmatrix} = \begin{pmatrix} \sigma^2_{cfol,o} & 0 \\ 0 & \sigma^2_{cfol,o} \end{pmatrix}. \tag{5.49}$$

We then have,

$$SIC = \frac{1}{2}\ln\frac{\left|\mathbf{B}\right|}{\left|\mathbf{A}\right|} = \frac{1}{2}\ln\left(1 + \frac{\sigma^2_{cfol,b}}{\sigma^2_{cfol,o}} + \frac{\sigma^2_{cfol,b}\eta_0^2}{\sigma^2_{cfol,o}}\right), \tag{5.50}$$

where $\eta_i = (1-\theta_{fol}) + f_{fol}(1-f_{auto})\zeta^i$. We see this is similar to equation (5.32) for the SIC of a single carbon pool observation but with an added term evolved by the linearised model. Here the second term is multiplied by η_0^2 which is the square of the first element of the linearised model $\mathbf{M}_0$. We can continue adding more observations at successive times. For three observations at successive times we have,

$$SIC = \frac{1}{2}\ln\left(1 + \frac{\sigma^2_{cfol,b}}{\sigma^2_{cfol,o}} + \frac{\sigma^2_{cfol,b}\eta_0^2}{\sigma^2_{cfol,o}} + \frac{\sigma^2_{cfol,b}\eta_0^2\eta_1^2}{\sigma^2_{cfol,o}}\right), \tag{5.51}$$

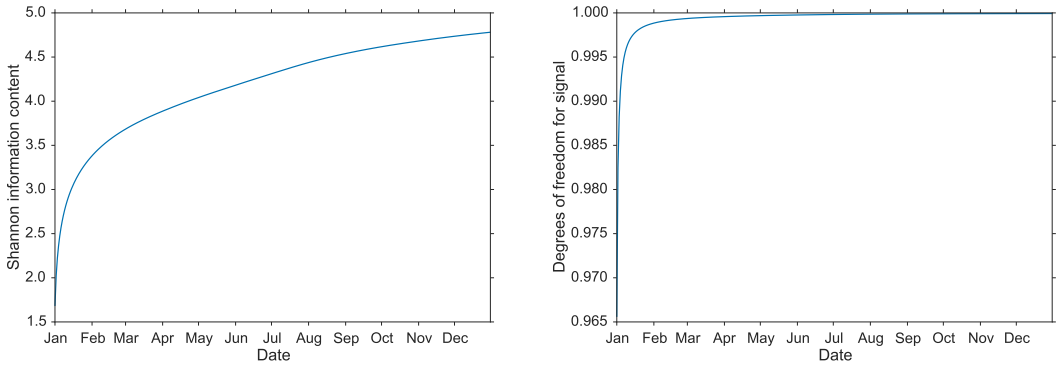
for four,

$$SIC = \frac{1}{2}\ln\left(1 + \frac{\sigma^2_{cfol,b}}{\sigma^2_{cfol,o}} + \frac{\sigma^2_{cfol,b}\eta_0^2}{\sigma^2_{cfol,o}} + \frac{\sigma^2_{cfol,b}\eta_0^2\eta_1^2}{\sigma^2_{cfol,o}} + \frac{\sigma^2_{cfol,b}\eta_0^2\eta_1^2\eta_2^2}{\sigma^2_{cfol,o}}\right). \tag{5.52}$$

Using a simple proof by induction we find that for n successive observations we have,

$$SIC \text{ for } n \text{ successive observations of } C_{fol} = \frac{1}{2}\ln\left(1+\frac{\sigma^2_{cfol,b}}{\sigma^2_{cfol,o}}\left(1+\sum_{k=0}^{n-2}\prod_{i=0}^{k}\eta_i^2\right)\right) \tag{5.53}$$

This demonstrates that SIC is additive for successive observations in time. In Figure 5.7 we have plotted the SIC and dfs for increasing numbers of observations of C_{fol}, using a year of meteorological driving data from a pine stand in Oregon. We see that as successive observations are added the information content tends to a limit where we are adding no new information with extra observations of C_{fol}. For dfs this limit is one as we are only observing a single degree of freedom so cannot constrain more than a single element of the state. For SIC we add a decreasing amount of information as observations are made further away from the initial state. We find similar results for all other carbon pools. This suggests making observations of any individual carbon pool for a forest site too often is not cost effective as after just a few observations the information you are adding to your system begins to decrease.

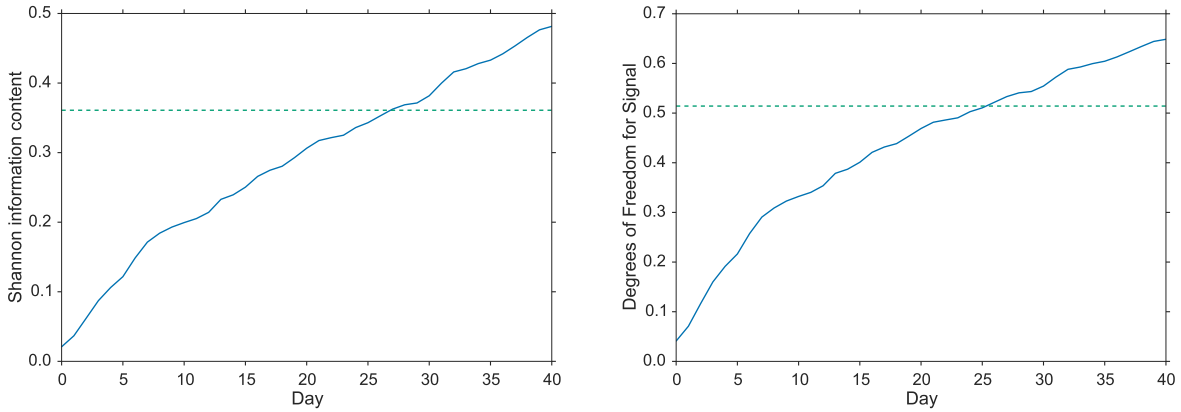


(a) SIC for successive C_{fol} observations

(b) dfs for successive C_{fol} observations

Figure 5.7: SIC and dfs for as successive C_{fol} observations are added throughout a year's window using driving data from a pine stand in Oregon taken in 2007.

In section 5.3.3.1 it was shown that observations of NEE made during the summer had significantly higher information content than those made during winter for an evergreen forest site. In figure 5.8 we show that 27 days of successive winter NEE observations (made from January 1st 2007) are required to give the same information content as a single summer observation of NEE (taken on 22nd June 2007).



(a) SIC for successive NEE observations

(b) dfs for successive NEE observations

Figure 5.8: Blue line: SIC and dfs for as successive NEE observations are added for 40 days from the 1st January 2007 using driving data from a pine stand in Oregon, green dotted line: SIC and dfs for a single summer observation of NEE made on 22nd June 2007.

5.3.3.4 Effect of time correlations between observation errors on information content

In chapter (ref. 1st results chapter) we have investigated the effect of including correlations in time between observation errors on the results from data assimilation with DALEC2. We can see the effect on the analytic representation of information content for two successive observations of NEE when including an off-diagonal correlation term in the matrix $\hat{\mathbf{R}}$. So that $\hat{\mathbf{R}} = \hat{\mathbf{D}}\mathbf{C}\hat{\mathbf{D}}^{\mathrm{T}}$, where $\hat{\mathbf{D}}$ is the diagonal matrix of observation standard deviations and $\mathbf{C}$ is a correlation matrix of the same shape. We then have

$$\hat{\mathbf{R}} = \hat{\mathbf{D}}\mathbf{C}\hat{\mathbf{D}}^{\mathrm{T}} = \begin{pmatrix} \sigma_{nee,o} & 0 \\ 0 & \sigma_{nee,o} \end{pmatrix} \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \begin{pmatrix} \sigma_{nee,o} & 0 \\ 0 & \sigma_{nee,o} \end{pmatrix} = \begin{pmatrix} \sigma^2_{nee,o} & \rho\sigma^2_{nee,o} \\ \rho\sigma^2_{nee,o} & \sigma^2_{nee,o} \end{pmatrix}, \tag{5.54}$$

with $0 \leq \rho < 1$.

We have not shown the analytic representation for the SIC here as it is too large. We instead use the symbolic Python package SymPy (Joyner et al., 2012) to plot the SIC for an increasing value of ρ in figure 5.9. Figure 5.9 shows that as the size of time correlation ρ approaches 1 the information content in the two observations of NEE decreases. This decrease in information content makes sense as including the correlation in time is decreasing the amount of independent information we are assimilating. This result is also seen in Järvinen et al. (1999) where including a serial correlations between observation errors is shown to reduce the weight given to the mean of the observations in the assimilation (equivalent to inflating the variance of the observations). This also supports the results found in section (ref. 1st results chapter) where we see that including

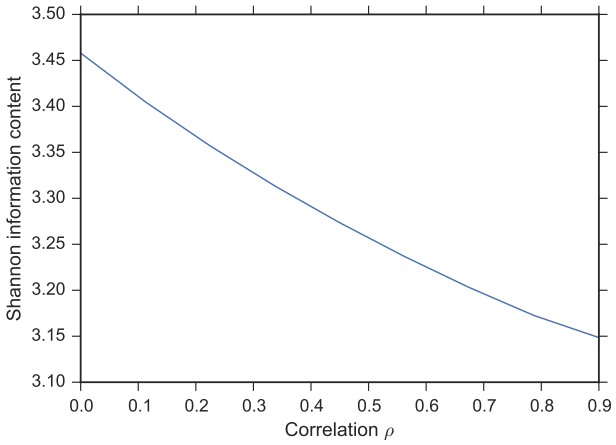


Figure 5.9: Shannon information content for two successive observations of NEE when a varying time correlation is included between observation errors.

correlations in time between observation errors reduces the fit to the assimilated observations in the analysis window.

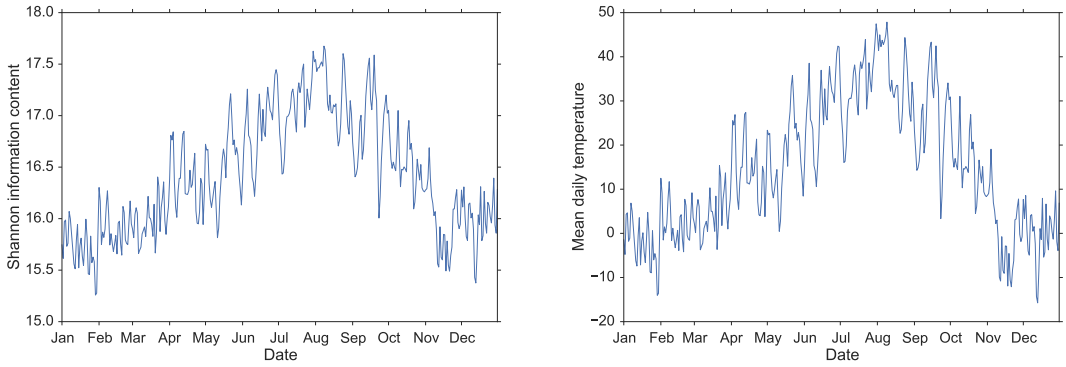
5.3.4 DALEC2 information content

5.3.4.1 Information content in observations for DALEC2

In this section we repeat and extend some of the results we have found for information content with the DALEC1 state estimation case in section 5.3.3 to the DALEC2 joint parameter and state estimation case. This means we now have an augmented state of 23 elements (17 parameters and 6 state variables) as opposed to just the 5 state members for DALEC1. For this reason we no longer examine the analytic representations of information content but instead consider the information content calculated numerically for DALEC2.

In section 5.3.3.1 it was shown that for DALEC1 the information content for a single observation of NEE was dependent on temperature. From Figure 5.10 we can see that this is still the case for DALEC2. However the value of SIC is higher for DALEC2 in Figure 5.10 than for DALEC1 in Figure 5.6 as the augmented state for the DALEC2 case also includes the parameters. This means that a single observation of NEE is giving us information about more elements of the state than for the DALEC1 state estimation case.

In figure 5.10 we have shown the information content varying for an evergreen forest site. As DALEC2 can also be parameterised and run for deciduous sites (with much work in this thesis being undertaken at Forest Research's deciduous study site, see sectionREF) it is important to



(a) SIC for single NEE observation

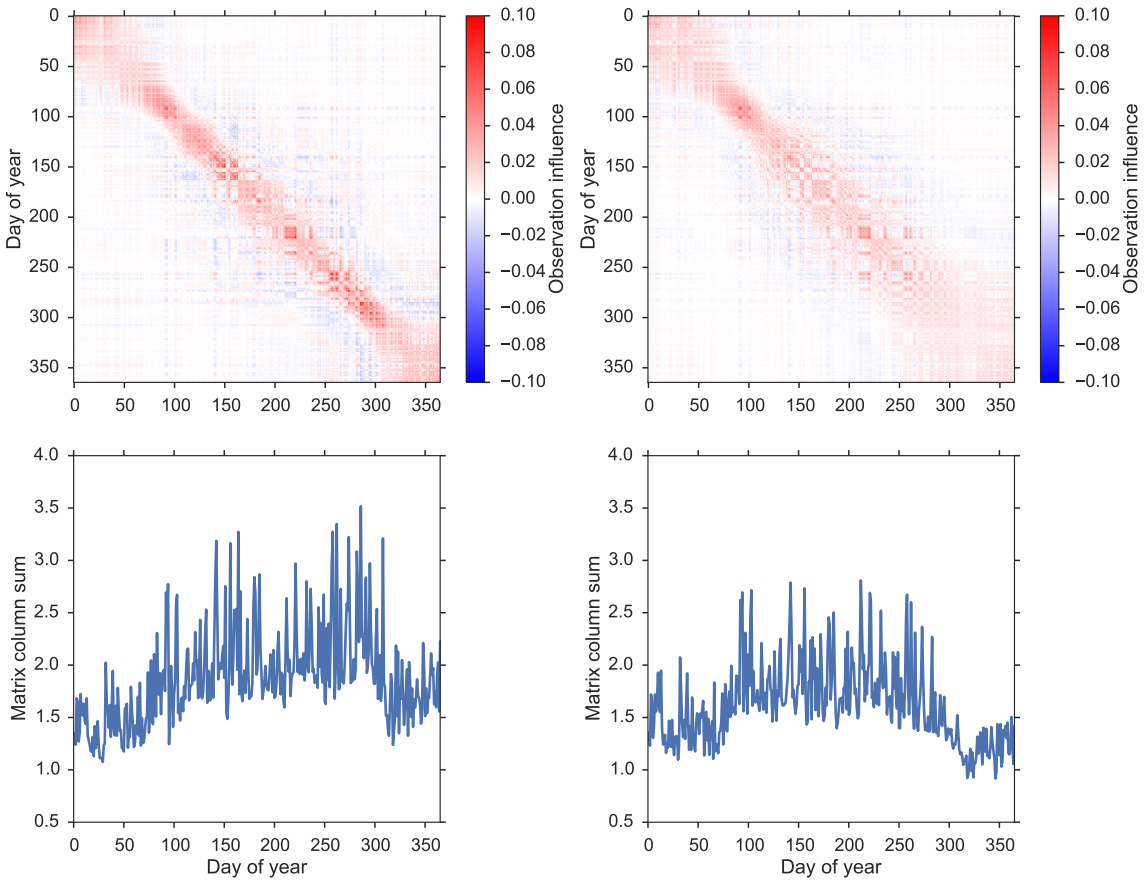
(b) Mean daily temperature for year's data

Figure 5.10: SIC a single NEE observation changing throughout a year's window using driving data from a pine stand in Oregon taken in 2007 (left). Mean daily temperature for the same site and period (right).

investigate the difference in information content between these cases. In order to visualise this difference, in figure 5.11 we show the analysis sensitivity to observations or influence matrix (Cardinali et al., 2004) as described in section 5.2.3.1, $\mathbf{S} = \mathbf{K}^T\mathbf{H}^T$, for a year's assimilation window with 365 observations of NEE. The influence matrix will depend on the initial augmented state we chose to linearise around, the driving data we use to run our model and the observations we specify for assimilation. In figure 5.11 we use an initial augmented state optimised for the Alice Holt deciduous forest and an initial augmented state optimised for an evergreen site in Oregon, we then use the same yearly driving data for both states so that it is only the difference between the initial augmented states of the sites effecting the difference between the influence matrices.

From figure 5.11 we can see that the influence of the assimilated observations of NEE is noticeably different between the deciduous and evergreen sites. However, in both cases at the beginning of the window there is a group of observations with similar influence. This makes sense as we are predicting the initial augmented state for DALEC2, so that observations closer to this initial state should have greater influence.

For the deciduous site in figure 5.11a we have groups of observations with high influence from around day 125 to day 175 and from day 250 to day 300. We also have some high influence observations between these two groups. High influence observations between these two groups would be consistent with the results showing that NEE observations have higher information content with higher temperatures, as the period between day 175 and 250 contains days with higher mean temperatures. For the evergreen site in figure 5.11b, although we have a group of observations at the beginning of the growing season with higher influence, we do not see a group



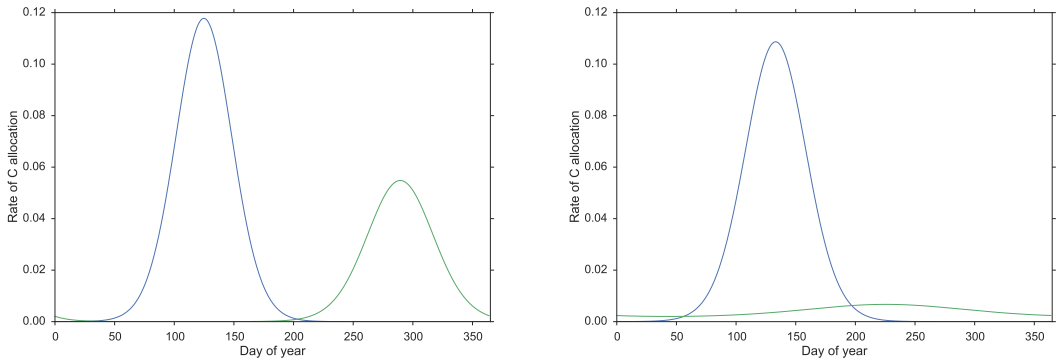
(a) Alice Holt deciduous site

(b) Oregon evergreen site

Figure 5.11: Influence matrices and column absolute value sums as described in section 5.2.3.1, showing the sensitivity of the modelled observations to the assimilated observations for a year's assimilation window starting at the beginning of January with 365 observations of NEE.

of with the same high influence between day 250 to day 300 as with the deciduous case. We still see observations of high influence corresponding to times of higher temperatures for the evergreen case.

In order to further investigate these groups of high influence observations we show the phenology functions controlling the rate of leaf-on and leaf-off for the DALEC2 model in figure 5.12. The description of phenology is the main difference between the more simplistic, evergreen only, DALEC1 and DALEC2 which can be parameterised for both deciduous and evergreen sites. It is logical that this is what is causing the difference in information content between the models and between the different sites. In figure 5.12 we see that the function controlling leaf-off for the deciduous site has a far larger peak than that of the evergreen site. This is expected as the deciduous site will drop all of its leaves at the end of the season. In both cases the forest puts most effort into putting on new leaves at the start of the growing season. This highlights the fact that the NEE for a deciduous site is highly controlled by phenology, as the forest cannot photosynthesise without



(a) Alice Holt deciduous site

(b) Oregon evergreen site

Figure 5.12: Phenology of DALEC2 model for a deciduous and evergreen forest. Blue line: function controlling rate of leaf-on (ϕ_{on}), green line: function controlling rate of leaf-off (ϕ_{off}).

leaves. Therefore the observations of NEE that help to constrain the phenology of the site should have a higher influence, as seen in figure 5.11a. Conversely for an evergreen site NEE is driven less by phenology and more by the climatic driving data. Seeing a greater relationship between temperature and information content for an evergreen site consequently makes sense and this can be seen in figure 5.11b.

5.3.4.2 Effect of time correlations on observation information content

In section 5.3.3.3 it was shown that, for the analytic DALEC1 case, when assimilating two successive observations of NEE the SIC decreased when including a correlation in time between NEE observation errors. It was noted that this was consistent with results found in section (ref. 1st results chapter) where including correlations between observation errors in time reduced the weight of the observations in the assimilation, in turn reducing the issue of overfitting to the assimilated observations. In figure 5.13 we repeat the experiment in section 5.3.3.3 but for DALEC2 with the year's worth of NEE observations assimilated in section (ref. 1st results chapter), in order to verify that including a correlation in time reduces the information content in assimilated observations. From figure 5.13 we see that we have similar results as in figure 5.9 where the information content in our observations decreases as we increase the time correlation between the assimilated observation errors. However, in figure 5.13 we have a higher value of SIC as we are assimilating many more observations than in figure 5.9. In figure 5.13 we have used the same correlation function as in section (ref. 1st results chapter) to create a correlated matrix $\hat{\mathbf{R}}$ and then varied the magnitude of the included correlation, ρ. The decreasing information content with an increasing correlation

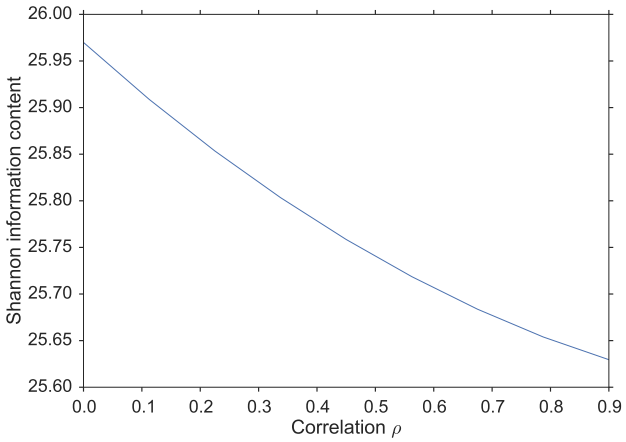


Figure 5.13: Shannon information content for 67 observations of NEE taken throughout a year's assimilation window when a varying time correlation is included between observation errors.

between observation errors in time supports the results in section 5.3.3.3 and section (ref. 1st results chapter). This is also consistent with the results of Järvinen et al. (1999) where including correlations between observation errors in time is shown to reduce the weight given to the mean of the observations in the assimilation (equivalent to inflating the variance of the observations).

5.4 Conclusions

In this chapter we have investigated both the observability and information content given the observations available to us. In section 5.3.1 and section 5.3.2 we have shown that for both DALEC1 and DALEC2 we have an observable system with the available observations, in this case NEE. An observable system in this case means that for data assimilation we can construct a locally unique solution from the observational information alone.

In section 5.3.3 we have seen that for the DALEC1 evergreen case the information content in observations of NEE is largely dependent on temperature, with higher temperatures meaning higher information content. This is important for informing planned maintenance or down time at flux tower sites measuring NEE. This dependance of information content on temperature is also seen for observations of ground respiration and total ecosystem respiration. When assimilating additional observations at the same time alongside NEE we have found that most information is added when the additional observation provides an orthogonal constraint to that of NEE. This is the case for root carbon (C_{roo}) and woody biomass carbon (C_{woo}), with the European Space Agency BIOMASS mission being launched soon this should add valuable information to current data as-

similation schemes. When using DALEC1 and assimilating successive observations in time it was shown that as observations are added further away from the initial state their impact is decreased. For two successive observation of NEE it was also shown that including a correlation in time between observation errors decreases the information content in the assimilated observations. This is consistent with results found in (ref. 1st results chapter) where including correlations in time between observation errors reduces overfitting to the assimilated observations.

In section 5.3.4 we again see the temperature dependence of information content in observations of NEE for DALEC2. However, for DALEC2 we also have varying information content based on the type of ecosystem we are observing. For a deciduous forest site we see that as well as temperature the information content in observations of NEE is strongly dependent on the time of growing season. Observations made at the time of leaf-on and leaf-off have higher influence on the results of the assimilation. This makes physical sense as for a deciduous ecosystem NEE is highly controlled by phenology, as the forest cannot photosynthesise without leaves. Therefore the observations of NEE that help to constrain the phenology of the site should have a higher influence. For an evergreen forest site we see much less dependence on phenology and have a greater relationship between temperature and information content. We again see similar results as in section 5.3.3 when including correlations in time between observation errors, with an increasing correlation, ρ, reducing the information content in the assimilated observations.

Chapter 6

Investigating the role of prior and observation error correlations

6.1 Abstract

Efforts to implement variational data assimilation routines with functional ecology models and land surface models have been limited, with sequential and Markov chain Monte Carlo data assimilation methods being prevalent. When data assimilation has been used with models of carbon balance, prior or "background" errors (in the initial state and parameter values) and observation errors have largely been treated as independent and uncorrelated. Correlations between background errors have long been known to be a key aspect of data assimilation in numerical weather prediction. More recently, it has been shown that accounting for correlated observation errors in the assimilation algorithm can considerably improve data assimilation results and forecasts. In this paper we implement a Four-Dimensional Variational data assimilation (4D-Var) scheme with a simple model of forest carbon balance, for joint parameter and state estimation and assimilate daily observations of Net Ecosystem CO_2 Exchange (NEE) taken at the Alice Holt forest CO_2 flux site in Hampshire, UK. We then investigate the effect of specifying correlations between parameter and state variables in background error statistics and the effect of specifying correlations in time between observation errors. The idea of including these correlations in time is new and has not been previously explored in carbon balance model data assimilation. In data assimilation, background and observation error statistics are often described by the background error covariance matrix and the observation error covariance matrix. We outline novel methods for creating

correlated versions of these matrices, using a set of previously postulated dynamical constraints to include correlations in the background error statistics and a Gaussian correlation function to include time correlations in the observation error statistics. The methods used in this paper will allow the inclusion of time correlations between many different observation types in the assimilation algorithm, meaning that previously neglected information can be accounted for. In our experiments we assimilate a single year of NEE observations and then run a forecast for the next 14 years. We compare the results using our new correlated background and observation error covariance matrices and those using diagonal covariance matrices. We find that using the new correlated matrices reduces the root mean square error in the 14 year forecast of daily NEE by 44% decreasing from 4.22 $gCm^{-2}day^{-1}$ to 2.38 $gCm^{-2}day^{-1}$.

6.2 Introduction

The land surface and oceans are responsible for removing around half of all human emitted carbon-dioxide from the atmosphere and therefore mediate the effect of anthropogenic induced climate change. Terrestrial ecosystem carbon uptake is the least understood process in the global carbon cycle (Ciais et al., 2014). It is therefore vital that we improve understanding of the carbon uptake of terrestrial ecosystems and their response to climate change in order to better constrain predictions of future carbon budgets. Observations of the Net Ecosystem Exchange (NEE) of CO_2 between terrestrial ecosystems and the atmosphere are now routinely made at flux tower sites world-wide, at sub-hourly resolution and covering multiple years (Baldocchi, 2008), providing a valuable resource for carbon balance model validation and data assimilation.

Data assimilation is the process of combining a mathematical model with observations in order to improve the estimate of the state of a system. Data assimilation has successfully been used in many applications to significantly improve model state and forecasts. Perhaps the most important application has been in numerical weather prediction where data assimilation has contributed to the forecast accuracy being increased at longer lead times, with the four day forecast in 2014 having the same level of accuracy as the one day forecast in 1979 (Bauer et al., 2015). This increase in forecast skill is obviously not solely due to data assimilation but also increased quality and resolution of observations along with improvements in model structure, however the introduction and evolution of data assimilation has played a large part (Dee et al., 2011). The current method implemented at many leading operational numerical weather prediction centres is known as Four-

Dimensional Variational data assimilation (4D-Var) (Bonavita et al., 2015; Clayton et al., 2013), which has been shown to be a significant improvement over its predecessor three-dimensional variational data assimilation (Lorenc and Rawlins, 2005). Variational assimilation techniques minimise a cost function to find the optimal state of a system given all available knowledge of errors in the model and observations. The minimisation routine typically requires the derivative of the model which can sometimes prove difficult to calculate. Using techniques such as automatic-differentiation (Renaud, 1997) can reduce the time taken to implement the derivative of a model.

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). This is possibly why Monte Carlo Markov chain (MCMC) data assimilation methods have been used more with ecosystem carbon cycle models. Smaller ecosystem models are much less computationally expensive to run than large numerical weather prediction models, meaning that MCMC methods (requiring many more model runs than variational assimilation methods) are more easily implemented. For larger scale and more complex ecosystem models variational methods represent a much more computationally efficient option for data assimilation. Variational data assimilation can be used for joint parameter and state estimation by augmenting the state vector with the parameters (Navon, 1998). By including the parameters in the state vector we must also specify error statistics and error correlations for them. Smith et al. (2009) show that the prescription of these error statistics and their correlations can have a significant impact on parameter-state estimates obtained from the assimilation.

Many different observations relevant to the carbon balance of forests have now been combined with functional ecology models, using data assimilation, in order to improve our knowledge of ecological systems (Zobitz et al., 2011; Fox et al., 2009; Richardson et al., 2010; Quaife et al., 2008; Zobitz et al., 2014b; Niu et al., 2014). Two such models that have been used extensively with data assimilation are the Data Assimilation Linked Ecosystem Carbon (DALEC) model (Williams et al., 2005) and the Simplified Photosynthesis and Evapo-Transpiration (SIPNET) model (Braswell et al., 2005). Nearly all data assimilation routines built with these models have used sequential and Monte Carlo Markov chain (MCMC) data assimilation methods with the exception of a variational routine being implemented for DALEC by Delahaies et al. (2013). There have been examples of global land surface models being implemented with variational methods such as the ORganizing

Carbon and Hydrology In Dynamic EcosystEms model (ORCHIDEE) (Krinner et al., 2005) and the Biosphere Energy Transfer HYdrology scheme (BETHY) in a Carbon Cycle Data Assimilation System (CCDAS) (Kaminski et al., 2013). These examples have mainly been used to assimilate data from satellite and atmospheric CO_2 observations with only a few cases where site level data has also been assimilated (Verbeeck et al., 2011; Bacour et al., 2015).

Forest carbon balance model parameters are often determined in advance of using the model for forecasting by calibration of the model against observations (Richardson et al., 2010; Bloom and Williams, 2015). Here we take the alternative approach of concurrent state-parameter estimation. A key difference between the joint state-parameter estimation approach and a priori calibration is the way that the observational data is used. Pre-calibration approaches train the model against historical data and so become infeasible when there is a lack of sufficient observational information prior to the model forecast period. Joint state-parameter estimation methods have the advantage that observations could be used as they arrive in real time, by sequential assimilation cycling. This approach also gives the possibility of adapting to changes in the forest (e.g., tree thinning, fires etc.) that may change the parameter values over time.

Background errors (describing our knowledge of error in prior model estimates before data assimilation) and observation errors have largely been treated as uncorrelated and independent in ecosystem model data assimilation schemes. In 3D and 4DVar schemes background and observation errors are represented by the error covariance matrices **B** and **R** respectively. The off-diagonal elements of these matrices indicate the correlations between errors in the parameter and state variables for **B** and the correlations between observation errors for **R**. In the assimilation, the off-diagonal terms in the **B** matrix act to spread information between the state and augmented parameter variables (Kalnay, 2003). This means that assimilating observations of one state variable can act to update different state and parameter variables in the assimilation when correlations are included in **B**. In 4D-Var the **B** matrix is propagated implicitly by the forecast model, so that even a propagated diagonal **B** matrix can develop correlations throughout an assimilation window. These correlations will only be in the propagated **B** matrix, with the **B** matrix valid at the initial time remaining unchanged. Including correlations in **B** has been shown to significantly improve data assimilation results in numerical weather prediction (Bannister, 2008).

Including correlations between observation errors has only started to be explored recently in numerical weather prediction, with **R** still often treated as diagonal (Stewart et al., 2013). Includ-

ing some correlation structure in **R** has been shown to improve forecast accuracy (Weston et al., 2014). Currently the correlations included in **R** have been mainly between observations made at the the same time rather than correlations between observations throughout time. When assimilating observations, data streams with many more observations can have a greater impact on the assimilation than those with fewer observations. In Richardson et al. (2010) this problem is discussed when assimilating large numbers of NEE observations along with smaller numbers of leaf area index and soil respiration observations. To address this problem Richardson et al. uses a cost function that calculates the product of the departures from the observations rather than a cost function which sums these departures, giving a relative rather than absolute measure of the goodness-of-fit to the observations. This problem is also encountered in Bacour et al. (2015) when assimilating daily eddy covariance data with weekly observations of the FrAction of Photosynthetically Active Radiation (FAPAR). In Bacour et al. (2015) the error in observations of FAPAR is divided by two in order to give these less frequent observations more weight in the assimilation algorithm. Specifying serial time correlations between observations represents another way of addressing this problem, whilst also adding valuable information to the data assimilation routine. Including serial correlations between observations of the same quantity decreases the impact of these observations (Järvinen et al., 1999) therefore increasing the impact of less frequent observations.

In this paper we implement the new version of DALEC (DALEC2 (Bloom and Williams, 2015)) in a 4D-Var data assimilation scheme for joint state and parameter estimation, assimilating daily NEE observations from the Alice Holt flux site in Hampshire, UK (Wilkinson et al., 2012). This assimilation scheme is then subjected to rigorous testing to ensure correctness. A new method is outlined for including parameter and state correlations in the background "prior" error covariance matrix. Currently parameter and state error statistics are largely treated as independent and uncorrelated when data assimilation has been used with models of carbon balance. We also introduce a novel method for including serial time correlations in the observation error covariance matrix. The idea of including time correlations between observation error statistics is new and has not been previously explored in carbon balance model data assimilation. These correlated matrices are then used in a series of experiments in order to examine the effect that including correlations in the assimilation scheme has on the results.

6.3 Model and Data Assimilation Methods

6.3.1 Alice Holt research forest

Alice Holt Forest is a research forest area managed by the UK Forestry Commission located in Hampshire, SE England. Forest Research has been operating a CO_2 flux measurement tower in a portion of the forest, the Straits Inclosure, since 1998 so it is one of the longer forest site CO_2 flux records, globally. The Straits Inclosure is a 90ha area of managed deciduous broadleaved plantation woodland, presently approximately 80 years old, on a surface water gley soil. The majority of the canopy trees are oak (*Quercus robur* L.), with an understory of hazel (*Corylus avellana* L.) and hawthorn (*Crataegus monogyna* Jacq.); but there is a small area of conifers (*Pinus nigra* J. F. Arnold) within the tower measurement footprint area in some weather conditions. Further details of the Straits Inclosure site and the measurement procedures are given in Wilkinson et al. (2012), together with analysis of stand-scale 30 minute average net CO_2 fluxes (NEE) measured by standard eddy covariance methods from 1998-2011. The data used here span from January 1999 to December 2013, and consist of the NEE fluxes and meteorological driving data of temperatures, irradiance and atmospheric CO_2 concentration. The original NEE data were subjected to normal quality control procedures, including u^* filtering to remove unreliable data when there were low turbulence night time conditions, as described in Wilkinson et al. (2012), but were not gap-filled. To compute daily NEE observations we take the sum over the 48 measurements made each day. We only select days where there is no missing data and over 90% of CO_2 flux observations have a quality control flag associated with the best observations and no observations associated with the worst from the EddyPro flux processing software (LI-COR, Inc., 2015).

6.3.2 The DALEC2 model

The DALEC2 model is a simple process-based model describing the carbon balance of a forest ecosystem (Bloom and Williams, 2015) and is the new version of the original DALEC (Williams et al., 2005). The model is constructed of six carbon pools (labile (C_{lab}), foliage (C_f), fine roots (C_r), woody stems and coarse roots (C_w), fresh leaf and fine root litter (C_l) and soil organic matter and coarse woody debris (C_s)) linked via fluxes. The aggregated canopy model (ACM) (Williams et al., 1997) is used to calculate daily gross primary production (*GPP*) of the forest, taking meteorological driving data and the modelled leaf area index (a function of C_f) as arguments. Figure 6.1 shows a

schematic of how the carbon pools are linked in DALEC2.

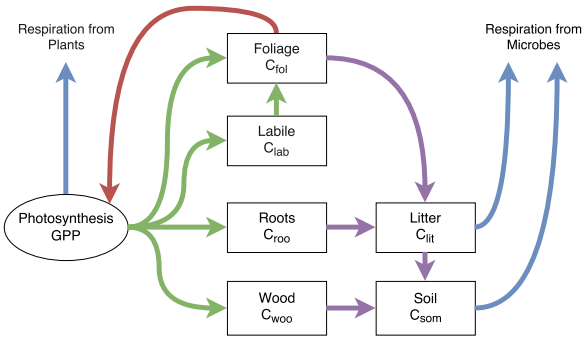


Figure 6.1: Representation of the fluxes in the DALEC2 carbon balance model. Green arrows represent C allocation, purple arrows represent litter fall and decomposition fluxes, blue arrows represent respiration fluxes and the red arrow represents the influence of leaf area index in the *GPP* function.

The model equations for the carbon pools at day i are as follows:

$$GPP^i = ACM(C_{fol}^{i-1}, c_{lma}, c_{eff}, \Psi) \tag{6.1}$$

$$C_{lab}^i = C_{lab}^{i-1} + (1 - f_{auto})(1 - f_{fol}) f_{lab} GPP^i - \Phi_{on} C_{lab}^{i-1}, \tag{6.2}$$

$$C_{fol}^i = C_{fol}^{i-1} + \Phi_{on} C_{lab}^{i-1} + (1 - f_{auto}) f_{fol} GPP^i - \Phi_{off} C_{fol}^{i-1}, \tag{6.3}$$

$$C_{roo}^i = C_{roo}^{i-1} + (1 - f_{auto})(1 - f_{fol})(1 - f_{lab}) f_{roo} GPP^i - \theta_{roo} C_{roo}^{i-1}, \tag{6.4}$$

$$C_{woo}^i = C_{woo}^{i-1} + (1 - f_{auto})(1 - f_{fol})(1 - f_{lab})(1 - f_{roo}) GPP^i - \theta_{woo} C_{woo}^{i-1}, \tag{6.5}$$

$$C_{lit}^i = C_{lit}^{i-1} + \theta_{roo} C_{roo}^{i-1} + \Phi_{off} C_{fol}^{i-1} - (\theta_{lit} + \theta_{min}) e^{\Theta T^{i-1}} C_{lit}^{i-1}, \tag{6.6}$$

$$C_{som}^i = C_{som}^{i-1} + \theta_{woo} C_{woo}^{i-1} + \theta_{min} e^{\Theta T^{i-1}} C_{lit}^{i-1} - \theta_{som} e^{\Theta T^{i-1}} C_{som}^{i-1}, \tag{6.7}$$

where T^{i-1} is the daily mean temperature, Ψ represents the meteorological driving data used in the *GPP* function and Φ_{on}/Φ_{off} are functions controlling leaf-on and leaf-off. Descriptions for each model parameter used in equations (6.1) to (6.7) are included in the appendix in table 6.3. DALEC2 differs from the original DALEC in that it can be parameterised for both deciduous and evergreen sites with Φ_{on} and Φ_{off} being able to reproduce the phenology of either type of site. The full details of this version of DALEC can be found in Bloom and Williams (2015).

6.3.3 4D-Var

Following the approach of Smith et al. (2011) for joint state and parameter estimation, we consider the discrete nonlinear dynamical system given by

$$\mathbf{z}_i = \mathbf{f}_{i-1\to i}(\mathbf{z}_{i-1}, \mathbf{p}_{i-1}), \tag{6.8}$$

where $\mathbf{z}_i \in \mathbb{R}^n$ is the state vector at time t_i, $\mathbf{f}_{i-1\to i}$ is the nonlinear model operator propagating the state at time t_{i-1} to time t_i for $i = 1,2,\ldots,N$ and $\mathbf{p}_{i-1} \in \mathbb{R}^q$ is a vector of q model parameters at time t_{i-1}. For DALEC2 the state vector $\mathbf{z}_i = (C^i_{lab}, C^i_{for}, C^i_{roo}, C^i_{woo}, C^i_{lit}, C^i_{som})^T$, with the parameters shown in table 6.3. Given a set of fixed parameters, the value of the forecast at time t_i is uniquely determined by the initial value. The model parameters are not updated by the nonlinear model operator, therefore the evolution of the parameters is given by,

$$\mathbf{p}_i = \mathbf{p}_{i-1}, \tag{6.9}$$

for $i = 1,2,\ldots,N$. We define the new vector $\mathbf{x}$ by joining the parameter vector $\mathbf{p}$ with the model state vector $\mathbf{z}$, giving us the augmented state vector

$$\mathbf{x} = \begin{pmatrix} \mathbf{p} \\ \mathbf{z} \end{pmatrix} \in \mathbb{R}^{q+n}. \tag{6.10}$$

We define the augmented system model by

$$\mathbf{x}_i = \mathbf{m}_{i-1\to i}(\mathbf{x}_{i-1}), \tag{6.11}$$

where

$$\mathbf{m}_{i-1\to i}(\mathbf{x}_{i-1}) = \begin{pmatrix} \mathbf{p}_{i-1} \\ \mathbf{f}_{i-1\to i}(\mathbf{z}_{i-1}, \mathbf{p}_{i-1}) \end{pmatrix} = \begin{pmatrix} \mathbf{p}_i \\ \mathbf{z}_i \end{pmatrix} \in \mathbb{R}^{q+n}. \tag{6.12}$$

The available observations at time t_i are represented by the vector $\mathbf{y}_i \in \mathbb{R}^{r_i}$ which are related to the augmented state vector through the equation

$$\mathbf{y}_i = \mathbf{h}_i(\mathbf{x}_i) + \boldsymbol{\varepsilon}_i, \tag{6.13}$$

where $\mathbf{h}_i : \mathbb{R}^{q+n} \to \mathbb{R}^{r_i}$ is the observation operator mapping the augmented state vector to observation space and $\boldsymbol{\varepsilon}_i \in \mathbb{R}^{r_i}$ represents the observation errors. These errors are usually assumed to be unbiased, Gaussian and serially uncorrelated with known covariance matrices $\mathbf{R}_i$.

In the 4D-Var data assimilation detailed here we aim to find the parameter and initial state values such that the model trajectory best fits the data over some time window, given some prior information about the system. The output from 4D-Var is an updated set of parameters, and an updated model state, valid at the beginning of the time window. The updated model state may be used as initial conditions for a forecast using the full nonlinear DALEC2 model. We assume that at time t_0 we have an initial estimate to the augmented state, usually referred to as the background vector denoted $\mathbf{x}^b$. This background is assumed to have unbiased, Gaussian errors with known covariance matrix $\mathbf{B}$. Adding the background term ensures that our problem is well posed and that we can find a locally unique solution (Tremolet, 2006). In 4D-Var we aim to find the initial state that minimises the weighted least squares distance to the background while minimising the weighted least squares distance of the model trajectory to the observations over the time window $t_0, \ldots, t_N$ (Lawless, 2013). We do this by finding the state $\mathbf{x}_0^a$ at time t_0 that minimises the cost function

$$J(\mathbf{x}_0) = \frac{1}{2}(\mathbf{x}_0 - \mathbf{x}^b)^T \mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}^b) + \frac{1}{2}\sum_{i=0}^{N}(\mathbf{y}_i - \mathbf{h}_i(\mathbf{x}_i))^T \mathbf{R}_i^{-1}(\mathbf{y}_i - \mathbf{h}_i(\mathbf{x}_i)), \tag{6.14}$$

subject to the augmented states $\mathbf{x}_i$ satisfying the nonlinear dynamical model (6.11). The state that minimises the cost function, $\mathbf{x}_0^a$, is commonly called the analysis. This state is found using a minimisation routine that takes as its input arguments the cost function, the background vector ($\mathbf{x}^b$) and also the gradient of the cost function given as,

$$\nabla J(\mathbf{x}_0) = \mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}^b) - \sum_{i=0}^{N} \mathbf{M}_{i,0}^T \mathbf{H}_i^T \mathbf{R}_i^{-1}(\mathbf{y}_i - \mathbf{h}_i(\mathbf{x}_i)) \tag{6.15}$$

where $\mathbf{H}_i = \frac{\partial \mathbf{h}_i(\mathbf{x}_i)}{\partial \mathbf{x}_i}$ is the linearized observation operator and $\mathbf{M}_{i,0} = \mathbf{M}_{i-1}\mathbf{M}_{i-2}\cdots\mathbf{M}_0$ is the tangent linear model with $\mathbf{M}_i = \frac{\partial \mathbf{m}_{i-1\to i}(\mathbf{x}_i)}{\partial \mathbf{x}_i}$. In practice $\nabla J(\mathbf{x}_0)$ is calculated using the method of Lagrange multipliers as shown in Lawless (2013). We can rewrite the cost function and its gradient to avoid the sum notation as,

$$J(\mathbf{x}_0) = \frac{1}{2}(\mathbf{x}_0 - \mathbf{x}^b)^T \mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}^b) + \frac{1}{2}(\hat{\mathbf{y}} - \hat{\mathbf{h}}(\mathbf{x}_0))^T \hat{\mathbf{R}}^{-1}(\hat{\mathbf{y}} - \hat{\mathbf{h}}(\mathbf{x}_0)) \tag{6.16}$$

and

$$\nabla J(\mathbf{x}_0) = \mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}^b) - \hat{\mathbf{H}}^T \hat{\mathbf{R}}^{-1}(\hat{\mathbf{y}} - \hat{\mathbf{h}}(\mathbf{x}_0)), \tag{6.17}$$

where,

$$\hat{\mathbf{y}} = \begin{pmatrix} \mathbf{y}_0 \\ \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_N \end{pmatrix}, \ \hat{\mathbf{h}}(\mathbf{x}_0) = \begin{pmatrix} \mathbf{h}_0(\mathbf{x}_0) \\ \mathbf{h}_1(\mathbf{m}_{0\to1}(\mathbf{x}_0)) \\ \vdots \\ \mathbf{h}_N(\mathbf{m}_{0\to N}(\mathbf{x}_0)) \end{pmatrix}, \ \hat{\mathbf{R}} = \begin{pmatrix} \mathbf{R}_{0,0} & \mathbf{R}_{0,1} & \dots & \mathbf{R}_{0,N} \\ \mathbf{R}_{1,0} & \mathbf{R}_{1,1} & \dots & \mathbf{R}_{1,N} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{R}_{N,0} & \mathbf{R}_{N,1} & \dots & \mathbf{R}_{N,N} \end{pmatrix} \text{ and } \hat{\mathbf{H}} = \begin{pmatrix} \mathbf{H}_0 \\ \mathbf{H}_1\mathbf{M}_0 \\ \vdots \\ \mathbf{H}_N\mathbf{M}_{N,0} \end{pmatrix}. \tag{6.18}$$

Solving the cost function in this form also allows us to build serial time correlations into the observation error covariance matrix $\hat{\mathbf{R}}$. The off-diagonal blocks of $\hat{\mathbf{R}}$ represent correlations in time between assimilated observations and are usually taken to be zero. In section 6.3.6 we show how these off-diagonal blocks can be specified. We can also calculate the posterior or analysis error covariance matrix after assimilation as,

$$\mathbf{A} = (\mathbf{B}^{-1} + \hat{\mathbf{H}}^T \hat{\mathbf{R}}^{-1} \hat{\mathbf{H}})^{-1}. \tag{6.19}$$

We can use this matrix to estimate the uncertainty in our parameter and initial state variables after assimilation.

6.3.4 Implementation and testing of 4D-Var system

In our DALEC2 4D-Var scheme we are performing joint parameter and state estimation. Typically MCMC techniques have been used for joint parameter and state estimation with functional ecology models, such as DALEC2. However 4D-Var has been used for joint parameter and state estimation with global carbon cycle models (Kaminski et al., 2013). The variational approach is computationally efficient and robust, making it particularly suited to large problems with complex models. The augmented state vector, $\mathbf{x}_0$, corresponds to the vector of the 17 model parameters and 6 initial carbon pool values, which can be found in the appendix in table 6.3. Here the nonlinear model (DALEC2) only updates the initial carbon pool values when evolving the augmented state vector forward in time with the parameters being held constant. To find the background estimate, $\mathbf{x}^b$, to the augmented state vector we can either use a previous DALEC2 model forecast estimate of the state of the system for the site (when available) or use expert elicitation to define likely state

and parameter values and ranges for the site. The background vector ($\mathbf{x}^b$) and its corresponding standard deviations (see table 6.3) used in this paper were provided from existing runs of the the CARbon DAta-MOdel fraMework (CARDAMOM) (Exbrayat et al., 2015). The CARDAMOM output is a dataset derived from satellite observations of leaf area index which provides a reasonable first guess to DALEC2 state and parameter values for the Alice Holt research site. In this paper we assimilate observations of daily NEE. From Richardson et al. (2008) the measurement error in observations of daily NEE is between 0.2 to 0.8 $\mathrm{gCm^{-2}day^{-1}}$. Richardson et al. (2008) also shows that flux errors are heteroscedastic. We assume a constant standard deviation of 0.5 $\mathrm{gCm^{-2}day^{-1}}$ in the assimilated observations of daily NEE as we found this standard deviation gave the best weighting to the observations in the assimilation algorithm, producing the best results for the forecast of NEE after assimilation. Assuming this constant standard deviation also allows for correlations in time between observation errors to be included more easily. Ignoring the heteroscedastic nature of NEE errors may influence results by giving observations of larger magnitude a higher weight than would be realistic. Future work should try to incorporate the heteroscedastic nature of NEE errors.

In order to find the tangent linear model (TLM) for DALEC2 it is necessary to find the derivative of the model at each time step with respect to the 17 model parameters and the 6 carbon pools. We use the AlgoPy automatic differentiation package (Walter and Lehmann, 2013) in Python to calculate the TLM at each time step. This package uses forward mode automatic differentiation to calculate the derivative of the model. In the following tests we use a diagonal approximation to the background and observation error covariance matrices so that, $\mathbf{B}_{diag} = \mathrm{diag}(\boldsymbol{\sigma}_b)^2$ and $\hat{\mathbf{R}}_{diag} = \mathrm{diag}(\boldsymbol{\sigma}_o)^2$, where $\boldsymbol{\sigma}_b$ is the vector of background standard deviations found in table 6.3 and $\boldsymbol{\sigma}_o$ is the vector of observational standard deviations, for a single observation of NEE $\sigma_o = 0.5$ $\mathrm{gCm^{-2}day^{-1}}$. To minimise the cost function we use the truncated Newton iteration method (Nocedal and Wright, 1999) from the Python package Scipy.optimize (Jones et al., 2001). This method uses a number of stopping criteria to ensure convergence to a minimum of our cost function. In sections 6.3.4.1 to 6.3.4.3 we show tests of our scheme.

6.3.4.1 Test of tangent linear model

The TLM is used in the calculation of the gradient of our cost function in 4D-Var. We can have confidence that our implementation of the TLM for DALEC2 is correct as it passes the following

relevant tests (Li et al., 1994). In 4D-Var we assume the tangent linear hypothesis,

$$\mathbf{m}_{0\to i}(\mathbf{x}_0 + \gamma\delta\mathbf{x}_0) \approx \mathbf{m}_{0\to i}(\mathbf{x}_0) + \gamma\mathbf{M}_{i,0}\delta\mathbf{x}_0, \tag{6.20}$$

where $\delta\mathbf{x}_0$ is a perturbation of the initial augmented state $\mathbf{x}_0$ and γ is a parameter controlling the size of this perturbation. The validity of this assumption depends on how nonlinear the model is, the length of the assimilation window and the size of the augmented state perturbation $\delta\mathbf{x}_0$. We can test this by rearranging equation (6.20) to find,

$$\frac{||\mathbf{m}_{0\to i}(\mathbf{x}_0 + \gamma\delta\mathbf{x}_0) - \mathbf{m}_{0\to i}(\mathbf{x}_0) - \gamma\mathbf{M}_{i,0}\delta\mathbf{x}_0||}{||\gamma\mathbf{M}_{i,0}\delta\mathbf{x}_0||} \to 0, \tag{6.21}$$

as $\gamma \to 0$ (here we are using the Euclidean norm). Equation (6.21) should hold if our implementation of the TLM is correct, even for a weakly non-linear model. Figure 6.2 shows equation (6.21) plotted for DALEC2 with i fixed at 731 days, a fixed 5% perturbation $\delta\mathbf{x}_0$ and values of γ approaching zero . Figure 6.2 shows that the TLM behaves as expected for values of γ approaching 0. This was also tested for different choices of $\mathbf{x}_0$ and sizes of perturbation with similar results.

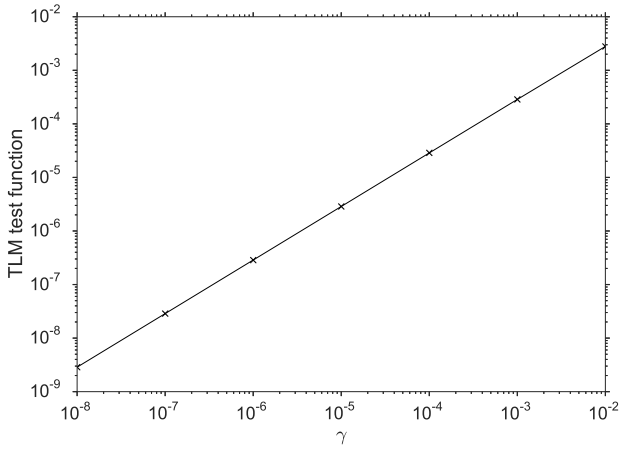


Figure 6.2: Plot of the tangent linear model test function (equation (6.21)) for DALEC2, for a fixed TLM evolving the perturbed augmented state 731 days forward in time and a fixed 5% perturbation, $\delta\mathbf{x}_0$.

It is also useful to show how the TLM behaves over a time window to see how the error in the TLM grows as we evolve the augmented state further forward in time. We again rearrange equation (6.20) with an additional error term to find,

$$\text{percentage error in TLM} = \frac{||m_{0\to i}(\mathbf{x}_0 + \gamma\delta\mathbf{x}_0) - m_{0\to i}(\mathbf{x}_0) - \gamma\mathbf{M}_{i,0}\delta\mathbf{x}_0||}{||\gamma\mathbf{M}_{i,0}\delta\mathbf{x}_0||} \times 100. \tag{6.22}$$

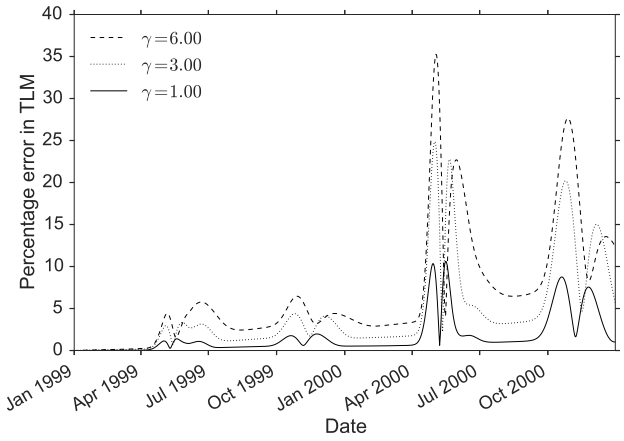


Figure 6.3: Plot of the percentage error in the tangent linear model (equation (6.22)) for DALEC2 when evolving the model state forward over a period of two years with three different values of γ and a fixed 5% perturbation $\delta\mathbf{x}_0$.

In figure 6.3 we plot the percentage error in the TLM tested throughout a two-year period as DALEC2 is run forward. From figure 6.3 we can see that the TLM for DALEC2 performs well after being run forward a year with less than a 7% error for all values of γ. By the second year we see some peaks in the error in spring and autumn. This is due to leaf on and leaf off functions in the TLM going out of phase with the nonlinear DALEC2. At these peaks the error reaches a maximum at 35% then coming back to around 10% before growing again in the autumn. Although this level of error is still acceptable we present results using a one year assimilation window in this paper as in practice we could cycle assimilation windows to make use of multiple years of data (Moodey et al., 2013).

6.3.4.2 Test of adjoint model

The adjoint model we have implemented for DALEC2 passes correctness tests. For the TLM $\mathbf{M}_{i,0}$ and its adjoint $\mathbf{M}_{i,0}^T$ we have the identity

$$< \mathbf{M}_{i,0}\delta\mathbf{x}_0, \mathbf{M}_{i,0}\delta\mathbf{x}_0 > = < \delta\mathbf{x}_0, \mathbf{M}_{i,0}^T\mathbf{M}_{i,0}\delta\mathbf{x}_0 > \tag{6.23}$$

for any inner product $<,>$ and perturbation $\delta\mathbf{x}_0$. This is derived from the adjoint identity (Lawless, 2013). Using the Euclidean inner product, equation (6.23) is equivalent to

$$(\mathbf{M}_{i,0}\delta\mathbf{x}_0)^T(\mathbf{M}_{i,0}\delta\mathbf{x}_0) = \delta\mathbf{x}_0^T(\mathbf{M}_{i,0}^T(\mathbf{M}_{i,0}\delta\mathbf{x}_0)). \tag{6.24}$$

We evaluated the left hand side and right hand side of this identity for differing values of $\mathbf{x}_0$ and size of perturbation $\delta\mathbf{x}_0$ and showed that they were equal to machine precision.

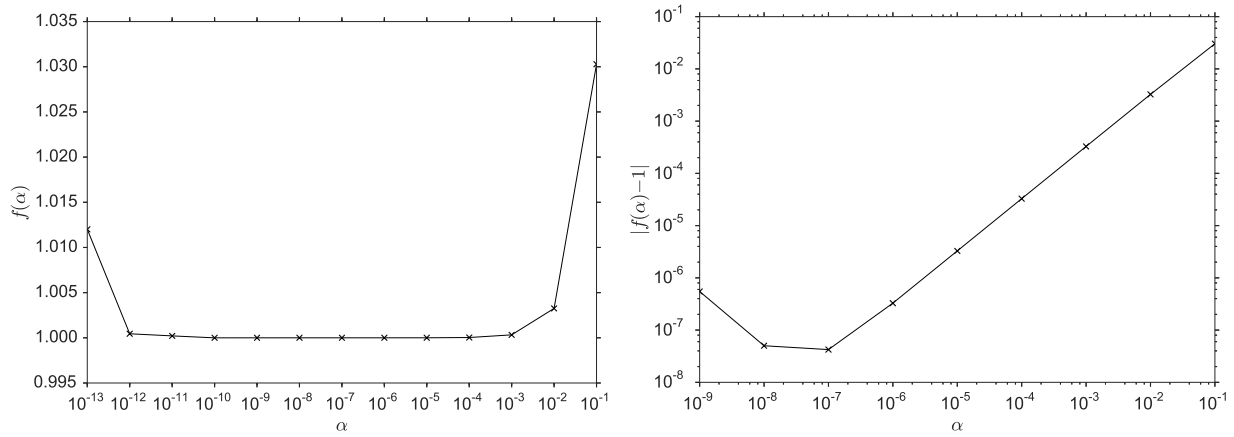
6.3.4.3 Gradient test

The 4D-Var system we have developed passes tests for the gradient of the cost function (Navon et al., 1992). In the implementation of the cost function and its gradient we regularise the problem using a variable transform (Freitag et al., 2010). For the cost function J and its gradient ∇J we can show that we have implemented ∇J correctly using the identity,

$$f(\alpha) = \frac{|J(\mathbf{x}_0 + \alpha\mathbf{b}) - J(\mathbf{x}_0)|}{\alpha\mathbf{b}^T\nabla J(\mathbf{x}_0)} = 1 + O(\alpha), \tag{6.25}$$

where $\mathbf{b}$ is a vector of unit length and α is a parameter controlling the size of the perturbation. For small values of α not too close to machine precision we should have $f(\alpha)$ close to 1. Figure 6.4a shows $f(\alpha)$ for a 365 day assimilation window with $\mathbf{b} = \mathbf{x}_0||\mathbf{x}_0||^{-1}$, we can see that $f(\alpha) \to 1$ as $\alpha \to 0$, as expected until $f(\alpha)$ gets too close to machine zero at order $\alpha = 10^{-11}$. This was also tested with $\mathbf{b}$ in different directions and similar results obtained.

We can also plot $|f(\alpha) - 1|$, where we expect $|f(\alpha) - 1| \to 0$ as $\alpha \to 0$. In figure 6.4b we have plotted $|f(\alpha) - 1|$ for the same conditions as in figure 6.4a, we can see that $|f(\alpha) - 1| \to 0$ as $\alpha \to 0$, as expected. This gives us confidence that the gradient of the cost function is implemented correctly.



(a) $f(\alpha)$ test

(b) $|f(\alpha) - 1|$ test

Figure 6.4: Tests of the gradient of the cost function for a 365 day assimilation window with $\mathbf{b} = \mathbf{x}_0||\mathbf{x}_0||^{-1}$.

6.3.5 Including correlations in the background error covariance matrix

As discussed in section 6.2, including correlations in $\mathbf{B}$ impacts how information from assimilated observations is spread between different types of analysis variables (Bannister, 2008). We explored a number of different methods in order to include parameter-state correlations in $\mathbf{B}$. In this paper we present a method using a set of ecological dynamical constraints, based on expert judgement, on model parameters and state variables from Bloom and Williams (2015). Bloom and Williams (2015) show that implementing these constraints in a Metropolis Hastings MCMC data assimilation routine improves results significantly. The constraints impose conditions on carbon pool turnover and allocation ratios, steady state proximity and growth and the decay of model carbon pools.

In order to create a correlated background error covariance matrix, $\mathbf{B}_{corr}$, using these constraints we create an ensemble of state vectors which we then take the covariance of to give us $\mathbf{B}_{corr}$. To create this ensemble we use the following procedure:

1. Draw a random augmented state vector, $\mathbf{x}_i$, from the multivariate truncated normal distribution described by

$$\mathbf{x}_i \sim \mathcal{N}(\mathbf{x}^b, \mathbf{B}_{diag}), \tag{6.26}$$

   where $\mathbf{B}_{diag}$ is the diagonal matrix described in section 6.3.4 and $\mathbf{x}_i$ is bound by the parameter and state ranges given in table 6.3 in the appendix.

2. Test this $\mathbf{x}_i$ with the ecological dynamical constraints (requiring us to run the DALEC2 model using this state).

3. If $\mathbf{x}_i$ passes it is added to our ensemble, else it is discarded.

Once we have a full ensemble we then take the covariance of the ensemble to find $\mathbf{B}_{corr}$. We chose an ensemble size of 1500 as a qualitative assessment using a larger ensemble showed little difference in correlations. In figure 6.5 we have plotted the correlation matrix or normalised error covariance matrix associated with $\mathbf{B}_{corr}$. This matrix includes both positive and negative correlations between parameter and state variables, with correlations of 1 down the diagonal between variables of the same quantity as expected. The largest positive off-diagonal correlation is 0.42 between f_{lab} and C_{lab}. This makes physical sense as f_{lab} is the parameter controlling the amount of GPP allocated to the labile carbon pool, C_{lab}.

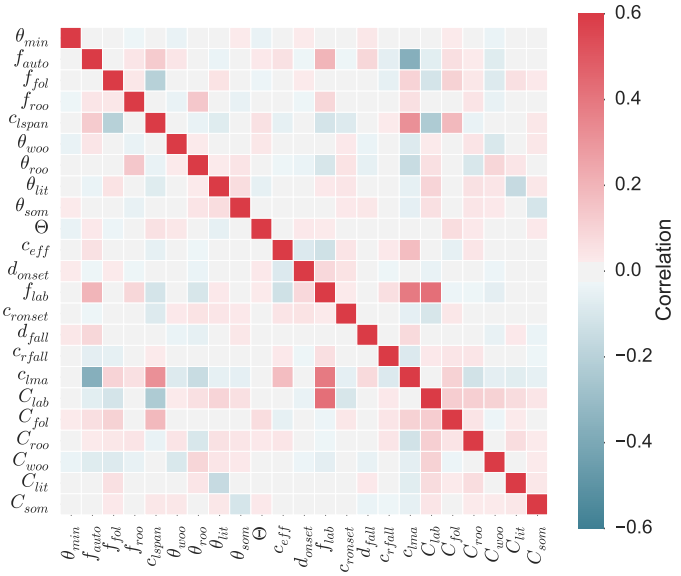


Figure 6.5: Background error correlation matrix created using method in section 6.3.5. Here the correlation scale for off-diagonal values ranges from -0.5 to 0.5 with the correlation along the diagonal being 1. For explanation of parameter and state variable symbols see table 6.3.

6.3.6 Specifying serial correlations in the observation error covariance matrix

The observation error covariance matrix does not only represent the instrumentation error for an observation but also the error in the observation operator (mapping the model state to the observation) and representativity error (error arising from the model being unable resolve the spatial and temporal scales of the observations). These other sources of error represented in $\hat{\mathbf{R}}$ can also lead to correlations between observation errors (Waller et al., 2014). Errors in NEE observations come from different sources such as instrument errors, sampled ecosystem structure from the variable footprint of the flux tower and turbulent conditions (when there is low turbulence and limited air mixing the magnitude of NEE is underestimated). These errors due to turbulence can still have effect even after u^* filtering (Papale et al., 2006b). Due to this dependence on atmospheric conditions we expect the errors in observations of NEE to be serially correlated, as the atmospheric signal itself is serially correlated (Daley, 1992). If we were assimilating half hourly observations of NEE we would expect stronger correlations between observation errors, as atmospheric conditions are more constant at this time scale, with correlations between observation errors getting weaker with lower frequency observations. Although some studies suggest that the correlation between NEE measurement errors on the scale of a day is negligible (Lasslop et al., 2008), it is also likely that error in the observation operator and representativity error will lead to observation error correlations for NEE (Waller et al., 2014).

In section 6.3.3 we have re-written the 4D-Var cost function in equation (6.16) in order to allow the specification of serial observation error correlations in our assimilation scheme. These serial correlations are represented by the off-diagonal blocks of $\hat{\mathbf{R}}$. In work carried out with spatial correlations it has been shown that the structure of the correlation is not critical and that it is better to include some estimate of error correlation structure in the observation error covariance matrix than wrongly assume that errors are independent (Stewart et al., 2013; Healy and White, 2005). As a first attempt we try including temporal correlations on the scale of the observation frequency. We adapt the simple Gaussian model found in Järvinen et al. (1999) (a second order autoregressive correlation function was also tested but is not presented here). The correlation r between 2 observations at times t_1 and t_2 is given as,

$$r = \begin{cases} a\exp\left[\frac{-(t_1-t_2)^2}{\tau^2}\right] + (1-a)\delta_{t_1-t_2} & |t_1 - t_2| \leq \eta \\ 0 & \eta < |t_1 - t_2| \end{cases}, \tag{6.27}$$

where τ is the e-folding time in days, a controls the strength of correlation, δ is the Kronecker delta and η is the cut off time after which the correlation between two observation errors is zero. We have incorporated a cut off for correlations between observation errors as the assumed correlation length scale for the assimilated observations is short. This cut off along with the form of correlation function using the Kronecker delta helps ensure $\hat{\mathbf{R}}$ is positive definite and therefore invertible, as required in the assimilation process. The standard deviation assumed in the observations of NEE is 0.5 $\text{gCm}^{-2}\text{day}^{-1}$ as described in section 6.3.4.

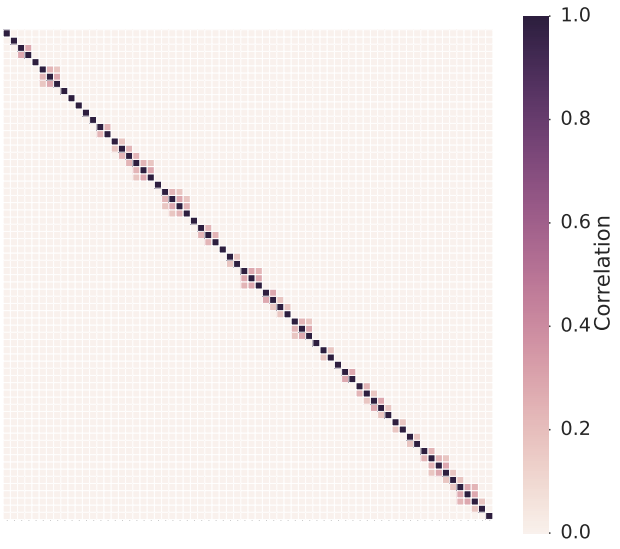


Figure 6.6: Observation error correlation matrix for the 67 observations used in assimilation created using method in section 6.3.6 with $\tau = 4$, $a = 0.3$ and $\eta = 4$.

Figure 6.6 shows the correlation matrix for $\hat{\mathbf{R}}$ created using equation (6.27). Here observations made on adjacent days will have an error correlation of 0.3; this will then decay exponentially for observations farther apart in time. There are 67 NEE observations in this one year assimilation window. These observations are not all on adjacent days and this is evident in the structure of $\hat{\mathbf{R}}$. The effect of the short e-folding time chosen here ($\tau = 4$) provides the desired structure.

6.4 Results

6.4.1 Experiments

In the following sections we present the results of four experiments where we vary the representations of $\mathbf{B}$ and $\hat{\mathbf{R}}$ while assimilating the same NEE observations in the window from the beginning of January 1999 to the end of December 1999. As shown in figure 6.3 the performance of the tangent linear model deteriorates after the first year. We then forecast the NEE over the next 14 years (Jan 2000 - Dec 2013) and compare with the observed data. Using this shorter analysis window with a long forecast allows us to see the effect of including correlations in the error statistics more clearly, as we have a longer time-series of data with which to judge our forecast after data assimilation. These experiments are outlined in table 6.1 where $\mathbf{B}_{diag}$ and $\hat{\mathbf{R}}_{diag}$ are the diagonal matrices of the parameter and state variances and the observations variances respectively and $\mathbf{B}_{corr}$ and $\hat{\mathbf{R}}_{corr}$ are the matrices as specified in section 6.3.5 and section 6.3.6 respectively.

Experiment	$\mathbf{B}_{diag}$	$\hat{\mathbf{R}}_{diag}$	$\mathbf{B}_{corr}$	$\hat{\mathbf{R}}_{corr}$
A	×	×		
B		×	×	
C	×			×
D			×	×

Table 6.1: The combination of error covariance matrices used in each data assimilation experiment.

6.4.2 Experiment A

In this experiment $\mathbf{B}_{diag}$ and $\hat{\mathbf{R}}_{diag}$ were used in the assimilation as described in section 6.4.1. Because these contain no correlations this experiment forms the baseline by which the subsequent results from assimilation experiments are judged.

Figure 6.7a shows assimilation and forecast results for NEE. We can see that assimilating the

observations of NEE has improved the background with the analysis trajectory (green line) fitting well with the observations during the assimilation window (Jan 1999- Dec 1999). The analysis trajectory then diverges in the forecast (Jan 2000 - Dec 2013). This can be seen more clearly in figue 6.8a, where there is an over prediction of respiration in the winter and the seasonal cycle does not match that of the observations. This is also shown in figure 6.9a where we have plotted the model-data differences for a year's period averaged over the 14 years in the forecast period. Figure 6.9a shows that the largest errors in our posterior model forecast occur as a result of not capturing the phenology of the site correctly, in particular the start of the season from April to June.

To see how well the forecast performs after assimilation we show a scatter plot of modelled NEE against observed NEE in figure 6.10b. From table 6.2 the predictions have a Root-Mean-Square Error (RMSE) of 4.22 $\text{gCm}^{-2}\text{day}^{-1}$ and a bias of -0.3 $\text{gCm}^{-2}\text{day}^{-1}$ for the forecast of NEE, whereas the analysis (Jan 1999 - Dec 1999) has a RMSE of 1.36 $\text{gCm}^{-2}\text{day}^{-1}$ and a bias of -0.03 $\text{gCm}^{-2}\text{day}^{-1}$. The background trajectory is the model trajectory for DALEC2 when run using the prior estimate of the parameter and initial state values described in section 6.3.4. The background or prior model trajectory has a RMSE of 3.86 $\text{gCm}^{-2}\text{day}^{-1}$ and a bias of -1.60 $\text{gCm}^{-2}\text{day}^{-1}$ in the analysis window (Jan 1999 - Dec 1999) and the same RMSE of 3.86 $\text{gCm}^{-2}\text{day}^{-1}$ but a bias of -1.36 $\text{gCm}^{-2}\text{day}^{-1}$ during the forecast period (Jan 2000 - Dec 2013). Although using $\mathbf{B}_{diag}$ and $\hat{\mathbf{R}}_{diag}$ in the assimilation has considerably reduced the RMSE in the analysis period, it has also increased the RMSE in the forecast of NEE. However it has reduced the bias in the model forecast considerably from -1.36 $\text{gCm}^{-2}\text{day}^{-1}$ to -0.3 $\text{gCm}^{-2}\text{day}^{-1}$. The bias in the background is due to the background model predicting less negative values of NEE than observed (i.e. above the 1:1 line shown in figure 6.10a). This leads to considerably worse results for the background trajectory than the analysis and its forecast for total forest carbon uptake. It is important to compare our results here with the background trajectory. The background acts as our initial prior model estimate and is the starting point for our minimisation in 4D-Var. Comparing our assimilation results with our background trajectory give us confidence that our 4D-Var scheme is improving the results of our model after assimilation.

6.4.3 Experiment B

Here $\mathbf{B}_{corr}$ (as defined in section 6.3.5) and $\hat{\mathbf{R}}_{diag}$ are used in the assimilation. Figure 6.7b shows assimilation and forecast results for NEE. In figure 6.8b we can see that the forecast performs considerably better than in experiment A, with the analysis trajectory no longer over predicting winter respiration and matching the observed seasonal cycle of NEE more closely in the forecast period (Jan 2000 - Dec 2013). This can be seen more clearly in figure 6.9b where the improvement in the period April-June is considerable as we capture green-up at the site more closely. Even though we have improved the representation of leaf-on in our model significantly here we can see from figure 6.8b that this is still where we have the largest uncertainty for our model after assimilation. From figure 6.10c and table 6.2 we see that the forecast RMSE has almost halved (now 2.56 $\text{gCm}^{-2}\text{day}^{-1}$) with a reduction in bias also, now -0.2 $\text{gCm}^{-2}\text{day}^{-1}$. In comparison using $\mathbf{B}_{corr}$ in the assimilation very slightly degrades the fit for the analysis (Jan 1999 - Dec 1999), with a RMSE of 1.42 $\text{gCm}^{-2}\text{day}^{-1}$ and a bias of -0.04 $\text{gCm}^{-2}\text{day}^{-1}$, as shown in table 6.2.

As discussed in section 6.2 previous work has shown the importance of specifying parameter-state correlations when using variational data assimilation for joint parameter and state estimation (Smith et al., 2009). In 4D-Var the initial correlation structure is evolved implicitly through time. However, in order to make full use of the observations it is essential to specify an accurate estimate to the initial correlation structure. Therefore by not specifying these correlations in experiment A we allow the parameter and state variables to attain unrealistic values in order to find the best fit to the observations in the analysis window (Jan 1999 - Dec 1999), leading to the divergence seen in the forecast (1999-2014) in experiment A.

We can see the effect that including correlations in $\mathbf{B}$ has on the analysis update in figure 6.11. For some variables including correlations in $\mathbf{B}$ has had a large impact on the analysis update after assimilation. This is particularly clear for the f_{lab} parameter. The largest positive off-diagonal correlation in $\mathbf{B}_{corr}$ is between C_{lab} and f_{lab}, with f_{lab} also having a large positive correlation with c_{lma} as shown in section 6.3.5. The effect of these correlations has been to change the analysis increment for f_{lab} from being slightly positive in experiment A to being strongly negative by following the analysis update of its correlated variables C_{lab} and c_{lma}. From figure 6.11 we can also see some of the possible reasons for the improved fit to the observations in experiment B. From figure 6.9a the largest errors in our model forecast of NEE in experiment A stem from a misrepresentation of the phenology of the site in the months April-June. We see that the parameter controlling day of leaf

on (d_{onset}) has been updated slightly differently in comparison to experiment A, with day of leaf on now being slightly later in the year (day 124 instead of 119), again this is due to the included correlations in $\mathbf{B}$. Even this small change in d_{onset} appears to reduce the errors at the start of the season for experiment B as seen from figure 6.9b. The forecast is also no longer over predicting winter respiration to the same extent as in experiment A. From figure 6.11 we see that the main parameters controlling ecosystem respiration in NEE (f_{auto}, θ_{lit}, θ_{som}, Θ) have been reduced in comparison with experiment A, which we believe have led to an improved fit to observations in experiment B. In experiment A we also had an over prediction of peak carbon uptake in summer which has been improved in this experiment. From figure 6.11 we see that one of the parameters controlling the magnitude of gross primary productivity (c_{eff}) has been decreased in comparison to experiment A. This appears to lead to less extreme predictions of peak summer carbon uptake than in experiment A. Two parameters with a significant change from experiment A are f_{fol} and C_{lit}; however in Chuter (2013) the DALEC model prediction of NEE is shown to be largely insensitive to variations in these parameters.

The added constraints provided by the correlations in $\mathbf{B}_{corr}$ acts to regularise the data assimilation problem and avoid overfitting to the assimilated data by limiting the parameter space of the problem (Smith et al., 2009). These correlations have been diagnosed using the EDC's from Bloom and Williams (2015), as shown in section 6.3.5, and help to limit unrealistic behaviour for a mature forest site. Although this has led to a slightly degraded fit to the observations in the analysis window (Jan 1999 - Dec 1999) it has also significantly improved the fit to observations for the forecast (Jan 2000 - Dec 2013).

6.4.4 Experiment C

Here we use $\mathbf{B}_{diag}$ and $\hat{\mathbf{R}}_{corr}$ (as defined in section 6.3.6) in the assimilation. Results shown in figure 6.7c, 6.8c and 6.9c appear similar to those in section 6.4.2 however there are some differences. From table 6.2 and figure 6.10d we see a slight reduction in RMSE for the forecast (now 4.09 $\mathrm{gCm^{-2}day^{-1}}$) in comparison with experiment A. As in experiment B the fit to the observations in the analysis window (Jan 1999 - Dec 1999) is very slightly degraded as the added correlations in $\hat{\mathbf{R}}_{corr}$ act to reduce the weight of the observations in the assimilation (Järvinen et al., 1999). The changes seen when using $\hat{\mathbf{R}}_{corr}$ in the assimilation are less than when using $\mathbf{B}_{corr}$ as the correlations specified in $\hat{\mathbf{R}}_{corr}$ are on a short timescale and much weaker than those in $\mathbf{B}_{corr}$. In

figure 6.11 we can see that the changes between experiment A and C in the analysis increment are much less than when using $\mathbf{B}_{corr}$.

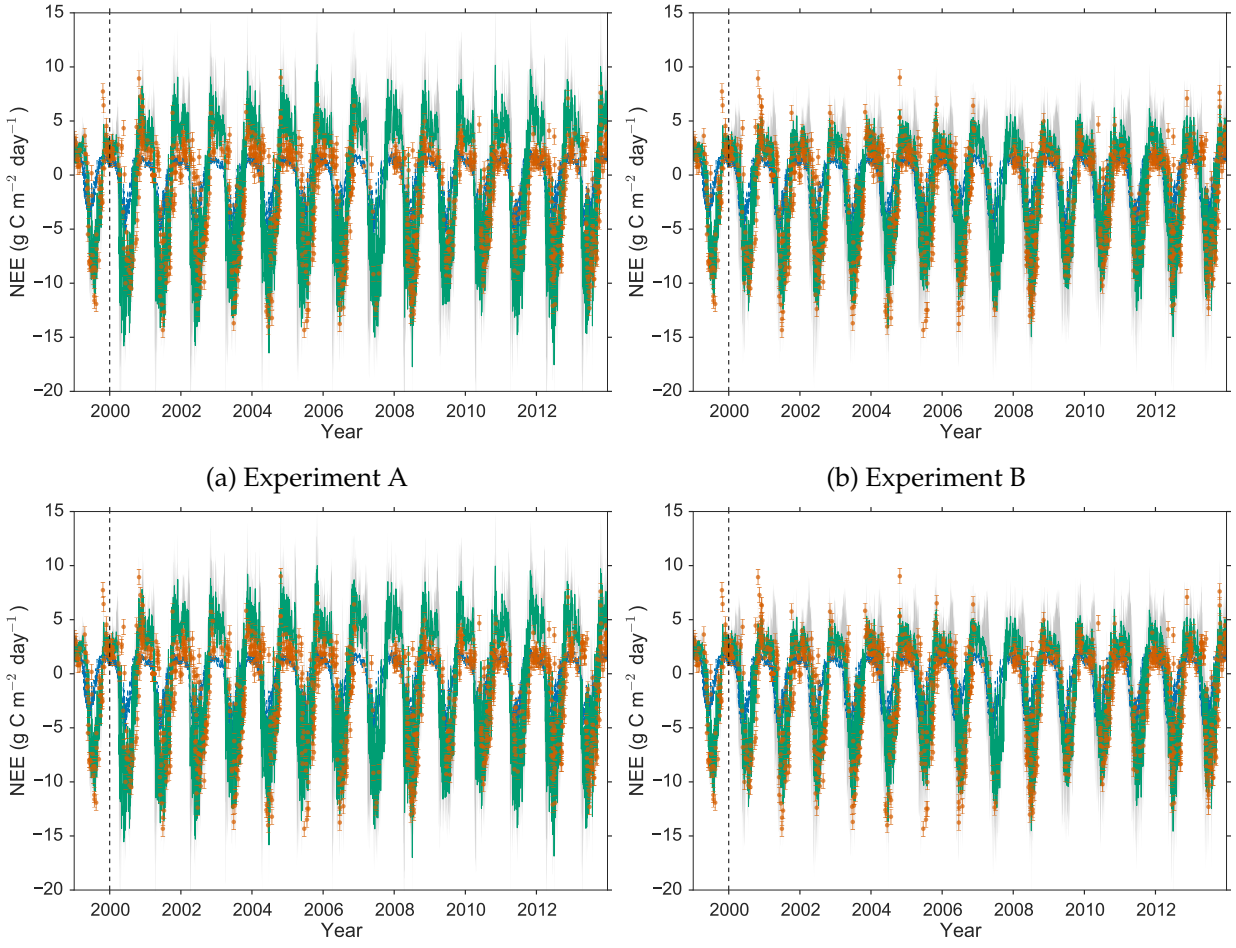
We also expect that specifying time correlations in $\hat{\mathbf{R}}$ will help when assimilating other less frequently sampled data streams along with NEE as the serial correlations reduce the weight given to the mean of the more frequently sampled observations (here NEE) and also reduce the information content of these more frequently sampled observations (Järvinen et al., 1999; Daley, 1992), meaning that less frequently sampled data streams can have more impact on the assimilation.

6.4.5 Experiment D

In the final experiment we use $\mathbf{B}_{corr}$ and $\hat{\mathbf{R}}_{corr}$ in the assimilation. Figure 6.8d, figure 6.8b and 6.9a shows that using both correlated matrices gives similar results as experiment B when $\mathbf{B}_{corr}$ is used with $\hat{\mathbf{R}}_{diag}$. However using $\hat{\mathbf{R}}_{corr}$ in addition to $\mathbf{B}_{corr}$ provides similar improvements as in experiment C. From table 6.2 and figure 6.10e we see the forecast RMSE is slightly reduced from results in experiment B to 2.38 $\text{gCm}^{-2}\text{day}^{-1}$. Using both matrices appears to combine the beneficial effects described in both section 6.4.3 and section 6.4.4. In figure 6.11 we can see that the analysis increment is very similar to experiment B.

6.4.6 Summary

In our experiments we have shown that both $\mathbf{B}_{corr}$ and $\hat{\mathbf{R}}_{corr}$ have the effect of improving the model forecast of NEE. As it can be difficult to inspect the skill of a certain model by only plotting model trajectories, in figure 6.12 we show Taylor diagrams displaying a statistical comparison of the four experiment and background analysis (Jan 1999 - Dec 1999) and forecast (Jan 2000 - Dec 2013) results with the observations of NEE. Here the radial distances from the origin to the points are proportional to the standard deviations of the observations and modelled observations and the azimuthal positions give the correlation coefficient between the modelled and observed NEE (Taylor, 2001). If a model predicted the observations perfectly it would have a correlation coefficient of 1 and a radial distance matching that of the observations (represented by the dotted line). Figure 6.12a shows that all the experiments give very similar results in the analysis window (Jan 1999 - Dec 1999) with all the experiment points closely grouped on top of each other, whereas figure 6.12b shows the significant difference between the experiment results in the forecast (Jan



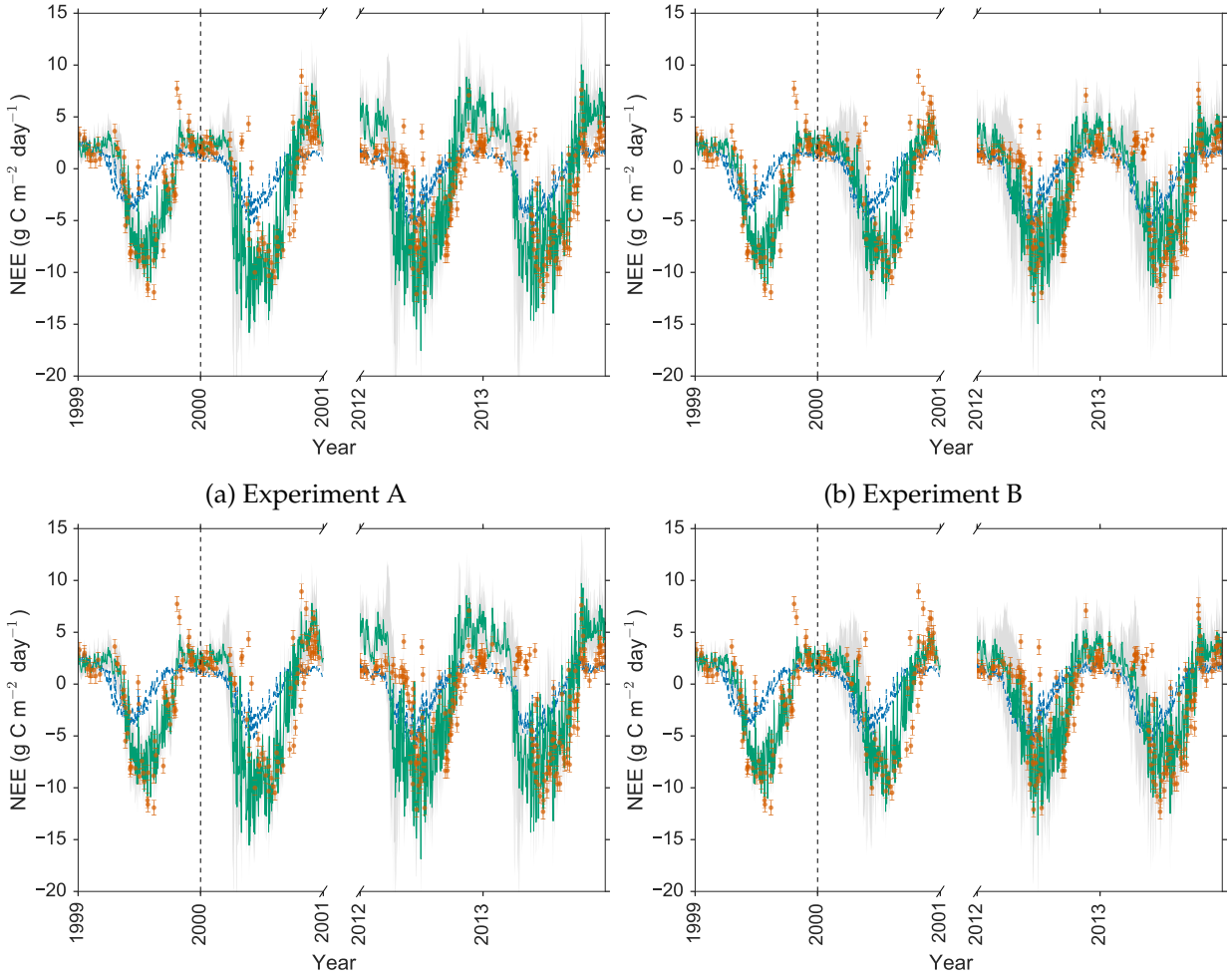
(a) Experiment A

(b) Experiment B

(c) Experiment C

(d) Experiment D

Figure 6.7: One year assimilation and fourteen year forecast of Alice Holt NEE with DALEC2, blue dotted line: background model trajectory, green line: analysis and forecast after assimilation, grey shading: Error in model after assimilation (+/- 3 standard deviations), red dots: observations from Alice Holt flux site with error bars.



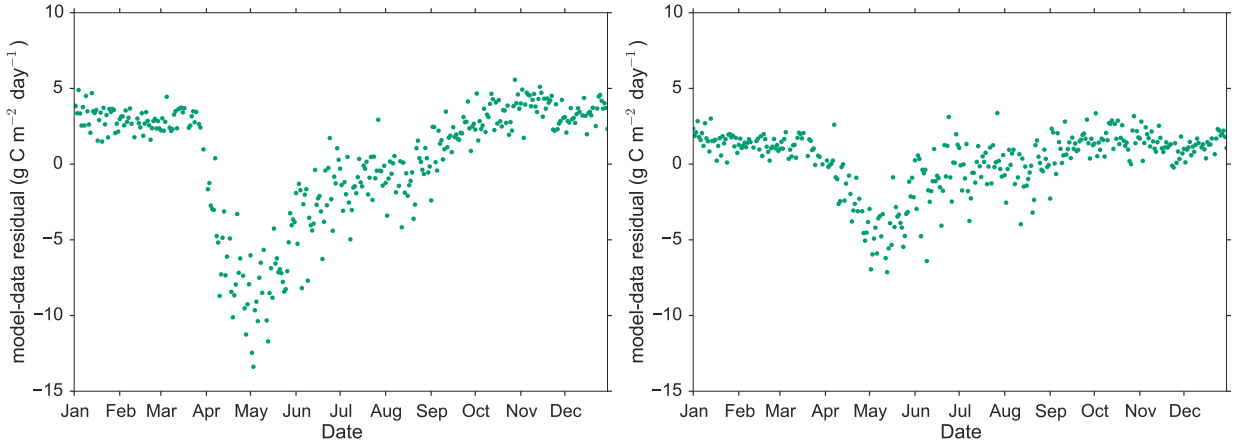
(a) Experiment A

(b) Experiment B

(c) Experiment C

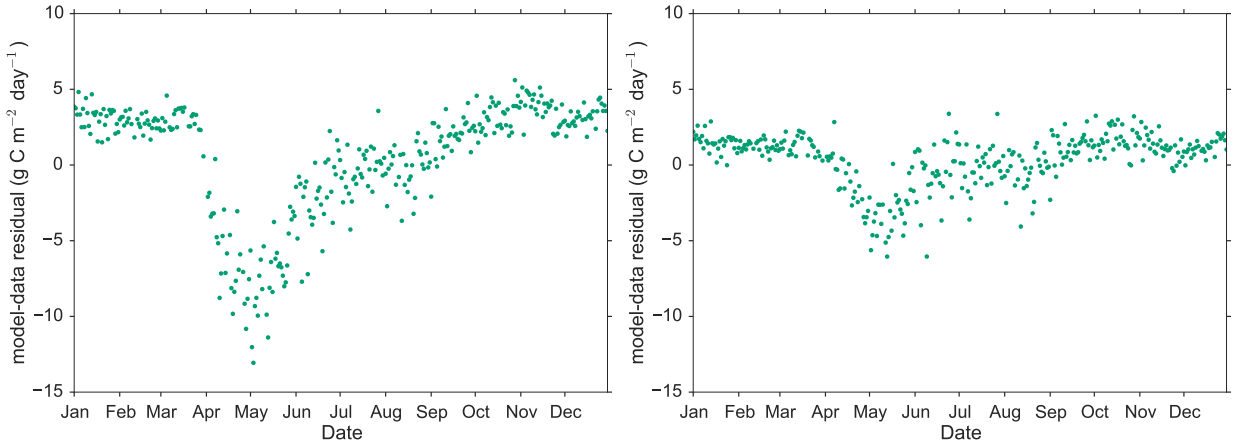
(d) Experiment D

Figure 6.8: As figure 6.7 but only showing the first and final two years results from the one year assimilation and fourteen year forecast of Alice Holt NEE with DALEC2, blue dotted line: background model trajectory, green line: analysis and forecast after assimilation, grey shading: Error in model after assimilation (+/- 3 standard deviations), red dots: observations from Alice Holt flux site with error bars.



(a) Experiment A

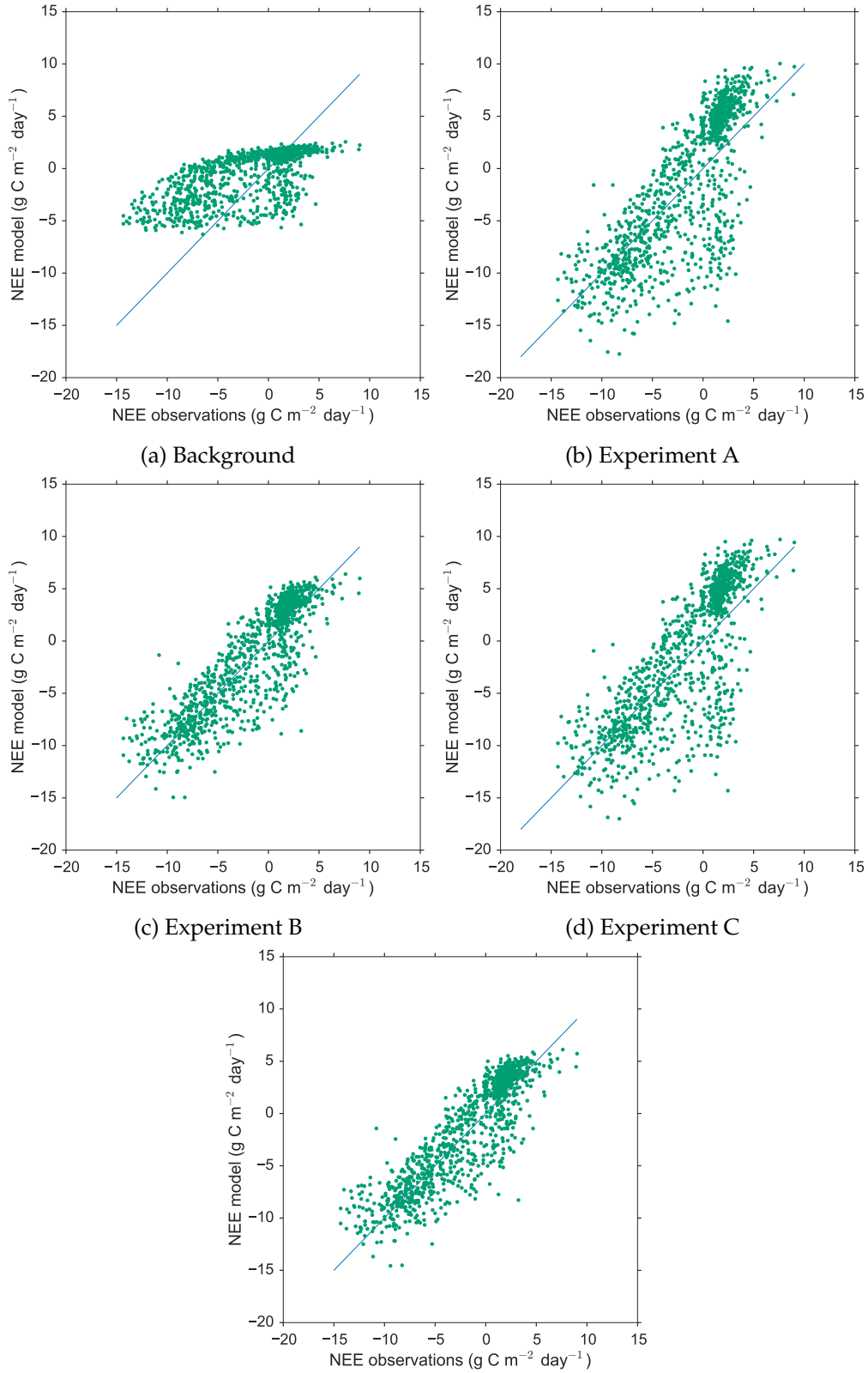
(b) Experiment B



(c) Experiment C

(d) Experiment D

Figure 6.9: Net ecosystem exchange model-data differences for the four experiments. Here each point corresponds to the mean model-data difference for that day of the year over the 14 year model forecast (Jan 2000 - Dec 2013).



(a) Background

(b) Experiment A

(c) Experiment B

(d) Experiment C

(e) Experiment D

Figure 6.10: Forecast scatter plot of modelled daily NEE vs. observations for Jan 2000 - Dec 2013 (green dots). Blue line represents the 1-1 line.

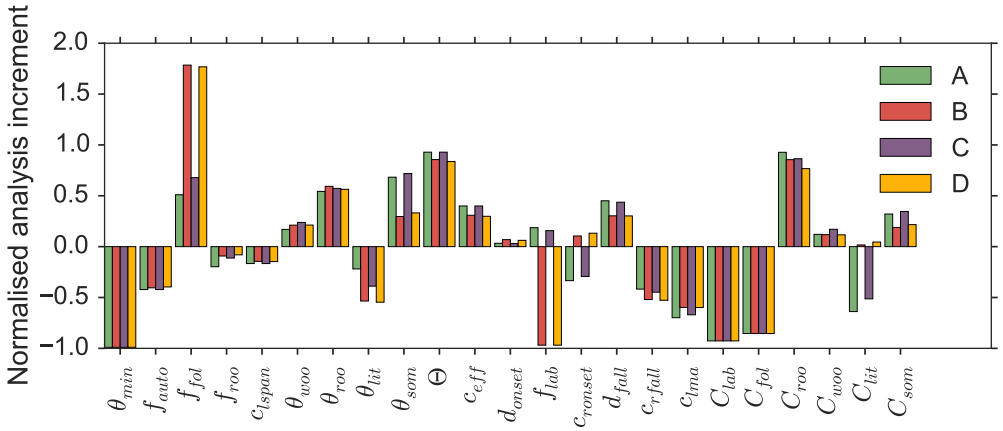
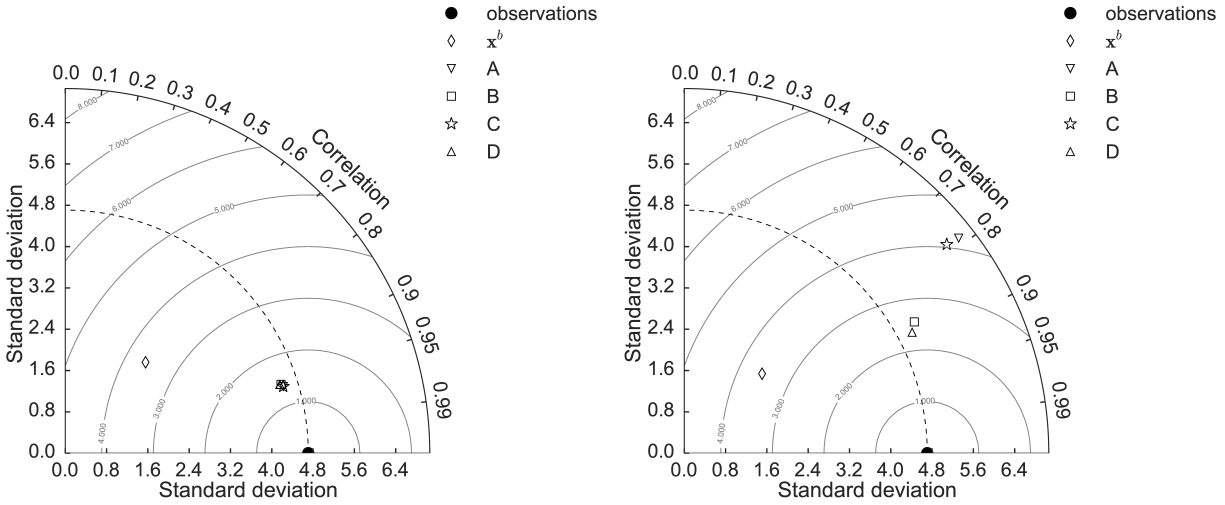


Figure 6.11: Normalised analysis increment $\left(\frac{(\mathbf{x}^a - \mathbf{x}^b)}{\mathbf{x}^b}\right)$ for the four experiments. Explanation of parameter and state variable symbols in table 6.3.

2000 - Dec 2013), with experiments B and D being closer to the dotted line. In all our experiments we find that θ_{min}, C_{lab} and C_{fol} reach the bounds after assimilation. In the case of θ_{min} this is most likely due to the fact that we do not have enough information to recover this parameter when only assimilating observations of NEE, as the DALEC model prediction of NEE is insensitive to variations in this parameter (Chuter, 2013). Assimilating more distinct data streams could help avoid this edge-hitting behaviour. For C_{lab} and C_{fol} this could suggest a flaw in the model or the fact that the prescribed bounds need to be relaxed slightly for the studied ecosystem. Our hypothesis is that the mechanism by which C_{lab} is distributed to the leaves is over simplified; we intend to test this in future work. In table 6.4 we show the standard deviations for our parameter and state variables after assimilation. We can see that we have improved our confidence for most of these variables after assimilation when compared with the standard deviations in table 6.3.

6.5 Discussion

In this paper we have implemented the DALEC2 functional ecology model in a 4D-Var data assimilation scheme, building an adjoint of the DALEC2 model and applying rigorous tests to our scheme. Using 4D-Var can provide much faster assimilation results than MCMC techniques as we have knowledge of the derivative of the model. For our experiments the 4D-Var routine has taken in the order of 10^2 function evaluations to converge to a minimum, whereas MCMC techniques using the same model take in the order of 10^8 function evaluations (Bloom and Williams, 2015). However, we do assume that the statistics of the problem are Gaussian whereas MCMC techniques do not. We have shown that 4D-Var is a valid tool for improving the DALEC2 model



(a) Analysis (Jan 1999 - Dec 1999)

(b) Forecast (Jan 2000 - Dec 2013)

Figure 6.12: Taylor diagrams displaying statistical comparison of the four experiment and background analysis (Jan 1999 - Dec 1999) and forecast (Jan 2000 - Dec 2013) results with observations of NEE ($\mathrm{gCm^{-2}day^{-1}}$). The dotted line represents the standard deviation of the observations and the contours represent values of constant root mean square error between model and observations.

estimate of NEE and that even when assimilating only a single year of NEE observations we can improve the forecast significantly. If more than one year was required, this type of data assimilation routine could be run in cycling mode, allowing for the assimilation of multiple years of data (Moodey et al., 2013). This also avoids any possible unstable behaviour associated with much longer single assimilation windows. However, here our aim is to investigate the effect of specifying correlations in background and observation error statistics on the forecast of NEE. We have therefore assimilated just one year of NEE observations and produced a long 14 year forecast in order to see more clearly the effect of including these correlations on the forecast when judging against observations. The observations of daily NEE from the Alice Holt flux site are quite variable year to year, peak summer uptake varies from -14.35 $\mathrm{gCm^{-2}day^{-1}}$ to -9.04 $\mathrm{gCm^{-2}day^{-1}}$, and therefore provide a reasonable test for the ability of the DALEC2 model forecast, especially over a 14 year period.

We then considered the nature of background and observation errors. The effect of specifying parameter-state correlations in the background information and serial correlations between the observation errors was explored.

The technique presented here to specify $\mathbf{B}_{corr}$ has been shown to have significantly improved forecasts of NEE over using a diagonal representation of $\mathbf{B}$. In section 6.4.3 we discuss how the correlations in $\mathbf{B}_{corr}$ impact the analysis update for the parameter and state variables (see figure 6.11)

Analysis (Jan 1999 - Dec 1999)			
Experiment	RMSE ($gCm^{-2}day^{-1}$)	Bias ($gCm^{-2}day^{-1}$)	Correlation coefficient
Background	3.86	-1.60	0.70
A	1.36	-0.03	0.96
B	1.42	-0.04	0.95
C	1.37	-0.09	0.96
D	1.43	-0.09	0.95
Forecast (Jan 2000 - Dec 2013)			
Experiment	RMSE ($gCm^{-2}day^{-1}$)	Bias ($gCm^{-2}day^{-1}$)	Correlation coefficient
Background	3.86	-1.36	0.66
A	4.22	-0.30	0.79
B	2.56	-0.20	0.87
C	4.09	-0.51	0.78
D	2.38	-0.33	0.88

Table 6.2: Analysis (Jan 1999 - Dec 1999) and forecast (Jan 2000 - Dec 2013) results for experiments and background when judged against observed NEE.

causing the seasonal cycle of carbon uptake and magnitude of fluxes to fit more closely with the observations than when using a diagonal **B** in the assimilation algorithm. These results agree with those of Smith et al. (2009) where the importance of specifying parameter-state correlations when performing joint parameter and state estimation with variational data assimilation was shown. The added constraint provided by including correlations in the prior error covariance matrix, **B**, acts to regularise the assimilation problem. Hence, the parameter and initial state values we retrieve from our data assimilation are more likely to be realistic, leading to better insight into the studied system. For example we see from figure 6.11 that when using $\mathbf{B}_{corr}$ in our assimilation we find a much longer labile release period (c_{ronset}) than when using $\mathbf{B}_{diag}$. This means that the period of green-up in our study site is possibly much longer than we would have estimated had we based our analysis on our assimilation results using a matrix **B** with no correlations. The method for specifying $\mathbf{B}_{corr}$ in this paper used a series of ecological dynamical constraints taken from Bloom and Williams (2015). Implementing correlations in the prior error statistics in this way may prove difficult for models where these type of constraints are not available; however there are other methods to build correlations into **B**. One technique we also tested (not presented here) to create a correlated **B** involved evolving an ensemble of state vectors over the length of the chosen assimilation window using the model (DALEC2) and then taking the covariance of the evolved ensemble. This gave us a **B** with parameter-state and state-state correlations, but no parameter-parameter correlations as the parameters are not updated by the model. Using the **B** created with this method also improved assimilation results significantly over using a diagonal

B. A larger number of different tests were run using different background vectors and variances and it was found that specifying some form of correlation structure in **B** always made an improvement to the results of the assimilation. As this work has only considered a single deciduous site, it would be useful to apply the techniques detailed here for an evergreen site. Evergreen ecosystems usually have less extreme seasonal variation, it will therefore be of interest to see if a similar improvement for evergreen ecosystem forecast results is found when using a $\mathbf{B}_{corr}$ created using the same method.

The purpose of this exercise was to see how well we could forecast NEE while also investigating the effect of including correlations between error statistics. It was not an attempt to recover all the parameters and state variables with a high level of accuracy. However, it is still instructive to look at these values and compare with data where available. In Meir et al. (2002) an observed range is given for leaf mass per area (c_{lma}) for the Alice Holt flux site of between 40 to 80 gCm^{-2}. The background value for c_{lma} in our experiments is 128.5 gCm^{-2}. When using diagonal error covariance matrices in experiment A we find a value of 38.7 gCm^{-2} for c_{lma} after assimilation which is almost within the range given by Meir et al. (2002). In experiment D when using error covariance matrices including correlations c_{lma} has a value of 51.6 gCm^{-2} after assimilation, this is well within the observed range given by Meir et al. (2002). From observations made by Forest Research we also have estimates of the above and below ground woody carbon pool (C_{woo}) at the start of 1999, with an observed value of 14258 gCm^{-2}. It is not clear how uncertain this estimate is. The background value for C_{woo} in our experiments is 6506 gCm^{-2}. When using diagonal error covariance matrices in experiment A we find a value of 7291 gCm^{-2} for C_{woo}, an increase but still far away from the observed estimate. In experiment D when using error covariance matrices including correlations C_{woo} has a value of 7262 gCm^{-2} a similar result as experiment A. Here the assimilation has not been able to recover a value of C_{woo} similar to that of the observed estimate. This is not necessarily of concern as we are not able to quantify the error in this observation. Also we are assimilating observations of daily NEE only; NEE is the difference between Gross Primary Productivity (GPP) and Total ecosystem respiration (RT), (NEE = RT - GPP), with neither GPP nor RT being direct functions of C_{woo}. Therefore it is unlikely that we will recover an accurate value of C_{woo}, as the assimilated observations are not significantly impacted by large changes in this state variable; this result is also discussed in Fox et al. (2009). This may also explains why we are able to recover a reasonable value of c_{lma} from the assimilation, as from equation (6.1) we can see that c_{lma} is one of the input arguments taken by GPP. The function calculating NEE will therefore be

sensitive to variations in the c_{lma} parameter and so assimilating observations of NEE could help to constrain this parameter.

In numerical weather prediction it has been shown that including correlations in $\mathbf{R}$ can help improve data assimilation results (Weston et al., 2014; Stewart et al., 2013). However the specified correlations have most commonly been satellite interchannel correlations with observations errors still being considered independent in time. In this paper we have shown that including correlations between observation errors in time can also improve data assimilation results, here providing a slight improvement for the DALEC2 model forecast of NEE. Here we only see a small impact on our results when using $\hat{\mathbf{R}}_{corr}$ in the assimilation as the correlations we have included are weak (especially in comparison to those included in $\mathbf{B}_{corr}$) and on a short time-scale. We expect including correlations in $\hat{\mathbf{R}}$ will have more of an impact on data assimilation results when assimilating data with stronger error correlations (i.e. finer temporal-resolution observations). We also expect including these serial correlations to have an even greater impact when assimilating more than one data stream as discussed in section 6.2. Using the form of $\hat{\mathbf{R}}$ given in this paper for specifying serial correlations will also allow us to specify serial correlations between different observation types. When running the DALEC2 model with a day-night time step, instead of the daily time step used for this paper, this will allow us to build in the type of correlations investigated by Baldocchi et al. (2015) between ecosystem respiration and canopy photosynthesis. More work is needed to investigate the effect of including correlations between observations error statistics when assimilating multiple data streams.

The $\hat{\mathbf{R}}_{corr}$ presented in this paper has a weak correlation ($a = 0.3$ as shown in section6.3.6) in time between observations of NEE, this representation of $\hat{\mathbf{R}}_{corr}$ has slightly improved the model forecast of NEE. However other choices of $\hat{\mathbf{R}}_{corr}$ (with much stronger correlations between observations) tested for this paper degraded the forecast. This is probably due to the specified correlations being unrealistic and highlights the fact that a reasonable estimate of the true correlation structure for $\hat{\mathbf{R}}_{corr}$ is needed to have a positive impact on results. The development of a more diagnostic approach for the calculation of serial correlations in $\hat{\mathbf{R}}$ would be useful. One option would be to adapt the Desroziers et al. (2005) diagnostic, which has been used successfully in numerical weather prediction for diagnosing observation error correlations for observations taken at the same time (Weston et al., 2014), and extending this technique to diagnose serial correlations.

6.6 Conclusion

Functional ecology and land surface model data assimilation routines largely treat prior estimates of parameter and state uncertainties and observation error statistics as independent and uncorrelated. In this paper we have shown the importance of including estimates of such correlations, especially between background parameter and state error statistics when performing joint parameter and state estimation.

When performing joint parameter and state estimation including correlations in the background error covariance matrix significantly improves the forecast after assimilation, in comparison to using a diagonal representation of $\mathbf{B}$. Specifying serial time correlations between observation errors in $\hat{\mathbf{R}}$ also improves the forecast and we expect these correlations to have a greater impact when assimilating more than one data stream. More work is needed to investigate the effect of including these correlations when assimilating multiple data streams. The development of a more diagnostic tool for the calculation of the error correlation structure in $\hat{\mathbf{R}}$ is also important.

When including both parameter-state correlations in $\mathbf{B}$ and time correlations between observation errors in $\hat{\mathbf{R}}$ and assimilating only a single year of NEE observations we can forecast 14 years of NEE observations with a root-mean square error of 2.38 $\mathrm{gCm^{-2}day^{-1}}$ and a correlation coefficient of 0.88. This is a significant 44% reduction in error from the results when using a $\mathbf{B}$ and $\hat{\mathbf{R}}$ with no specified correlations of 4.22 $\mathrm{gCm^{-2}day^{-1}}$ and a correlation coefficient of 0.79.

6.7 Acknowledgements

This work was funded by the UK Natural Environment Research Council (NE/K00705X/1) with a CASE award from the UK Forestry Commission. This work was also partly funded by the National Centre for Earth Observation. We are grateful to Luke Smallman for providing the data from the CARDAMOM system.

6.8 Appendix

Parameter	Description	Background vector ($\mathbf{x}^b$)	Standard deviation	Rang
θ_{min}	Litter mineralisation rate (day^{-1})	9.810×10^{-4}	2.030×10^{-3}	10^{-5} -
f_{auto}	Autotrophic respiration fraction	5.190×10^{-1}	1.168×10^{-1}	0.3 –
f_{fol}	Fraction of GPP allocated to foliage	1.086×10^{-1}	1.116×10^{-1}	0.01 -
f_{roo}	Fraction of GPP allocated to fine roots	4.844×10^{-1}	2.989×10^{-1}	0.01 -
c_{lspan}	Determines annual leaf loss fraction	1.200×10^{0}	1.161×10^{-1}	1.000
θ_{woo}	Woody carbon turnover rate (day^{-1})	1.013×10^{-4}	1.365×10^{-4}	$2.5 \times$
θ_{roo}	Fine root carbon turnover rate (day^{-1})	3.225×10^{-3}	2.930×10^{-3}	10^{-4} -
θ_{lit}	Litter carbon turnover rate (day^{-1})	3.442×10^{-3}	3.117×10^{-3}	10^{-4} -
θ_{som}	Soil and organic carbon turnover rate (day^{-1})	1.113×10^{-4}	1.181×10^{-4}	10^{-7} -
Θ	Temperature dependance exponent factor	4.147×10^{-2}	1.623×10^{-2}	0.018
c_{eff}	Canopy efficiency parameter	7.144×10^{1}	2.042×10^{1}	$10 - 1$
d_{onset}	Leaf onset day (day)	1.158×10^{2}	6.257×10^{0}	$1 - 36$
f_{lab}	Fraction of GPP allocated to labile carbon pool	3.204×10^{-1}	1.145×10^{-1}	0.01 -
c_{ronset}	Labile carbon release period (days)	4.134×10^{1}	1.405×10^{1}	$10 - 1$
d_{fall}	Leaf fall day (day)	2.205×10^{2}	3.724×10^{1}	$1 - 36$
c_{rfall}	Leaf-fall period (days)	1.168×10^{2}	2.259×10^{1}	$10 - 1$
c_{lma}	Leaf mass per area (gCm^{-2})	1.285×10^{2}	6.410×10^{1}	$10 - 4$
C_{lab}	Labile carbon pool (gCm^{-2})	1.365×10^{2}	6.626×10^{1}	$10 - 1$
C_{fol}	Foliar carbon pool (gCm^{-2})	6.864×10^{1}	3.590×10^{1}	$10 - 1$
C_{roo}	Fine root carbon pool (gCm^{-2})	2.838×10^{2}	2.193×10^{2}	$10 - 1$
C_{woo}	Above and below ground woody carbon pool (gCm^{-2})	6.506×10^{3}	7.143×10^{3}	100 –
C_{lit}	Litter carbon pool (gCm^{-2})	5.988×10^{2}	5.450×10^{2}	$10 - 1$
C_{som}	Soil and organic carbon pool (gCm^{-2})	1.936×10^{3}	1.276×10^{3}	100 –

Table 6.3: Parameter values and standard deviations for background vector used in experiments.

Parameter	A	B	C	D
θ_{min}	1.822×10^{-6}	3.742×10^{-7}	1.519×10^{-6}	3.854×10^{-7}
f_{auto}	2.913×10^{-3}	1.428×10^{-3}	2.937×10^{-3}	1.510×10^{-3}
f_{fol}	5.459×10^{-3}	4.581×10^{-3}	6.797×10^{-3}	4.591×10^{-3}
f_{roo}	7.907×10^{-2}	9.141×10^{-3}	8.199×10^{-2}	9.149×10^{-3}
c_{lspan}	4.884×10^{-7}	5.894×10^{-4}	5.304×10^{-7}	5.469×10^{-4}
θ_{woo}	1.849×10^{-8}	8.365×10^{-9}	1.849×10^{-8}	8.365×10^{-9}
θ_{roo}	6.870×10^{-6}	3.494×10^{-6}	7.326×10^{-6}	3.508×10^{-6}
θ_{lit}	3.144×10^{-6}	4.808×10^{-7}	2.242×10^{-6}	4.635×10^{-7}
θ_{som}	1.178×10^{-8}	6.848×10^{-9}	1.210×10^{-8}	6.850×10^{-9}
Θ	7.905×10^{-5}	6.808×10^{-5}	8.010×10^{-5}	6.978×10^{-5}
c_{eff}	3.755×10^{2}	2.625×10^{2}	3.724×10^{2}	2.608×10^{2}
d_{onset}	3.552×10^{1}	3.755×10^{1}	3.649×10^{1}	3.766×10^{1}
f_{lab}	1.220×10^{-2}	3.209×10^{-3}	1.225×10^{-2}	3.203×10^{-3}
c_{ronset}	8.304×10^{1}	1.642×10^{2}	1.100×10^{2}	1.644×10^{2}
d_{fall}	5.992×10^{2}	5.294×10^{1}	5.772×10^{2}	6.145×10^{1}
c_{rfall}	1.540×10^{2}	1.521×10^{2}	1.604×10^{2}	1.599×10^{2}
c_{lma}	2.134×10^{2}	2.209×10^{2}	2.503×10^{2}	2.372×10^{2}
C_{lab}	6.142×10^{2}	5.709×10^{2}	8.586×10^{2}	5.618×10^{2}
C_{fol}	7.971×10^{2}	1.212×10^{2}	8.029×10^{2}	1.285×10^{2}
C_{roo}	3.984×10^{4}	2.539×10^{4}	4.114×10^{4}	2.553×10^{4}
C_{woo}	5.075×10^{7}	2.764×10^{7}	5.075×10^{7}	2.764×10^{7}
C_{lit}	4.157×10^{4}	5.416×10^{4}	7.179×10^{4}	5.532×10^{4}
C_{som}	1.454×10^{6}	1.106×10^{6}	1.482×10^{6}	1.105×10^{6}

Table 6.4: Standard deviations for each experiment after assimilation, calculated using equation 6.19.

Chapter 7

Using data assimilation to understand the effect of disturbance of the carbon dynamics of the Alice Holt forest

Insert Chapter here

Chapter 8

Conclusion

chapter 8 goes here

BIBLIOGRAPHY

Anderson, J. L. and S. L. Anderson, 1999: A monte carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Monthly Weather Review*, **127 (12)**, 2741–2758, doi:10.1175/1520-0493(1999)127¡2741:AMCIOT¿2.0.CO;2.

Bacour, C., et al., 2015: Joint assimilation of eddy-covariance flux measurements and FAPAR products over temperate forests within a process-oriented biosphere model. *Journal of Geophysical Research: Biogeosciences*, n/a–n/a, doi:10.1002/2015JG002966, URL http://doi.wiley.com/10.1002/2015JG002966.

Baldocchi, D., 2008: Turner review no. 15.'breathing'of the terrestrial biosphere: lessons learned from a global network of carbon dioxide flux measurement systems. *Australian Journal of Botany*, **56 (1)**, 1–26.

Baldocchi, D., C. Sturtevant, and F. Contributors, 2015: Does day and night sampling reduce spurious correlation between canopy photosynthesis and ecosystem respiration? *Agricultural and Forest Meteorology*, **207**, 117–126, doi:10.1016/j.agrformet.2015.03.010, URL http://linkinghub.elsevier.com/retrieve/pii/S016819231500088X.

Baldocchi, D., et al., 2001: Fluxnet: A new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities. *Bulletin of the American Meteorological Society*, **82 (11)**, 2415–2434.

Bannister, R. N., 2008: A review of forecast error covariance statistics in atmospheric variational data assimilation. i: Characteristics and measurements of forecast error covariances. *Quarterly Journal of the Royal Meteorological Society*, **134 (637)**, 1951–1970.

Barnett, S. and R. Cameron, 1985: *Introduction to Mathematical Control Theory*. Oxford Applied Mathematics and Computing Science Series, Clarendon Press.

Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525 (7567)**, 47–55.

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

Bloom, A. A., J.-F. Exbrayat, I. R. van der Velde, L. Feng, and M. Williams, 2016: 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.

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

Bonavita, M., E. Hlm, L. Isaksen, and M. Fisher, 2015: The evolution of the ecmwf hybrid data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, n/a–n/a, doi: 10.1002/qj.2652, URL http://dx.doi.org/10.1002/qj.2652.

Booth, B. B. B., et al., 2012: High sensitivity of future global warming to land carbon cycle processes. *Environmental Research Letters*, **7 (2)**, 024 002.

Braswell, B. H., W. J. Sacks, E. Linder, and D. S. Schimel, 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 (2)**, 335–355.

Bréda, N. J., 2003: Ground-based measurements of leaf area index: a review of methods, instruments and current controversies. *Journal of experimental botany*, **54 (392)**, 2403–2417.

Brown, S., 2002: Measuring carbon in forests: current status and future challenges. *Environmental pollution*, **116 (3)**, 363–372.

Cardinali, C., S. Pezzulli, and E. Andersson, 2004: Influence-matrix diagnostic of a data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, **130 (603)**, 2767–2786, doi: 10.1256/qj.03.205, URL http://dx.doi.org/10.1256/qj.03.205.

Carvalhais, N., et al., 2010: 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, doi:10.1111/j.1365-2486.2010.02173.x, URL http://dx.doi.org/10.1111/j.1365-2486.2010.02173.x.

Chuter, A. M., 2013: A Qualitative Analysis of the Data Assimilation Linked Ecosystem Carbon Model , DALEC. Ph.D. thesis, University of Surry, Guildford.

Ciais, P., et al., 2014: Carbon and other biogeochemical cycles. *Climate change 2013: the physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge University Press, 465–570.

Clayton, A. M., A. C. Lorenc, and D. M. Barker, 2013: Operational implementation of a hybrid ensemble/4d-var global data assimilation system at the met office. *Quarterly Journal of the Royal Meteorological Society*, **139 (675)**, 1445–1461, doi:10.1002/qj.2054, URL http://dx.doi.org/10.1002/qj.2054.

Courtier, P., et al., 1998: The ecmwf implementation of three-dimensional variational assimilation (3d-var). i: Formulation. *Quarterly Journal of the Royal Meteorological Society*, **124 (550)**, 1783–1807.

Cover, T. M. and J. A. Thomas, 1991: Elements of information theory. new york: J. *Wiley and Sons*, **5**, 5.

Cox, P. M., R. A. Betts, C. D. Jones, S. A. Spall, and I. J. Totterdell, 2000: Acceleration of global warming due to carbon-cycle feedbacks in a coupled climate model. *Nature*, **408 (6809)**, 184–187, URL http://dx.doi.org/10.1038/35041539.

Dahdouh-Guebas, F. and N. Koedam, 2006: Empirical estimate of the reliability of the use of the point-centred quarter method (pcqm): Solutions to ambiguous field situations and description of the pcqm+ protocol. *Forest Ecology and management*, **228 (1)**, 1–18.

Daley, R., 1992: The Effect of Serially Correlated Observation and Model Error on Atmospheric Data Assimilation. 164–177 pp., doi:10.1175/1520-0493(1992)120¡0164:TEOSCO¿2.0.CO;2.

Dee, D., et al., 2011: The era-interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, **137 (656)**, 553–597.

Delahaies, S., I. Roulstone, and N. Nichols, 2013: A regularization of the carbon cycle data-fusion problem. *EGU General Assembly Conference Abstracts*, Vol. 15, 4087.

Demarty, J., F. Chevallier, A. D. Friend, N. Viovy, S. Piao, and P. Ciais, 2007: Assimilation of global modis leaf area index retrievals within a terrestrial biosphere model. *Geophysical Research Letters*, **34 (15)**, n/a–n/a, doi:10.1029/2007GL030014, l15402.

Desroziers, G., L. Berre, V. Chabot, and B. Chapnik, 2009: A posteriori diagnostics in an ensemble of perturbed analyses. *Monthly Weather Review*, **137 (10)**, 3420–3436.

Desroziers, G., L. Berre, B. Chapnik, and P. Poli, 2005: Diagnosis of observation, background and analysis-error statistics in observation space. *Quarterly Journal of the Royal Meteorological Society*, **131 (613)**, 3385–3396.

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

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

Engelen, R. and G. Stephens, 2004: Information content of infrared satellite sounding measurements with respect to co2. *Journal of Applied Meteorology*, **43 (2)**, 373–378.

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

Exbrayat, J.-f., T. L. Smallman, A. A. Bloom, and M. Williams, 2015: Using a data-assimilation system to assess the influence of fire on simulated carbon fluxes and plant traits for the Australian continent. *EGU General Assembly*, **17**, 6421.

Eyre, J., 1990: The information content of data from satellite sounding systems: A simulation study. *Quarterly Journal of the Royal Meteorological Society*, **116 (492)**, 401–434.

Falkowski, P., et al., 2000: The global carbon cycle: A test of our knowledge of earth as a system. *Science*, **290 (5490)**, 291–296, doi:10.1126/science.290.5490.291.

Fassnacht, K. S., S. T. Gower, J. M. Norman, and R. E. McMurtric, 1994: A comparison of optical and direct methods for estimating foliage surface area index in forests. *Agricultural and Forest Meteorology*, **71 (1)**, 183–207.

Fisher, M., 2003: *Estimation of entropy reduction and degrees of freedom for signal for large variational analysis systems*. European Centre for Medium-Range Weather Forecasts.

Fowler, A. and P. Jan Van Leeuwen, 2012: Measures of observation impact in non-Gaussian data assimilation. *Tellus A*, **64 (0)**, doi:10.3402/tellusa.v64i0.17192, URL http://www.tellusa.net/index.php/tellusa/article/view/17192.

Fowler, A. and P. J. Van Leeuwen, 2013: Observation impact in data assimilation: The effect of non-gaussian observation error. *Tellus, Series A: Dynamic Meteorology and Oceanography*, **65 (1)**, 1–16, doi:10.3402/tellusa.v65i0.20035.

Fox, A., et al., 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 (10)**, 1597–1615.

Freitag, M. A., N. K. Nichols, and C. J. Budd, 2010: L1-regularisation for ill-posed problems in variational data assimilation. *Pamm*, **10 (1)**, 665–668, doi:10.1002/pamm.201010324, URL http://doi.wiley.com/10.1002/pamm.201010324.

Hamill, T. M., J. S. Whitaker, and C. Snyder, 2001: Distance-dependent filtering of background error covariance estimates in an ensemble kalman filter. *Monthly Weather Review*, **129 (11)**, 2776–2790, doi:10.1175/1520-0493(2001)129¡2776:DDFOBE¿2.0.CO;2.

Healy, S. and A. White, 2005: Use of discrete Fourier transforms in the 1D-Var retrieval problem. *Quarterly Journal of the Royal Meteorological Society*, **131 (605)**, 63–72, doi:10.1256/qj.03.193, URL http://doi.wiley.com/10.1256/qj.03.193.

Husch, B., T. W. Beers, and J. A. Kershaw Jr, 2002: *Forest mensuration*. John Wiley & Sons.

Jacquez, J. A. and P. Greif, 1985: Numerical parameter identifiability and estimability: Integrating identifiability, estimability, and optimal sampling design. *Mathematical Biosciences*, **77 (1-2)**, 201–227, doi:10.1016/0025-5564(85)90098-7.

Järvinen, H., E. Andersson, and F. Bouttier, 1999: Variational assimilation of time sequences of surface observations with serially correlated errors. *Tellus A*, **51 (4)**, 469–488.

Johnson, C., B. J. Hoskins, and N. K. Nichols, 2005: A singular vector perspective of 4d-var: Filtering and interpolation. *Quarterly Journal of the Royal Meteorological Society*, **131 (605)**, 1–19.

Jonckheere, I., S. Fleck, K. Nackaerts, B. Muys, P. Coppin, M. Weiss, and F. Baret, 2004: Review of methods for in situ leaf area index determination Part I. Theories, sensors and hemispherical photography. *Agricultural and Forest Meteorology*, **121 (1-2)**, 19–35, doi: 10.1016/j.agrformet.2003.08.027.

Jones, C., et al., 2013: Twenty-first-century compatible co2 emissions and airborne fraction simulated by cmip5 earth system models under four representative concentration pathways. *Journal of Climate*, **26 (13)**, 4398–4413.

Jones, E., T. Oliphant, P. Peterson, et al., 2001: SciPy: Open source scientific tools for Python. URL `http://www.scipy.org/`, [Online; accessed 2015-12-04].

Joyner, D., O. Čertík, A. Meurer, and B. E. Granger, 2012: Open source computer algebra systems: Sympy. *ACM Commun. Comput. Algebra*, **45 (3/4)**, 225–234, doi:10.1145/2110170.2110185.

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

Kalnay, E., 2003: *Atmospheric modeling, data assimilation, and predictability*. Cambridge university press.

Kaminski, T., et al., 2013: The BETHY/JSBACH Carbon Cycle Data Assimilation System: Experiences and challenges. *Journal of Geophysical Research: Biogeosciences*, **118 (4)**, 1414–1426, doi: 10.1002/jgrg.20118.

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

Keenan, T. F., E. Davidson, A. M. Moffat, W. Munger, and A. D. Richardson, 2012: 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, doi:10.1111/j.1365-2486.2012.02684.x.

Keenan, T. F., E. A. Davidson, J. W. Munger, and A. D. Richardson, 2013: 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, doi:10.1890/12-0747.1.

Kerr, G. and J. Haufe, 2011: Thinning practice: A silvicultural guide. *Forestry Commission*, 54.

Kimmins, J., 1973: Some statistical aspects of sampling throughfall precipitation in nutrient cycling studies in british columbian coastal forests. *Ecology*, 1008–1019.

Knorr, W. and M. Heimann, 2001: 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.

Krinner, G., et al., 2005: A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system. *Global Biogeochemical Cycles*, **19 (1)**, 1–33, doi:10.1029/2003GB002199.

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

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

Lasslop, G., M. Reichstein, J. Kattge, and D. Papale, 2008: Influences of observation errors in eddy flux data on inverse model parameter estimation. *Biogeosciences Discussions*, **5 (1)**, 751–785, URL `https://hal.archives-ouvertes.fr/hal-00297973`.

Law, B., P. Thornton, J. Irvine, P. Anthoni, and S. Van Tuyl, 2001a: Carbon storage and fluxes in ponderosa pine forests at different developmental stages. *Global Change Biology*, **7 (7)**, 755–777.

Law, B., S. Van Tuyl, A. Cescatti, and D. Baldocchi, 2001b: Estimation of leaf area index in open-canopy ponderosa pine forests at different successional stages and management regimes in oregon. *Agricultural and Forest Meteorology*, **108 (1)**, 1–14.

Lawless, A. S., 2013: Variational data assimilation for very large environmental problems. *Large scale Inverse Problems: Computational Methods and Applications in the Earth Sciences, Radon series on Computational and Applied Mathematics*, M. J. P. Cullen, M. A. Freitag, S. Kindermann, and R. Scheichl, Eds., De Gruyter, 55–90.

Le Quéré, C., et al., 2015: Global carbon budget 2015. *Earth System Science Data*, **7 (2)**, 349–396.

Le Toan, T., et al., 2011: The biomass mission: Mapping global forest biomass to better understand the terrestrial carbon cycle. *Remote sensing of environment*, **115 (11)**, 2850–2860.

Lefsky, M. A., D. Harding, W. Cohen, G. Parker, and H. Shugart, 1999: Surface lidar remote sensing of basal area and biomass in deciduous forests of eastern maryland, {USA}. *Remote Sensing of Environment*, **67 (1)**, 83 – 98, doi:http://dx.doi.org/10.1016/S0034-4257(98)00071-6, URL http://www.sciencedirect.com/science/article/pii/S0034425798000716.

Li, Y., I. M. Navon, W. Yang, X. Zou, J. R. Bates, S. Moorthi, and R. W. Higgins, 1994: Four-Dimensional Variational Data Assimilation Experiments with a Multilevel Semi-Lagrangian Semi-Implicit General Circulation Model. *Monthly Weather Review*, **122 (5)**, 966–983, doi:10.1175/1520-0493(1994)122¡0966:FDVDAE¿2.0.CO;2, URL http://journals.ametsoc.org/doi/abs/10.1175/1520-0493(1994)122%3C0966:FDVDAE%

LI-COR, Inc., 2015: *EddyPro 6 Help and User's Guide*. LI-COR, Inc. Lincoln, NE.

Ljung, L., 1998: *System Identification: Theory for the User*. Pearson Education.

Lorenc, A. C. and F. Rawlins, 2005: Why does 4d-var beat 3d-var? *Quarterly Journal of the Royal Meteorological Society*, **131 (613)**, 3247–3257.

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

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

Lüthi, D., et al., 2008: High-resolution carbon dioxide concentration record 650,000–800,000 years before present. *Nature*, **453 (7193)**, 379–382.

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

McKay, H., J. Hudson, and R. Hudson, 2003: Woodfuel resource in britain. *Forestry Commission Report*.

Meir, P., B. Kruijt, M. Broadmeadow, E. Barbosa, O. Kull, F. Carswell, A. Nobre, and P. Jarvis, 2002: Acclimation of photosynthetic capacity to irradiance in tree canopies in relation to leaf nitrogen concentration and leaf mass per unit area. *Plant, Cell & Environment*, **25 (3)**, 343–357.

Mercado, L. M., N. Bellouin, S. Sitch, O. Boucher, C. Huntingford, M. Wild, and P. M. Cox, 2009: Impact of changes in diffuse radiation on the global land carbon sink. *Nature*, **458 (7241)**, 1014–1017.

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

Mitchell, J. F., 1989: The greenhouse effect and climate change. *Reviews of Geophysics*, **27 (1)**, 115–139.

Moodey, A. J. F., A. S. Lawless, R. W. E. Potthast, and P. J. van Leeuwen, 2013: Nonlinear error dynamics for cycled data assimilation methods. *Inverse Problems*, **29 (2)**, 025 002, URL http://stacks.iop.org/0266-5611/29/i=2/a=025002.

Moore, D. J., J. Hu, W. J. Sacks, D. S. Schimel, and R. K. Monson, 2008: 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, doi:http://dx.doi.org/10.1016/j.agrformet.2008.04.013.

Moss, R. H., et al., 2010: The next generation of scenarios for climate change research and assessment. *Nature*, **463 (7282)**, 747–756.

Myneni, R., et al., 2002: Global products of vegetation leaf area and fraction absorbed par from year one of modis data. *Remote sensing of environment*, **83 (1)**, 214–231.

Navon, I., 1998: Practical and theoretical aspects of adjoint parameter estimation and identifiability in meteorology and oceanography. *Dynamics of Atmospheres and Oceans*, **27 (1)**, 55–79.

Navon, I. M., X. Zou, J. Derber, and J. Sela, 1992: Variational Data Assimilation with an Adiabatic Version of the NMC Spectral Model. *Monthly Weather Review*, **120 (7)**, 1433–1446, doi:10.1175/1520-0493(1992)120¡1433:VDAWAA¿2.0.CO;2, URL http://journals.ametsoc.org/doi/abs/10.1175/1520-0493(1992)120%3C1433%3AVDAWA

Niu, S., Y. Luo, M. C. Dietze, T. F. Keenan, Z. Shi, J. Li, and F. S. C. Iii, 2014: The role of data assimilation in predictive ecology. *Ecosphere*, **5 (5)**, art65, doi:10.1890/ES13-00273.1, URL http://www.esajournals.org/doi/abs/10.1890/ES13-00273.1.

Nocedal, J. and S. J. Wright, 1999: *Numerical Optimization*. Springer Science & Business Media, 198 pp., URL http://books.google.co.uk/books/about/Numerical_Optimization.html?id=epc5fX0l

Norman, J. and P. Jarvis, 1975: Photosynthesis in sitka spruce (picea sitchensis (bong.) carr.): V. radiation penetration theory and a test case. *Journal of Applied Ecology*, 839–878.

Oak Ridge National Laboratory Distributed Active Archive Center ORNL DAAC, 2013: Fluxnet maps & graphics web page. URL http://fluxnet.ornl.gov/maps-graphics, [USA Accessed November 5, 2013].

Paige, C., 1981: Properties of numerical algorithms related to computing controllability. *IEEE Transactions on Automatic Control*, **26 (1)**, 130–138, doi:10.1109/TAC.1981.1102563.

Palmer, T., R. Gelaro, J. Barkmeijer, and R. Buizza, 1998: Singular vectors, metrics, and adaptive observations. *Journal of the Atmospheric Sciences*, **55 (4)**, 633–653.

Pan, Y., et al., 2011: A large and persistent carbon sink in the world?s forests. *Science*, **333 (6045)**, 988–993.

Papale, D., et al., 2006a: Towards a standardized processing of net ecosystem exchange measured with eddy covariance technique: algorithms and uncertainty estimation. *Biogeosciences*, **3 (4)**, 571–583.

Papale, D., et al., 2006b: Towards a standardized processing of Net Ecosystem Exchange measured with eddy covariance technique: algorithms and uncertainty estimation. *Biogeosciences*, **3 (4)**, 571–583, doi:10.5194/bg-3-571-2006.

Pitman, R. and M. Broadmeadow, 2001: Leaf area, biomass and physiological parameterisation of ground vegetation of lowland oak woodland. *Forestry Commission, Edinburgh*.

Press, W., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, 2007: *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press.

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

Rabier, F., N. Fourrié, D. Chafäi, and P. Prunet, 2002: Channel selection methods for infrared atmospheric sounding interferometer radiances. *Quarterly Journal of the Royal Meteorological Society*, **128 (581)**, 1011–1027.

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

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

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

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

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

Rich, P. M., J. Wood, D. Vieglais, K. Burek, and N. Webb, 1999: Hemiview user manual, version 2.1. *Delta-T Devices Ltd., Cambridge, UK*, **79**.

Richardson, A. D., et al., 2008: Statistical properties of random {CO2} flux measurement uncertainty inferred from model residuals. *Agricultural and Forest Meteorology*, **148 (1)**, 38 – 50, doi: http://dx.doi.org/10.1016/j.agrformet.2007.09.001.

Richardson, A. D., et al., 2010: Estimating parameters of a forest ecosystem c model with measurements of stocks and fluxes as joint constraints. *Oecologia*, **164 (1)**, 25–40.

Rodgers, C. D., 1998: Information content and optimisation of high spectral resolution remote measurements. *Advances in Space Research*, **21 (3)**, 361–367.

Rodgers, C. D. et al., 2000: *Inverse methods for atmospheric sounding: Theory and practice*, Vol. 2. World scientific Singapore.

Running, S. W., R. R. Nemani, F. A. Heinsch, M. Zhao, M. Reeves, and H. Hashimoto, 2004: A continuous satellite-derived measure of global terrestrial primary production. *Bioscience*, **54 (6)**, 547–560.

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

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

Schimel, D., 2013: *Climate and ecosystems*. Princeton University Press.

Schlich, S. W. and W. Perrée, 1905: *Working plan for the Alice Holt Forest*. HM Stationery Office.

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

Scholze, M., W. Knorr, N. W. Arnell, and I. C. Prentice, 2006: A climate-change risk analysis for world ecosystems. *Proceedings of the National Academy of Sciences*, **103 (35)**, 13 116–13 120.

Schwalm, C. R., et al., 2010: A model-data intercomparison of co2 exchange across north america: Results from the north american carbon program site synthesis. *Journal of Geophysical Research: Biogeosciences*, **115 (G3)**.

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

Singh, K., A. Sandu, M. Jardak, K. Bowman, and M. Lee, 2013: A practical method to estimate information content in the context of 4d-var data assimilation. *SIAM/ASA Journal on Uncertainty Quantification*, **1 (1)**, 106–138.

Sitch, S., et al., 2008: Evaluation of the terrestrial carbon cycle, future plant geography and climate-carbon cycle feedbacks using five dynamic global vegetation models (dgvms). *Global Change Biology*, **14 (9)**, 2015–2039.

Smith, P. J., S. L. Dance, M. J. Baines, N. K. Nichols, and T. R. Scott, 2009: Variational data assimilation for parameter estimation: application to a simple morphodynamic model. *Ocean Dynamics*, **59 (5)**, 697–708.

Smith, P. J., S. L. Dance, and N. K. Nichols, 2011: A hybrid data assimilation scheme for model parameter estimation: Application to morphodynamic modelling. *Computers & Fluids*, **46 (1)**, 436 – 441, doi:http://dx.doi.org/10.1016/j.compfluid.2011.01.010, 10th {ICFD} Conference Series on Numerical Methods for Fluid Dynamics (ICFD 2010).

Stewart, L. M., S. Dance, and N. Nichols, 2008: Correlated observation errors in data assimilation. *International journal for numerical methods in fluids*, **56 (8)**, 1521–1527.

Stewart, L. M., S. L. Dance, and N. K. Nichols, 2013: Data assimilation with correlated observation errors: Experiments with a 1-D shallow water model. *Tellus, Series A: Dynamic Meteorology and Oceanography*, **65 (1)**, 1–14, doi:10.3402/tellusa.v65i0.19546.

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

Taylor, K. E., 2001: Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research*, **106 (D7)**, 7183, doi:10.1029/2000JD900719, URL http://doi.wiley.com/10.1029/2000JD900719.

Tremolet, Y., 2006: Accounting for an imperfect model in 4D-Var. *Quarterly Journal of the Royal Meteorological Society*, **132 (621)**, 2483–2504, doi:10.1256/qj.05.224, URL http://doi.wiley.com/10.1256/qj.05.224.

Trudinger, C. M., et al., 2007: Optic project: An intercomparison of optimization techniques for parameter estimation in terrestrial biogeochemical models. *Journal of Geophysical Research: Biogeosciences*, **112 (G2)**.

Verbeeck, H., P. Peylin, C. Bacour, D. Bonal, K. Steppe, and P. Ciais, 2011: Seasonal patterns of CO2 fluxes in Amazon forests: Fusion of eddy covariance data and the ORCHIDEE model. *Journal of Geophysical Research*, **116 (G2)**, 1–19, doi:10.1029/2010JG001544.

Wahba, G., 1985: Design criteria and eigensequence plots for satellite-computed tomography. *Journal of Atmospheric and Oceanic Technology*, **2 (2)**, 125–132.

Waller, J. A., S. L. Dance, A. S. Lawless, N. K. Nichols, and J. R. Eyre, 2014: Representativity error for temperature and humidity using the Met Office high-resolution model. *Quarterly Journal of the Royal Meteorological Society*, **140 (681)**, 1189–1197, doi:10.1002/qj.2207, URL http://doi.wiley.com/10.1002/qj.2207.

Walter, S. F. and L. Lehmann, 2013: Algorithmic differentiation in Python with AlgoPy. *Journal of Computational Science*, **4 (5)**, 334–344, doi:10.1016/j.jocs.2011.10.007, URL http://www.sciencedirect.com/science/article/pii/S1877750311001013.

Weston, P., W. Bell, and J. Eyre, 2014: Accounting for correlated error in the assimilation of high-resolution sounder data. *Quarterly Journal of the Royal Meteorological Society*, **140 (685)**, 2420–2429.

Wilkinson, M., P. Crow, E. Eaton, and J. Morison, 2015: Effects of management thinning on co_2 exchange by a plantation oak woodland in south-eastern england. *Biogeosciences Discussions*, **12 (19)**.

Wilkinson, M., E. Eaton, M. Broadmeadow, and J. Morison, 2012: Inter-annual variation of carbon uptake by a plantation oak woodland in south-eastern england. *Biogeosciences*, **9 (12)**, 5373–5389.

Williams, M., E. B. Rastetter, D. N. Fernandes, M. L. Goulden, G. R. Shaver, and L. C. Johnson, 1997: Predicting gross primary productivity in terrestrial ecosystems. *Ecological Applications*, **7 (3)**, 882–894.

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

Williams, M., et al., 2009: Improving land surface models with fluxnet data. *Biogeosciences*, **6 (7)**, 1341–1359.

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

Yuan, W., 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 (3)**, 189–207.

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

Zobitz, J., D. J. Moore, T. Quaife, B. H. Braswell, A. Bergeson, J. A. Anthony, and R. K. Monson, 2014a: 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.

Zobitz, J. M., D. J. P. Moore, T. Quaife, B. H. Braswell, A. Bergeson, J. a. Anthony, and R. K. Monson, 2014b: 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, **195-196**, 73–88, doi:10.1016/j.agrformet.2014.04.011, URL http://dx.doi.org/10.1016/j.agrformet.2014.04.011.

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

Zou, X., I. Navon, and F. Le Dimet, 1992: Incomplete observations and control of gravity waves in variational data assimilation. *Tellus A*, **44 (4)**, 273–296.