The Limitations and Robustness of Data Assimilation in Terrestrial Ecosystem Modelling

Edmund Ryan
Department of Applied Mathematics,
University of Sheffield

PhD by Research

July, 2013

Declaration

The content of this thesis is completely my own work, including all Matlab scripts, and it is written in my own words. This work was carried out under the supervision of Professor Shaun Quegan at the University of Sheffield.

Chapter four was accepted as a paper in July, 2011: Hill, T.C., E. Ryan and M. Williams (2012) *The use of CO_2 flux time series for parameter and carbon stock estimation in carbon cycle research.* **Global Change Biology** 18: 179-193. Equal time and effort between myself and T.C. Hill was spent on the design of the study. Hill carried out the computer runs for the Ensemble Kalman Filter, and I did those for the Metropolis algorithm. While Hill wrote the paper, I was heavily involved in checking through it and giving feedback. In addition, chapter four is written completely my own words, and the four other bits of extension work (the sensitivity analysis using GEM-SA, testing the forecasts under drought conditions, the lag plots and determining whether the true parameter set is at the global minimum of the mismatch between observed and modelled Net Ecosystem Exchange for the Metropolis algorithm) did not feature in the paper. Williams provided very useful feedback for the design of the study, in the analysis of the results, and in checking over the several versions of the final manuscript.

Acknowledgments

I would like to thank my supervisor, Professor Shaun Quegan, from whom I have learnt an enormous amount during the past few years due to his invaluable guidance, knowledge, and insight. I also appreciate greatly the time and effort he spent giving feedback on the thesis. I would also like to thank the others I have worked alongside during my PhD, in particular Professor Mat Willliams at the University of Edinburgh, Dr Tim Hill and the University of St Andrews, and Dr Tristan Quaife and the University of Reading. I have learnt a huge amount from all of these fantastic academics. I owe a huge amount to my wife Jacqueline, for her generosity when I needed to work late, for her encouragement particularly during the difficult times, and for her continued love. Finally, I would like to thank my family and friends for their support and love.

Contents

Abstract

Accurately estimating how much carbon is leaving the atmosphere and being taken up by plants, by processes such as photosynthesis, is critical in order to make accurate climate forecasts. There is a large uncertainty of this atmosphere-plant carbon flux, sometimes referred to as Net Ecosystem Exchange (NEE), therefore reducing this is essential. One way of doing this is through Data Assimilation (DA), the framework by which data and models are combined together in a statistically optimal way. A key aspect of DA is that the uncertainty on the estimate of interest, e.g. NEE, is less than either the uncertainty on using either the model or observations on their own. DA can also been used to estimate model parameters, which have traditionally been estimated from expert knowledge or from small scale studies. While DA has gained much interest as a powerful tool in estimating model parameters and quantities such as NEE, there are a number of issues surrounding its use, which are not yet properly understood. These include:

- When using DA to estimate parameters using ground observations:
    a. Understanding the limitations of DA and the conditions it performs best.
    b. Determining likely factors that cause variations in parameter estimates.
- Assessing the impact of assimilating satellite observations of leaf area index to improve the model states, and whether DA is robust against unrealistic features of the satellite data.

The aim of this PhD was to address and learn more about these issues. This was done by using the evergreen and deciduous versions of the Data Assimilation Linked ECosystem (DALEC) model. The main findings are summarised in the following four paragraphs:

The Ensemble Kalman Filter (EnKF) is good at estimating parameters using synthetic NEE data. It was found that between 2 and 5 years of this data was required in order for the parameters and NEE forecasts to be close to the truth. There was for the most part very little difference to the EnKF parameter estimates and NEE forecasts whether very noisy or very non-noisy observations were used, or whether 20% or 100% of the daily observations were present in the dataset.

For the Metropolis algorithm, most of the runs had to be discarded as it was found that the the algorithm was not converging for the global minimum for these runs; this

caused some other problems, in particular residuals between the modelled and observed NEE were autocorrelated. For these discarded runs, the parameter estimates tended to be far from the truth and the 90% posterior intervals rarely included the truth. For the remaining runs, where the converse of the above was found to be true, as dataset length increased from 1 to 5 years, the posterior parameter distribution coincided with the truth to a greater extent.

Using the Metropolis algorithm and assimilating three years daily NEE observations (with around 60% data coverage) and around 10 LAI observations during this period, it was found that parameter estimates were sensitive to the initial value of the labile carbon store. Moreover, the parameters were close to their true values if the true initial value of the labile C pool was used. It was also found that when these initial conditions were treated as parameters, although the modal value of the corresponding marginal posterior distributions were far from the truth, every other aspect of the model (parameters and trajectories of the model states) agreed well with the truth. This supported the common approach by many of the DA community that treating initial conditions as parameters is preferable than keeping them fixed (using site inventory data or from model spin-up). The novelty of this part of the thesis was for the first time an emulator was applied to a DA scheme.

Finally, the EnKF was used to estimate the LAI and NEE model states, using a fixed parameter set and LAI data from the MODIS satellite sensor. It was found that processing the MODIS LAI in order to correct for unrealistic features of the dataset, such as excessive temporal variation and very small uncertainties, improved the fit of the modelled to observed NEE after assimilation. The improvement in the fit was significantly better for Gross Primary Production (GPP).

Chapter 1

Motivation and aims of thesis: modelling and observing the terrestrial carbon cycle, and the need for Data Assimilation

1.1 The role of the terrestrial biosphere in the C cycle

Over the past few decades, more and more evidence has been built up to show that climate change is happening. Such evidence can be found in the various IPCC (Inter-Governmental Panel on Climate Change) documents that have been produced since 1990 (IPCC, 2007). The purpose of the IPCC is to assess whether climate change is happening, whether it is anthropogenic, and advising on mitigation and adaptation. It brings together climate change related research from varying academic disciplines and over a wide range of countries.

Joseph Fourier is regarded by some as the father of climate change research, since he was the first to hypothesise a concept called the "Greenhouse effect" - see Burgess (1837). The greenhouse effect is the effect of atmospheric gases referred as greenhouse gases (GHGs) in raising the surface temperature of the earth. To understand this we first describe the global energy balance. The description that follows is taken from Soloman et al. (2007). Of the incoming shortwave solar radiation that reaches the earth, about a half is either reflected back into space of absorbed by the atmosphere. The remainder is

absorbed by the surface, but is then transferred back to the atmosphere by thermals (the heating of the air close to the earth's surface), evapotranspiration and surface radiation (figure 1.1). Certain wavelengths of the outgoing longwave radiation emitted from the surface are absorbed by GHGs, such as CO_2, methane, and water vapour. This heating of the atmosphere caused by this absorption results in the atmosphere emitting more radiation, some of which heads towards the earth surface increasing its temperature. The more GHGs are present in the atmosphere, the higher the surface temperature becomes. The IPCC report (IPCC, 2007) states:

> *'Without the natural greenhouse effect, the average temperature at the Earth's surface would be below the freezing point of water. Thus the natural greenhouse effect makes life as we know it possible. However, human activities, primarily the burning of fossil fuels and the clearing of forests, have greatly intensified the natural greenhouse effect, causing global warming.'*

The IPCC report (IPCC, 2007) refers to carbon dioxide as 'the most important anthropogenic (i.e. caused by human activity) greenhouse gas'. The report states that atmospheric CO_2 levels have increased from the pre-industrial value of 280 ppm^3 (parts per million per unit volume of air) to 379 ppm^3 (in 2005). Physical evidence for this came from Petit et al. (1999) who used ice core data from Vostok in Antarctica to estimate atmospheric CO_2 over the past 400,000 years (past 10,000 years shown in the left panel of figure 1.2). The results showed that prior to 1800 atmospheric CO_2 levels were never above 300 ppm^3. However after 1800, there has been a steady and continuous increase in CO_2 with levels never dropping below 300 ppm^3. The sudden change in atmospheric CO_2 levels at around 1800 is observed by Etheridge (1998) who estimated CO_2 levels from 1006 to 1978, and Neftel (1994) who estimated atmospheric CO_2 levels from 1734 to 1983, with both studies using ice-core data.

Increases in CO_2 over the past 50 years have been backed up work by Charles Keeling (Keeling, 1960) who was the first person to make frequent and continuous measurements of CO_2. In 1958 he began taking atmospheric measurements of CO_2 conc-

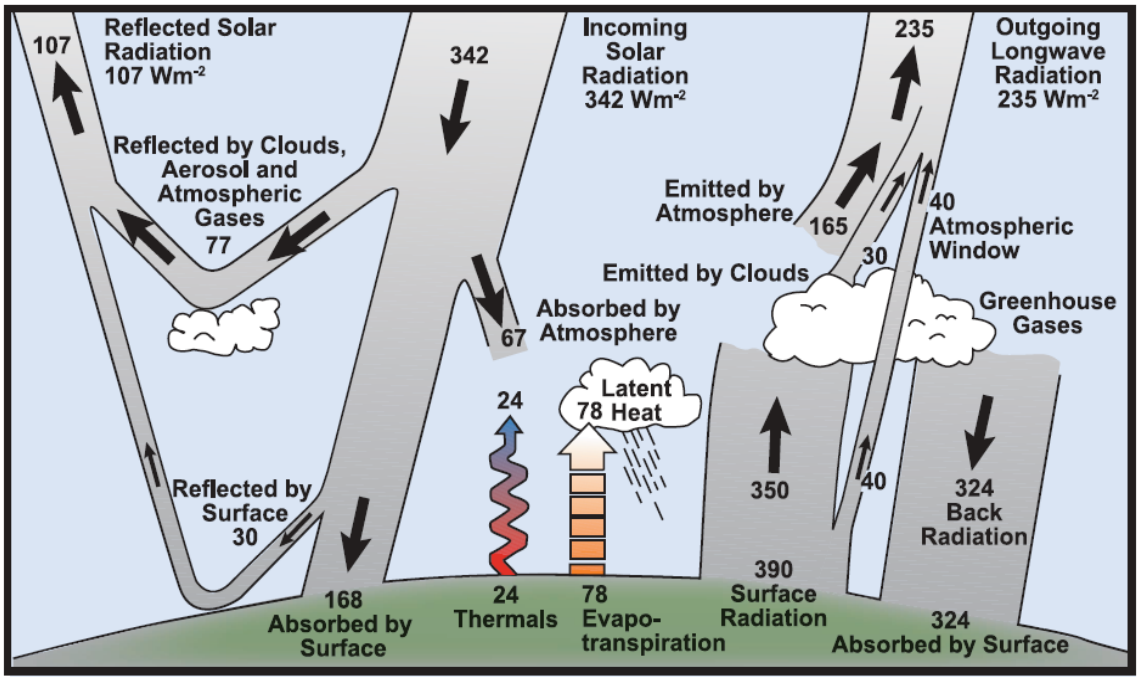


Figure 1.1 Estimate of the Earth's annual and global mean energy balance. Source: Solomon et al. (2007).

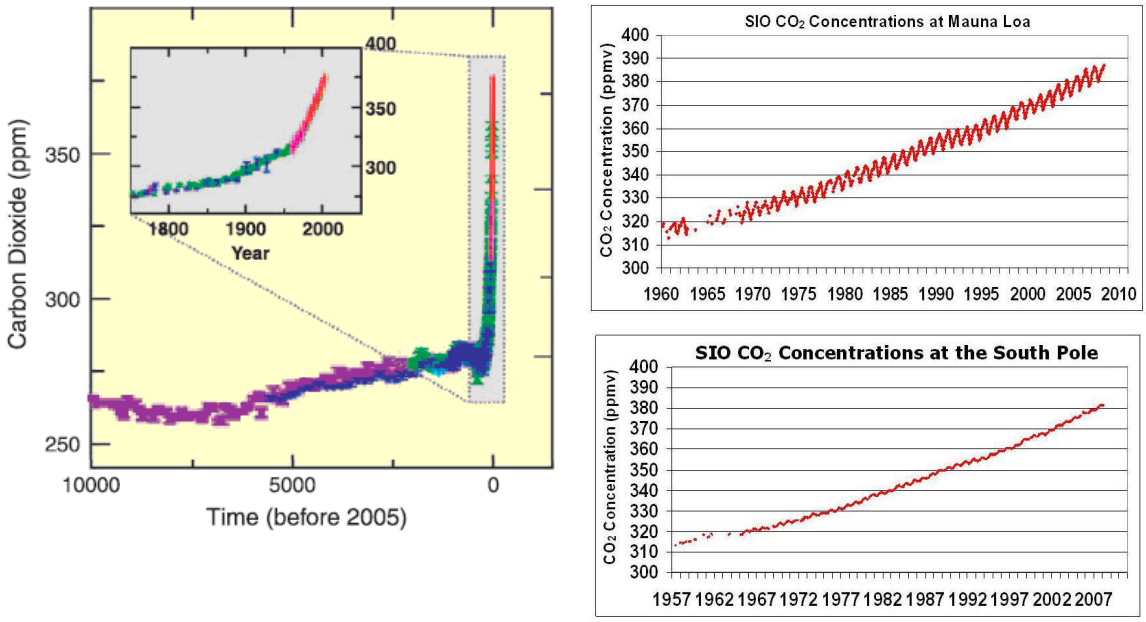


Figure 1.2 Left panel: Global average CO_2 levels for the past 10,000 years (IPCC, 2007). Top right panel: Atmospheric CO_2 levels at Mauna Loa in Hawaii from 1958-1974 by Charles Keeling and 1974-2007 by the National Oceanic Atmospheric Administration (Keeling et al., 2008); Bottom right panel: Atmospheric CO_2 levels at the South Pole for 1957 – 2007 (Keeling et al., 2008).

entrations in Mauna Loa, Hawaii, and these have continued to be made to the present day. The graphical representation of this data is known as the Keeling curve (figure 1.2, top right panel). A feature of the curve is the seasonal oscillations. Each annual segment peaks in the winter because there is more CO_2 in the atmosphere as a result of there being less photosynthetic activity, and conversely the troughs occur during the summer when photosynthesis is at its highest. If we were to make atmospheric CO_2 measurements in the Antarctic (figure 1.2, bottom right panel), we would find that this seasonal variation is very minimal because of the lack of vegetation (Keeling et al., 2008).

For parts of the globe which experience cooler annual average temperatures, an increase of a few degrees may be nothing bad. However, there are worrying consequences for the planet. One big consequence is the rising of the sea level caused by the melting of glaciers and ice-caps and the fact that water expands when it is heated. The consequences of these are catastrophic – countries such as the Netherlands and Bangladesh which are low-lying could become inhabitable and flooding is likely to increase in other places. There are numerous other consequences of rising CO_2 levels and rising temperatures, which include: (i) the increased acidification of the oceans (due to greater uptake of CO_2 from the atmosphere) and a decrease in the salinity of the oceans, which are having a major impact on the behaviour of many plant and animal organisms; (ii) the melting of permafrost in the boreal regions.

In order to better predict future atmospheric CO_2 levels and its implications not only on how surface temperatures will rise but also on how the climate will change and the effects on all plant and animal life, we need to gain a better understanding of the current climate. In particular better knowledge of the Carbon cycle is required, in order to determine the amount of carbon leaving and entering the atmosphere. Put simply, the Carbon cycle describes the movement of carbon between the atmosphere, the biosphere, the geosphere and the oceans. The rate at which carbon moves from one of the four earth compartments and within each one varies from seconds to 1000s of years (Post et al., 1990). Figure 1.3 displays a simplified diagram of how it operates. Carbon is transferred between the atmosphere, biosphere, etc… in a number of different ways.

The main natural fluxes of carbon (i.e. those not induced by human activity), are shown in black in figure 1.3, are:

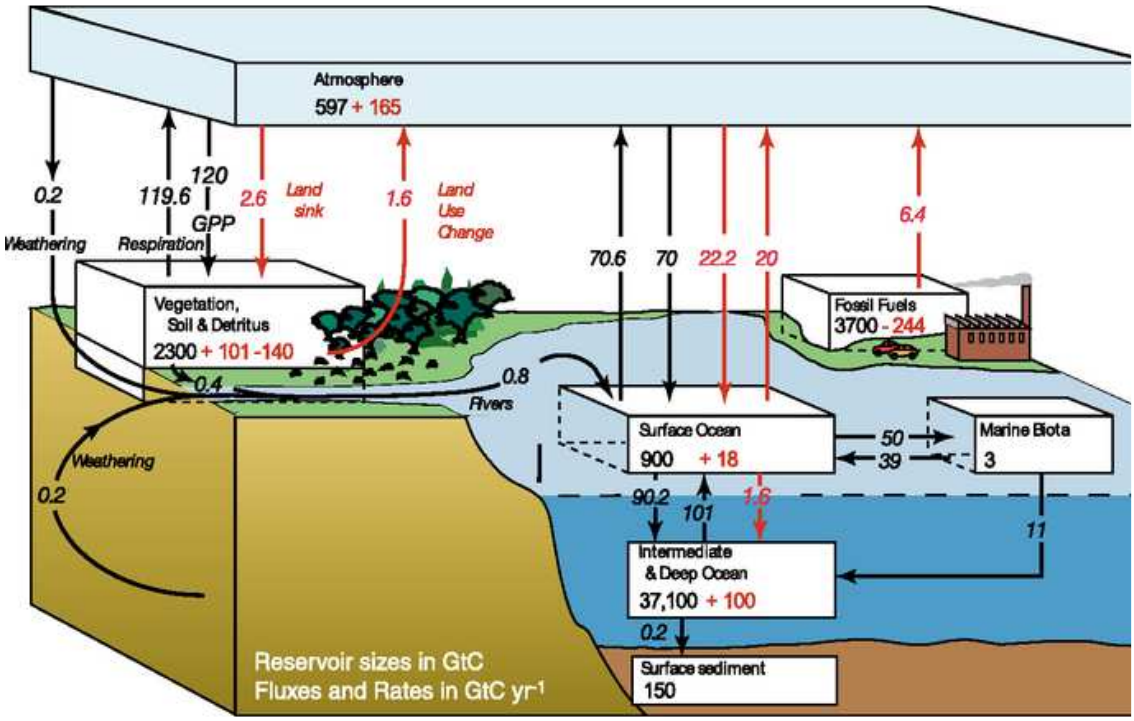


Figure 1.3: The global carbon cycle for the 1990s, which shows the pre-industrial 'natural' fluxes (black) and 'anthropogenic' fluxes (red), measured in GtC/yr. This figure is taken from IPCC (2007), with original sources being Sarmiento & Gruber (2006) and Sabine et al. (2004b)).

- **GPP**: The removal of CO_2 in the atmosphere by plants via photosynthesis, labelled as GPP (Gross Primary Production) in figure 1.3. On the terrestrial biosphere, photosynthesis is a mechanism used by plants to convert solar radiation into chemical energy or sugars that provides fuel for the plant to grow, requiring CO_2 and H_20, and excreting O_2. (Blankenship, 2002). Larcher (2003) states 'for every gram atomic weight of carbon taken up [by the plant via photosynthesis], 479 kJ of potential energy are obtained.' This is equivalent to approximately 113 kCal, which is around 5.7% of the 2000 kCal recommended daily energy required for a man. CO_2 and O_2 are exchanged between the outside air and the chloroplasts inside the plant predominantly by diffusion (Niinemets et al., 2005).
- **Respiration:** The respiring of carbon by plants and soils back to the atmosphere, denoted by the upward facing 'respiration' black arrow in figure 1.3. Carbon can be respired back the atmosphere from the biosphere in a number of different ways.

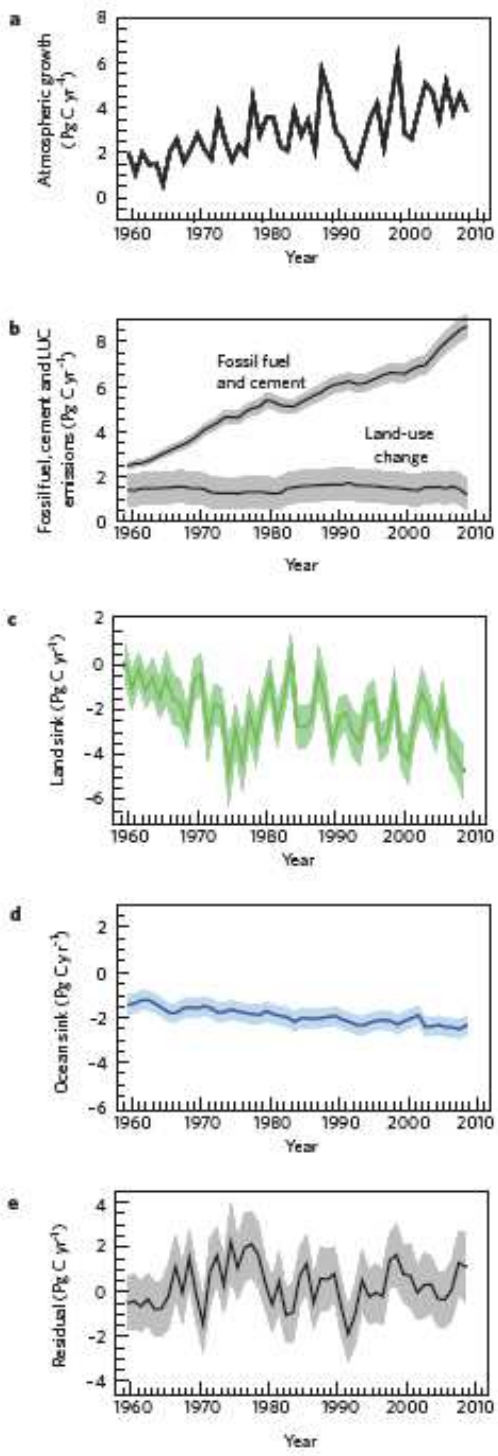
First of all, the sugars produced from photosynthesis need to be broken down metabolically in order to produce the energy required for plant maintenance, growth and reproduction (Larcher, 2003). Due to this using up of the sugars and using up oxygen and the production of CO_2, respiration is considered the opposite of photosynthesis (Blankenship, 2002). Respiration can also occur in the soil for the same reasons, but as a result of root growth and waste produced by microbes which feed on foliar or root litter and soil organic matter (Fang & Moncrieff, 1999)

- **Fires:** Although not shown in figure 1.3, vegetation fires can result in a significant amount of CO_2 to the atmosphere from year to year, but most of the CO_2 is taken up by the terrestrial biosphere over the course of a decade as the vegetation grows back (Bowman et al., 2009).
- **Photosynthesis in the oceans:** Phytoplankton in the oceans take up CO_2 in the atmosphere through photosynthesis, and this is represented by the downward black arrow by the ocean part of figure 1.3. Phytoplankton are organisms that are microscopic in size and because they obtain their energy through photosynthesis, they must live close to the surface of the ocean where it is well lit (Behrenfeld, 2009).
- **Respiration in the oceans:** The respiring of most of the carbon at the surface occurs through the decaying of organic matter by the presence of bacteria (Rivkin & Legendre, 2001). This is represented by the upward facing arrow between the surface ocean box and the atmosphere.
- **The biological pump:** Represented by the downward and upward black arrows in figure 1.3 between the 'surface ocean' box and the 'Intermediate and Deep Ocean' box, some of the photosynthesized carbon descends to intermediate and deep layers as particles. Most of the organic carbon is transported to the deeper part of the ocean by sinking particulate material, e.g. dead organisms (Post et al., 1990). The transport of carbon from the lower depths to the surface is achieved by a process known as 'upwelling'. This transport of carbon between the surface and lower depths of the ocean is commonly referred as the 'biological pump' (Post et al., 1990).

Perturbations in the natural carbon cycle occur as a result of human activity. These anthropogenic fluxes of carbon, represented by the red arrows in figure 1.3, are:

- CO_2 released back into the atmosphere predominantly by the burning of fossil fuels, such as oil and coal. This is represented by the red arrow on the furthest right of figure 1.3.
- Represented by the red arrow in figure 1.3 labelled 'land-use change', CO_2 is also released back to the atmosphere due to land-use change which is mainly deforestation. This occurs because deforestation results in fewer trees which means less CO_2 is being taken up by the land via photosynthesis (Van der Werf et al., 2009). Although to a lesser extent, CO_2 can also be released back to the atmosphere by the burning of forests initiated anthropogenically.
- Extra CO_2 taken up plants by photosynthesis due to the extra CO_2 in the atmosphere some of which, but not all, is respired by the plants back into the atmosphere (Sabine et al., 2004b). This results in a net flux of CO_2 from the atmosphere to the geosphere, and this is denoted by the 'land sink' arrow in figure 1.3.
- Extra CO_2 taken up the phytoplankton in the oceans by photosynthesis, again due to the extra CO_2 in the atmosphere (Sabine et al., 2004a). Much of this, but not all, is also respired back to atmosphere and is represented by the two red arrows above the ocean part of figure 1.3.

Figure 1.4 shows the global estimates of how carbon is entering the atmosphere (the sources) and how it is staying in or leaving the atmosphere (the sinks) from 1960 to 2008 (Le Quere et al., 2009). Panel (a) shows the increase in atmospheric CO_2, and despite the very large intra-annual variability, there is a positive trend. Canadell et al. (2007) state that the growth rate between 2000 and 2006 was 1.93 ppm yr^{-1} (parts per million per year), compared to the lower rates in the 1980s and 1990s of 1.58 ppm yr^{-1} and 1.49 ppm yr^{-1} respectively. To gain an appreciation of the uncertainty of atmospheric CO_2 measurements, Tans & Conway (2011) estimated the uncertainty of the mean growth rate of measurements made at Mauna Loa as being 0.11 ppm yr^{-1}. In terms of Pg yr^{-1}, Canadell et al. (2007) estimate this to be 0.04 Pg yr^{-1}; this is low because of the high precision of the atmospheric CO_2 measurements and because of the fast mixing time-scale of the atmosphere. Le Quere et al. (2008) describe the increase in atmospheric CO_2 levels in terms of the growth rate of the air-borne fraction of CO_2. They state that between 1959



	GCP (2011) & Le Quere et al. (2009)	**Canadell et al. (2007)**
	2000 - 2008	2000-2006
a	4.1 ± 0.2 PgC yr^{-1}	4.1 ± 0.04 PgC yr^{-1}
	2000 - 2008	2000-2006
b	FF: 7.7 ± 0.5 PgC yr^{-1}	7.6 ± 0.38 PgC yr^{-1}
	1990 - 2005	2000-2006
	LU: 1.4 ± 0.7 PgCyr^{-1}	1.5 ± 0.5 PgC yr^{-1}
	(FF = Fossil Fuel & cement; LU = Land-Use change)	
	2000–2008&1990–1999	2000-2006
c	3.0 ± 0.7 PgC yr^{-1}	2.8 ± 0.7 PgC yr^{-1}
	2000–2008&1990–1999	2000-2006
d	2.3 ± 0.4 PgC yr^{-1}	2.2 ± 0.4 PgC yr^{-1}

Figure 1.4 The five left panels show the Global Carbon flux budget for the 1959-2008. Units are petagrams of Carbon per year. (Le Quere et al., 2009). The boxes on the right show the amounts of C in Petagrams of Carbon per year (Pg C yr^{-1}) for the period 2000 – 2008 (estimates taken from the Global Carbon Project (2011) and uncertainties taken from Le Quere et al., 2009) and 2000 – 2006 (Canadell et al., 2007). For panels (c) and (d), the uncertainties for the 2000 – 2008 period are taken from the 1990-1999 period.

and 2008, 43% on average of the total CO_2 emissions each year has remained in the atmosphere. Furthermore, this fraction increased by 0.3 ± 0.2% yr^{-1} between 1959 and 2008. The large relative uncertainty of 0.2% is due to the large uncertainty in CO_2 emitted to the atmosphere as a result of land use change, estimated to be ± 0.7 Pg C yr^{-1} for 2000 – 2008 (figure 1.4b). Le Quere et al. (2009) state that this was revised up from ± 0.5 Pg C yr^{-1} due to recent discrepancies between conversion statistics used to calculate deforestation and agricultural rates. Presumably, this error was found after the publication of Canadell et al. (2007) as the uncertainty shown here is 0.5 Pg C yr^{-1}.

This source of CO_2 to the atmosphere was estimated to be 1.4 Pg C yr^{-1} for the period 1995 – 2006 (Le Quere et al., 2009) or 1.5 Pg C yr^{-1} during 2000 to 2006 (Canadell et al., 2007). The estimates over the past 50 years indicate that this source has remained fairly constant over this period though Le Quere et al. (2009) and Canadell et al. (2007) do not suggest why. The uncertainty on this land use source is high because we do not have accurate global measurements of rates of deforestation and degradation (Achard et al., 2002). This is mainly due to incomplete and inaccurate data taken from the ground, partly because obtaining reliable estimates from individual countries and / or landowners can be difficult as programmes of deforestation can carry a high economic benefit for that country. There are other sources of CO_2 to the atmosphere due to land use change, and these include logging and cultivation of crop-land soils, but deforestation (which include fire emissions) of tropical forests accounted for the greatest proportion (Le Quere et al., 2009).

Between 1850 and 1960 fossil fuel emissions and cement production steadily increased from close to 0 Pg C yr^{-1} to around 2 Pg C yr^{-1} (Boden et al., 2010). However since 1960, we can see from figure 1.4b that this source of CO_2 to the atmosphere has increased at a much faster rate. It is likely that this has been the major driver behind the atmospheric CO_2 increase of 1.93 ppm yr^{-1} for 2000-2006. As a percentage, this increase is around 0.4%, calculated using ((385–315)÷315)×100, where 315 and 385 are the atmospheric CO_2 concentration in ppm for 1959 and 2008 (Tans & Conway, 2011). What is interesting is that this 0.4% (ignoring the very small uncertainty) is larger than the yearly increase in air-borne fraction of CO_2 of 0.3% from 1959 to 2008 suggesting that plants are taking up more CO_2 from the atmosphere as atmospheric CO_2 increases,

however the large relative uncertainty of the 0.3% (+/-0.2% yr^{-1}) makes this difficult to determine for sure at the moment. Improvements in estimating the contribution of CO_2 from the land use change will reduce this uncertainty. An alternative way of reducing this uncertainty is to directly estimate the take up of CO_2 by the land and oceans. This is exactly what is shown in panels (c) and (d) of figure 1.4.

In both panels (c) and (d), we can see from the general trend of the line plots that between 1959 and 2008, there has been a greater uptake of CO_2 from the atmosphere by the land and ocean. While the rise in the uptake by ocean has been fairly steady, the uptake by land has been very variable over the past 50 years. Le Quere et al. (2009) state that this is mainly driven by variability in precipitation, surface temperature and radiation. The uncertainties in the land and ocean sink were determined from the spread of estimates given by five global vegetation model and four ocean general circulation models. One reason for the smaller uncertainty on the ocean sink estimate is because the ocean is more uniform in terms of CO_2 uptake than the terrestrial biosphere (Gruber et al., 2009). However, the ocean is as a complex system as the terrestrial biosphere and there is some suggestion that the method used to estimate the atmosphere–ocean C fluxes may not be as robust as assumed before (Keeling, 2005); therefore quantifying the uncertainty based on the spread of models is arguably unsatisfactory. Nonetheless the fact that there is a larger spread of estimates of the land sink, compared to the ocean sink, suggests that providing consistent estimates between different global models is more of a challenge.

Panel (e) of figure 1.4 shows the residual of the global carbon budget, since what goes into the atmosphere must equal what comes out. If all of the estimates of the sources and sinks of CO_2 into the atmosphere were correct, the residual would be a flat horizontal line at zero line on the y-axis. The line is far from flat and will only approach this ideal state with improved our estimates of the sources and sinks. The uncertainty on the residual is also large and estimated to be of length 2.1 Pg C yr^{-1}. Le Quere et al. (2009) do not define how the uncertainties were obtained, but presumably they represent the minimum and maximum values of the residuals based on the upper and lower bounds of the sources and sinks.

Correctly forecasting the carbon budget for the future is essential if we are to predict what the earth's future climate will be under anthropogenic influences (such as

continuing to emit CO_2 into the atmosphere from the burning of fossil fuels) and we can best plan and mitigate against these changes. However, in order to correctly forecast the carbon budget we must be able to first correctly estimate it for the current and past time periods. Reducing the absolute value of the residual to be closer to zero and reducing the corresponding uncertainty would be a strong indication that our estimates of the sinks and sources of CO_2 to the atmosphere for the past 50 years would be closer to the correct values. There is scope for improvement of all the sinks and sources, but it is of critical importance to improvement the estimates of those sinks and sources which we are most uncertain about. From figure 1.4, the source or sink with the highest relative uncertainty is the source of CO_2 to the atmosphere due to land use change. For 2000 – 2006, this has an uncertainty of ± 0.7 Pg C yr^{-1} or 50% relative to its estimate of 1.4 Pg C yr^{-1}. As stated previously, this large uncertainty is because we do not have accurate global measurements of rates of deforestation and degradation, mainly due to incomplete and inaccurate data. For 2000 – 2010, the land sink (panel c of figure 1.4) has the second largest relative uncertainty of just over 23% (calculated by dividing 0.7 by 3.0 and multiplying by 100%). The other sources and sinks are less critical for the following reasons:

- The increase in atmospheric CO_2 has a very small relative uncertainty of around 1% (panel (a) of figure 1.4).
- Although not as small, the source of CO_2 to the atmosphere from the burning of fossil fuels and from cement also has a small relative uncertainty of around 6.5% (panel (b) of figure 1.4).
- The ocean sink has a relative uncertainty of 17%, which although is not much less than that of the land sink, the ocean sink has been much less variable over the past 50 years (panels (c) and (d) of figure 1.4) and so predicting what it will be in the future could be argued to be much easier.

As a result of the previous paragraph, the following questions that arise:

(i) How can we reduce the uncertainty on the land use source?

(ii) How is the land taking up more carbon in response to increased atmospheric CO_2 from human activity? In the future will there be a saturation point of this land sink?

(iii) How can we reduce the uncertainty on the land sink estimate?

Since the land use source is predominantly due to deforestation, the main way (i) can be addressed is to get accurate measurements of deforestation rates across the globe. Currently the only way this can feasibly be done is by inferring them by using satellite measurements. This is because, as stated previously, taking measurements from the ground is likely to be unreliable. Also, while using a model to forecast future deforestation rates might be worthwhile, using a model to estimate current and past global deforestation rates requires accurate observations which as stated above are not available to an accurate enough degree on the ground. Measurements of biomass change have been made from space since 2005 using the polarimetric Phased Array L-band Synthetic Aperture Radar (PALSAR) on board the Advanced Land Observing Satellite (ALOS) (Whittle et al., 2012), and are helping to reduce the uncertainties of the land use source. Furthermore, one of the aims of BIOMASS mission (Quegan et al., 2012), due to be launched in the next five years, is to monitor forest disturbance and help to reduce this uncertainty to do the source of C to the atmosphere as a result of land use change.

Answering (ii) and (iii) is more difficult, but just as urgent as (i), because the variation in the land sink over the past 50 years has much greater than that of the land use source (figure 1.4b). Also, unlike deforestation rates, it is almost impossible to measure photosynthesis and plant respiration across the entire globe. This is because if they are measured from the ground they are limited by the spatial and temporal frequency they can be made, and if measured from space they may only have coarse resolution and we are not able to make all types of measurements, e.g. measuring the respiration of carbon from the soil. Therefore, in order to accurate estimate and predict this land sink, we need to use models of the terrestrial carbon cycle as well as observations. Such models describe our best knowledge of the physics of the processes that occur, and have the ability to provide global estimates and fine resolution. However, in order for such models to be of value they need to be able to validated against what we observe through the measurements.

Therefore, this thesis is motivated by how data and models can be used together to accurately estimate and forecast this land carbon sink. Before discussing the types of

models are data that are used, a brief overview is first given of the terrestrial carbon cycle and the main quantities of interest that require estimating.

1.2 Modelling and measuring the terrestrial C cycle

In this section, methods of modelling and measuring the terrestrial C cycle are presented. These are discussed in greater detail in chapter two but given that data assimilation, the merging of data and models, is at the heart of this thesis it is relevant and important to discuss some of the key aspects of modelling and measuring the terrestrial C cycle. The section begins with an overview of the terrestrial C cycle and the main quantities that are of interest in estimating. It then moves onto how measurements are made and then to how we might go about modelling the terrestrial C cycle. This then leads into section 1.3, which introduces the argument for the need for data assimilation and the basic principles behind it.

1.2.1 The terrestrial C cycle

The terrestrial carbon cycle is the part of the carbon which only involves exchanges of carbon between the atmosphere and geosphere. We now summarise the main quantities of interest in this exchange, based on Steffen et al. (1998), before giving a more full description of these quantities including their sizes, uncertainties, and temporal and spatial variability. Of the total amount of carbon taken up by the plant via photosynthesis (referred to as Gross Primary Production or GPP), around half is respired autotrophically from the plant due to plant maintenance and growth. What remains is referred to as Net Primary Production (NPP). Some of this is shed as litter entering the soil as foliar litter or already exists in the soil as root litter (i.e. dead roots). This litter decomposes in the soil, releasing nutrients to the soil and CO_2 to the atmosphere. The remaining carbon left in the geosphere is referred to as Net Ecosystem Production or NEP. Much of this is lost due to fire, insect damage, harvesting and deforestation. What remains is then called Net Biome Production (NBP), a tiny fraction of GPP, and this represents the long-term (decadal) store of carbon. This process is depicted graphically in figure 1.5. In this figure,

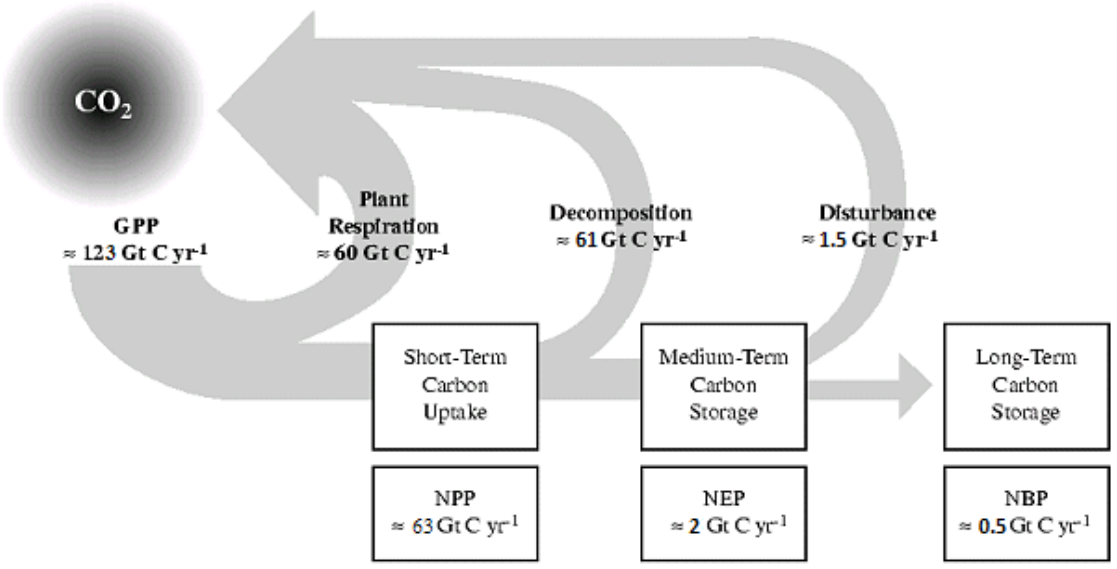


Figure 1.5 Global terrestrial carbon uptake. Plants respire CO_2 autotrophically into the atmosphere, reducing GPP to NPP. Heterotrophic respiration results from the release of CO_2 to the atmosphere from the decomposition of litter and soils, reducing NPP to NEP. Disturbance from anthropogenic sources (e.g. deforestation) leads to further release of CO_2 to the atmosphere which leads to long-term carbon storage. The figure is taken from Stefan et al. (1998) with the numerical values updated based on Beer et al. (2010), Mooney, Roy & Saugier (2001), Harmon et al. (2011) and Canadell et al. (2007).

values of GPP, NPP, NEP and NBP are given. These values along with their uncertainties and temporal and spatial variability are discussed next.

GPP

GPP is estimated to be 123 Gt C yr^{-1} with an uncertainty, represented as a standard deviation, of ± 8 Gt C yr^{-1}, based on the period 1998 to 2005 (Beer et al., 2010). Figure 1.6 shows how the spatial variation of GPP (left panel) and detailed estimates of it based on different latitudes. Areas around the equator, which predominantly include tropical forests and savannahs account for 60% of the total GPP across the globe. The most noticeable observation about figure 1.6 is the large variation in GPP estimates between the process models for latitudes close to the equator. The authors of Beer et al. (2010) do not state the number of sites that were used to compute the data driven median estimates and to calibrate the process median estimates for the different biomes, but they do state that this spread could be reduced with more data for the equatorial areas. The authors of

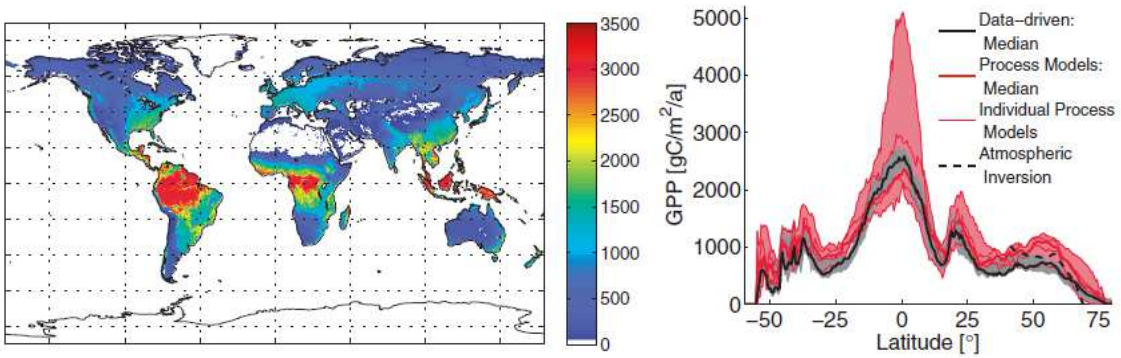


Figure 1.6 The left panel shows the estimates of GPP for different parts of the globe, calculated using the median of five different techniques to extrapolate site level GPP to other areas. The right panel shows the variation of GPP among different latitudes, where the black line corresponds to the amounts depicted in the left panel, and the red line and red shaded area correspond to the median and spread of GPP estimates by latitude computed using five different process models. Both diagrams are taken from Beer et al. (2010).

Beer et al. (2010) did also not explicitly state how the uncertainty of ±8 Gt C $yr^{-1}$was calculated, only that it was determined by propagating the uncertainties of pre-processing of the tower CO_2 flux measurements, the flux partitioning into GPP, the climate and remote sensing data sets, and the structural uncertainties in the diagnostic models. No information is currently available on the inter-annual variability of GPP for the global estimate of GPP and for different parts of the world, however this information is available for NPP, which is described next.

NPP

In the short term, the amount of C take up by plants from the atmosphere is represented by NPP. In order to be able to more accurately how much C is taken up over longer time scales (i.e. NEP, NBP), getting NPP correct will make this much easier. The annual global estimate of NPP is approximately 60 Gt C yr^{-1}, based on extrapolation of field measurements, with Atjay et al. (1979) and Mooney, Roy & Saugier (2001) estimating it to be 59.9 Gt C yr^{-1} and 62.6 Gt C yr^{-1} respectively. These agree with estimates using remote sensing data (e.g. Ruimy et al., 1994; Knorr and Heimann, 1995). However all these studies admit to large uncertainties in their estimates, with Ruimy et al. (1994) approximating this to be around ± 10 Gt C yr^{-1}. Figure 1.5 uses the estimate by Mooney,

	NPP (Gt C yr^{-1})	**Global Carbon stocks (Gt C)**		
	MRS	MRS	IGBP	
		Plants	Soil	Total
Tropical forests	21.9	340	213	553
Temperate forests	8.1	139	153	292
Boreal forests	2.6	57	338	395
Tropical savannas & grasslands	14.9	79	247	326
Temperate grasslands & shrublands	7.0	23	176	199
Deserts and semi deserts	3.5	10	159	169
Tundra	0.5	2	115	117
Croplands	4.1	4	165	169
Wetlands	-	-	-	-
Total	62.6	654	1567	2221

Table 1.1 Estimates of NPP and terrestrial carbon stocks (summed up globally by biome). Atjay is Atjay et al. (1979); MRS is Mooney, Roy & Saugier (2001); IGBP is International Geosphere-Biosphere Programme with the soil carbon layer (Carter and Scholes, 2000) overlaid with the current vegetation map (De Fries et al., 1999) to give average ecosystem soil carbon.

Roy & Saugier (2001) as this is the most recent of all the estimates and is what is used for the IPCC report. Based on this estimate of 63 Gt C yr^{-1} to the nearest whole number, respiration by plants (autotrophic respiration) is computed by deducting this from the GPP estimate of 123 Gt C yr^{-1} from the previous page.

Goetz et al. (2000) computed the inter-annual variation of NPP for the period 1982 to 1989, for different latitudes and biomes. The authors found that annually integrated global NPP was found to vary as much as 12% between years and was very sensitive to air temperature. The authors also found that there was extreme seasonal and moderate inter-annual variation (10 – 60%) in NPP of middle- to high-latitude regions (temperate and boreal forests) and found evidence that this could increase year on year. This is in contrast to the tropical forests and tropical savannas & grasslands, which have a much smaller interannual variability, despite having the largest NPP. A critical biome is

Temperate forests because as stated at the end of the previous page it has one of the largest inter-annual variability meaning that it is more difficult to forecast. It also has the 3rd highest global aggregated NPP and second highest storage of carbon in plant biomass of all the biomes based on estimates from Mooney, Roy & Saugier (2001), meaning that it is not small enough to ignore or pay less attention to (table 1.1). Therefore, one way of reducing the large uncertainty of global aggregated NPP is by improving estimates of NPP in the temperate forests biome.

NEP and NBP

On seasonal time scale, most of the C stored as NPP gets returned to the atmosphere through heterotrophic respiration, where microbes feed on soil organic matter such as leaf and root litter, respiring CO_2 in the process. Heterotrophic respiration (R_H) is estimated to be 55 ± 2 Gt C yr^{-1} (Harmon et al., 2011), however this does not include respiration from decomposition of woody debris (denoted R_{WD}). Bond-Lamberty et al. (2004) state that in an undisturbed forest, R_{WD} is approximately 10% of R_H, thus factoring this in we can approximate the global aggregated heterotrophic respiration to be 61 ± 3 Gt C yr^{-1} with the uncertainty increased to reflect additional uncertainty of the exact contribution R_{WD} to R_H. It should also be noted that R_{WD} can increase to a much larger amount in the first 5 to 10 years after a disturbance – Kurz et al. (2008) found this to be true with a beetle bug outbreak and Bond-Lamberty et al. (2004) found this after a forest fire. However changes to heterotrophic respiration as a result of a disturbance are included in the calculations for NBP, so are ignored for the calculation for NEP. Thus, subtracting heterotrophic respiration as stated above (i.e. 61 Gt C yr^{-1}) from the aggregated global estimate of NPP (i.e. 63 Gt C yr^{-1}) gives an global aggregated estimate of 2 Gt C yr^{-1}. The uncertainty of this NEP estimate is reasonably large, partly because the global aggregated NPP has a large uncertainty (around ± 10 Gt C yr^{-1} as stated previously) and based on other very different NEP estimates, for example Stefan et al. (1998) estimates it to be 10 Gt C yr^{-1}.

As stated in figure 1.4, NEP has a high inter-annual variability, and as stated on page 12, this is mainly driven by variability in precipitation, surface temperature and radiation (Le Quere et al., 2009). In other words, GPP is highest when solar radiation is

sufficiently high and precipitation is at a certain level. A similar relationship between the weather and respiration is also evident – for instance it is widely known that microbes in the soil are most active (and so respire the most) when the soil temperature are at a certain levels (Cable et al., 2009). For these reasons, accurately estimating NEP is a great challenge however very necessary if we are to appreciate the extent to which atmospheric CO_2 is being taken up by the terrestrial biosphere.

Eventually, the net gain C to the terrestrial biosphere from the atmosphere will be NBP, which is NEP minus C returned to the atmosphere from disturbances such as deforestation or forest fires. Using the estimate of the gain of C to the atmosphere from disturbances to be 1.5 ± 0.5 Gt C yr^{-1} from figure 1.4 (Canadell et al., 2007) NBP is approximately 0.5 Gt C yr^{-1}. The uncertainty on NBP is hard to estimate, but based on other estimates (IPCC, 2007, and Stefan et al, 1998), it is probably 0.5 – 1 Gt C yr^{-1}. Therefore, while we are ultimately interested in NBP above the other quantities (such as NEP, NPP and GPP), there are several reasons why it is more important to estimate NEP. First of all, if we wish to correctly estimate NBP it is imperative that we get NEP right. Secondly, the mechanisms which drive NEP (and so how we model NEP) are very different to the mechanisms that determine the land source due to disturbances. In other words, NEP is modelled using theoretical and empirical principles of photosynthesis and plant and soil respiration (Parton et al., 2007), whereas disturbances such as forest fires or beetle bug outbreaks are often modelled in a more stochastic way (Williams et al., 2012). Finally, there is a large global network of flux towers which measure NEP on a continuous basis (Baldocchi, 2003). For these reasons, this thesis focuses on improving estimates of NEP. At this point, it should be noted that Net Ecosystem Exchange (NEE) is sometimes used instead of NEP. NEE is simply –NEP and represents the net loss of CO_2 from the atmosphere to the terrestrial biosphere via plants. Traditionally, NEP is used by those modelling the carbon cycle from the terrestrial perspective (how much C is the forest is gaining), and NEE is used by those modelling the carbon from the perspective of the atmosphere (how much C the atmosphere is losing). However, many of the modellers from the terrestrial perspective use NEE (e.g. Williams et al., 2005). This is probably due to the fact that those collecting data on C accumulation to the land via plants through the network of Eddy-Covariance flux towers use NEE instead NEP

(Vickers, 2009), perhaps causing some of the modelling community to also adopt NEE. The likely reason why NEE is used instead of NEP where measured from the flux towers is because the eddy-covariance instrumentation which make the measurements are usually located high up on the flux tower and so could be thought of looking down onto the forest or plants, i.e. observing the net C gain to the terrestrial biosphere from the perspective of the atmosphere. NEP is also used by those measuring and modelling the terrestrial carbon cycle (e.g. Quaife et al., 2007).

Throughout this thesis various notations of the quantities of interest will be used, the main ones represented in figure 1.5. Definitions of these are:

- Gross Primary Production (GPP) = the total amount of carbon taken up by plants through photosynthesis.
- Autotrophic Respiration (R_a) = the amount of carbon respired from the plant. R_a can be further sub-divided into maintenance respiration and growth respiration; abbreviated as R_m and R_g in Turner et al. (2005).
- Heterotrophic Respiration (R_h) = the amount of carbon respired from soils and dead plant matter (e.g. litter).
- Total Respiration (R_{tot}) = $R_a + R_h$;
- Net Primary Production (NPP) = GPP – R_a;
- Net Ecosystem Production (NEP) = GPP – R_{tot};
- Net Ecosystem Exchange (NEE) = R_{tot} – GPP = –NEP;
- Net Biome Production (NBP) = GPP – R_{tot} – disturbances (e.g. deforestation).

The units to these quantities depend on the spatial scale one is working at. For this thesis, the most appropriate units are grams of Carbon / day (g C day^{-1}).

As stated earlier, the main quantity we are interested in estimating in this thesis is NEE = –NEP. If, over a particular period of time e.g. a day, GPP > R_{tot}, then NEE will be negative and represents the amount of carbon lost from the atmosphere by the land (referred to as a sink of carbon). Conversely, if GPP < R_{tot}, then NEE will be positive and will indicate the amount of carbon gained by the atmosphere. We are also in interested in estimating GPP, R_a, and R_h correctly as all of these could be poorly estimated but NEE estimated well.

1.2.2 Ecosystem Observations and Data

DATA MEASURED OR INFERRED FROM MEASUREMENTS FROM GROUND.

Obtaining data of ground based NEE, GPP and R_{tot} is important as they can be made continuously over a long period of time, and so allow us to understand how NEE (for example) changes from year to year, and particularly how certain types of weather (for example if one year is particularly hot, or dry or wet) influence NEE. NEE is measured continuously using the Eddy-Covariance technique, which involves calculating the covariance between the vertical wind speed and the CO_2 concentration (full details given in chapter two), and is typically measured at different heights of the forest canopy by attaching the measuring equipment to a steel scaffold, or flux tower as they are commonly known (Baldocchi, 2003). Although GPP and R_{tot} are not measured directly in forests (on a continuous basis), such data can be inferred using partitioning models (Reichstein et al., 2005)

However, it should be noted that where there is vegetation of a very low height, such as grasses or shrubs, GPP and R_{tot} can be measured almost directly through a different measuring technique. This involves placing a small static chamber over the vegetation, which contains within it one or two circulating fans and a photosynthetically active sensor (Bachman et al., 2010). This measures NEE in the same way as the sensors do on the flux towers do for the Eddy-Covariance technique, however if the flask is covered so as to block out all solar radiation penetrating the flask, then the measured flux is then R_{tot}. If measurements of NEE and R_{tot} are made consecutively, then under the assumption that the two measurements were made at the same time, GPP can be inferred by subtracting the NEE from the R_{tot} measurements (Bachman et al., 2010). Such flask measurements are not measured continuously over a long period of time as they require a significant amount of human labour. For forests, continuous long-term measurements of NEE can only be made using the eddy-covariance technique on flux-towers, and are much less labour intensive. Thus for forests, data for GPP and R_{tot} can only be inferred using partitioning models as stated above.

Currently, data for NEE, GPP and R_{tot} are available on half-hourly intervals at over 200 sites across the globe (Baldocchi, 2003). This network of sites is called Fluxnet,

with data freely available via an internet based database. The Fluxnet database holds other measurements, which include:

- Leaf Area Index (m^2/m^2), or LAI, which is defined as the 'one-sided green leaf area per unit ground area, in broadleaf canopies, and as the projected needle leaf area in coniferous canopies' (Myneni et al., 2002). LAI is an indicator how much photosynthesis is taking place, among other things, and so is an important ecological variable (Breda, 2003). LAI can be inferred from direct measurements of foliar biomass which are obtained by stripping the leaves off a tree and feeding them through a scanner. LAI is then inferred by dividing the foliar carbon mass by Leaf Mass per Area (LMA); see Williams et al. (2005), De Kauwe (2008), Fox et al. (2009), Richardson et al. (2012) and Spadavecchia et al. (2011). In reality LMA can vary between leaves on a tree, especially between different levels of canopy (Poorter et al., 2009), but for computational ease it is often treated as fixed (Fox et al., 2009; Williams et al., 2005; Quaife et al., 2007). Obtaining inferred measurements of LAI by direct means is the most accurate among methods, but it can be time-consuming and it is also very destructive (Gower et al., 1999). More commonly, LAI is determined by indirect methods, and from the ground this is done by measuring the gap fraction of the canopy using optical techniques such as fish-eye photography (Gower et al., 1999). A fuller description is given in chapter 2. Although obtaining LAI measurements using indirect methods is not as labour intensive as making woody, litter and root biomass measurements (described next), it is still time consuming and for this reason LAI is typically only measured a handful of times throughout the year (Williams et al., 2005).
- Woody, litter and root biomass (gC/m^2), which indicate how much carbon is stored in the tree trunks, leaf litter and tree roots for every unit area. Above ground woody biomass is normally estimated by first recording the dimensions of the tree and then estimating the biomass using allometric relationships determined from the destructive harvest of a variety of trees covering a range of different sizes (Law et al., 2001). Biomass from litter is determined by manually collecting the litter into multiple trays and separating it into foliar litter and woody litter. Fine root biomass can be estimated by destructive means by extracting the roots

contained in the soil using soil cores and then extracted in the laboratory after cleaning off any soil (Law et al., 2001). Due to these measurements being labour intensive, they are usually only made around five or six times a year, sometimes less frequently (Law et al., 2001).

- Soil moisture content, which is a measure of how much water is present in a unit amount of soil. Measurements are made using a frequency domain reflectometer probe (Xu et al., 2004) either at a selection of depths such as 5cm, 10cm and 20cm, or a selection of ranges of depths, such as 0-15cm, 15-30cm, etc…

Meteorological data are also available from Fluxnet, and these can include: atmospheric CO_2 concentrations, precipitation, transpiration (mm) and solar radiation (W/m^2).

DATA INFERRED FROM MEASUREMENTS MADE FROM SATELLITE SENSORS

As stated in the previous section, satellites are increasingly being used to make ecological measurements, via their sensors, as they provide frequent and global coverage. Common ecological quantities that are inferred from satellite data are: GPP, LAI, the fraction of photosynthetically active radiation (fPAR) and soil moisture. The sensors on the satellites do not measure these quantities directly, but instead measure the sun's reflection on the vegetation. The ecological quantities are then inferred using a 3D radiative transfer model (RTM). The RTM has, among other things, canopy reflectance of solar radiation (or simply 'reflectance') as its output and uses various biophysical parameters, such as LAI, as inputs (Lewis, 1999). The RTM is inverted to find the parameters (LAI, etc…) that produce the best fit between the modelled and observed reflectance; for example, for the LAI inferred from the MODIS sensor of the TERRA satellite, see Knyazikhin et al. (1998) and Myneni et al. (2002). To avoid multiple solutions (e.g. if inferring LAI and if there are multiple values of LAI, which produce near optimal fits of modelled to observed reflectance) a biome classification map is also used in the retrieval. From this point on, we refer to 'satellite data' as being data of ecological quantities inferred from the processing of canopy reflectance data made from satellite sensors, using models.

DATA CHALLENGES

Despite their advantages, satellite data can contain gaps and important quantities may be unmeasurable. Using data on their own can also make it difficult to understand particular ecosystem processes. Furthermore, it is not possible to make realistic forecasts of the ecosystem under different scenarios using solely data, for example under varying atmospheric CO_2 levels. Models can overcome these problems, and these are described and discussed next.

However a data challenge which, for the most part, models cannot overcome is how to quantify the uncertainty of data. This is important because if untruthworthy data is being used to make inferences (e.g. make regional estimates of GPP) or to validate model outputs, then these inferences and model outputs will likely be wrong. Some progress has been made with quantifying data uncertainty. For example, Hollinger and Richardson (2005) suggested the random error of Eddy-Covariance measurements followed a Laplacian distribution and provided a method for estimating the statistical parameters for this distribution, based on pairs of measurements made at two nearby flux towers. Richardson & Hollinger (2006) refined the methodology so as the parameters could be estimated using only one flux tower, based on pairs of measurements separated by 24 hours and made under similar meteorological conditions. Despite this progress, quantifying systematic errors of NEE data remains a challenge. For example, it is well known that nighttime eddy-covariance respiration data tends to be under-estimated because the wind-speed is too low (Baldocchi, 2003), but there is little known on the size of this under-estimation.

Despite studies that have validated satellite data with comparable ground based data (e.g. Garrigues et al., 2008), there has been less attention paid to quantifying the uncertainty of these satellite data. A possible exception to this is the current version of the radiative transfer model used to estimate LAI from the MODIS sensor of the TERRA satellite; see De Kauwe et al. (2011) for description of algorithm. Here, a standard deviation value is given for each of the MODIS LAI data points, but this refers to the spread of retrieved LAI that result in the output of the radiative transfer model (i.e. the canopy reflectances) used having a good fit to the observed reflectances (Knyazikhin et al., 1998). It is reasonable to assume that many modellers would treat these standard

deviation values as measures of data uncertainty, but it could be argued that these s.d. values do not really include a representation of uncertainty of the radiative transfer model nor of the raw radiance measurements.

ATMOSPHERIC MEASUREMENTS AND OTHER RELEVANT EO DATA

In this subsection, we briefly describe examples of atmospheric measurements. This is important as they are relevant to inferring C fluxes. The Scanning Image Absorption Spectrometer for Atmospheric Chartography (SCHIMACHY) and the Global Greenhouse Gas Observation by Satellite (GOSAT) are two satellite spectrometers that measure trace gas concentrations in the atmosphere (Breon & Ciais, 2010). Trace gas concentrations are inferred by determining the absorption of the wavelengths of the solar radiance received by the satellite sensor reflected from different trace gases (Breon & Ciais, 2010). A radiative transfer model is then inverted to estimate the trace gas concentration. This is because the radiative model represents the radiance spectrum as a function of trace gas concentration; by observing the changes in the reflected solar radiance spectra it is possible to infer changes in the gas concentration (Barkley et al., 2006).

The fraction of Photosynthetically active radiation (fAPAR) is defined as the fraction of incoming solar radiation that is absorbed by a plant. It is an important climate variable because it is directly related to GPP and used by some models to infer assimilation of CO_2 in plant material (Knyazikhin et al., 1998). fAPAR can be inferred from satellite measurements of solar radiation in the same way as LAI (Knyazikhin et al., 1998). fAPAR can also be used as an indicator of the health and evolution of vegetation cover; as a result it replaces the Normalized Difference Vegetation (NDVI) assuming it is estimated well. NDVI was (and still is) useful as it can help to identify simple climate characteristics such as spring vegetation green-up and subsequent summer and fall dry-down (Tucker et al., 2001). NDVI was one of the first climate variables inferred from space and so it is useful in that it is one of the longest data streams of all remotely sensed data (Tucker et al., 2001). NDVI is an equation which takes into account the amount of infrared reflected by plants. It is connected to vegetation because healthy vegetation reflects very well in the near-infrared part of the electromagnetic spectrum (Tucker et al., 2001).

Global spaceborne observations of solar induced chlorophyll fluorescence is becoming increasingly important, due to its strong linear correlation with GPP (Frankenberg et al., 2011). Traditionally GPP has been inferred from space using remotely sensed vegetation indices such as LAI or fAPAR (as described above), but using florescence data has enabled far better predictive abilitity of GPP (Frankenberg et al., 2011). A drawback of the florescence data from the GOSAT satellite is that it is incredibily noisy (Guanter et al., 2012), and so methods need to be employed to extract the true signal.

1.2.3 Ecosystem models

In this section, an overview is presented of different models that could be used in this thesis, along with the advantages and disadvantages of using each one. The models vary according their size, complexity, and what their main use is. In chapter two, a more comprehensive description is given for the different models, and then an argument is made for which model is most appropriate for this thesis based on the context and aims of this work. For this subsection, we begin by describing common features contained in process models, the purpose of the different models of the terrestrial carbon cycle, then a brief overview is given for 12 models, outlining the advantages and disadvantages of adopting each.

The central role of a model is to try to represent and understand reality, and they can be of different forms. Models can be statistical in nature, i.e. built using data, and carry certain assumptions about how the independent variables (e.g. time) and dependent variables (e.g. GPP) are related, e.g. in a linear way. To estimate quantities such as NEE, GPP and R_{tot}, process models are often used. These models are typically based on differential equations which describe our best knowledge of the plant processes and physics of the system. These processes are controlled by parameters, which are physical quantities giving information about what the system is doing. Such models are normally dynamic since we often wish to understand how these quantities of interest change over time. Regardless of their size, all ecosystem process models will have several things in common, including the following:

- Pools, which are the stores of carbon, for example the foliar carbon pool, the soil carbon pool, the woody carbon pool, etc…
- Fluxes, which describe the flow of carbon between the pools.
- A set of parameters, for example the loss rate of woody carbon.
- States, which are the important quantities we wish to estimate, such as NEE, NBP, GPP, R_{tot}, R_a, R_h, but also the pools and the fluxes.

There is a whole array of different models of the terrestrial carbon cycle, and these vary according to their purpose. While all aim to estimate the states of the terrestrial carbon cycle (e.g. NEE), some are designed to make multiple estimates across the globe requiring hundreds or thousands or parameters due to the different types of vegetation that need to be modelled differently. Other models are designed to make site level or regional estimates only and often built with a particular vegetation type in mind (e.g. Evergreen trees, or crops, …) and so have fewer parameters that those that are used to make global estimates. It should be noted here that global models can also be put into a site level or regional level mode, where they only make estimates for a particular site or region. In this way, only a subset of parameters would be used and the global model in effect becomes a site level or regional model (Kennedy et al., 2008).

The models also vary according to how complex the processes (e.g. vegetation dynamics, photosynthesis, …) are modelled. For example, some might use the Farquahar equations (Farquhar & Caemmerer, 1982) to model photosynthesis, e.g. the SPA model (Williams et al., 1996). Alternatively, others might model photosynthesis in a simpler way, such as using an approximation to the Farquhar equations (e.g. the Aggregated Canopy model used in the DALEC model, Williams et al., 2005), or estimate it empirically. Models may also vary in size according to the number of processes that are included. For example, the SiPNET model (Braswell et al., 2005) simulates the flow of carbon from photosynthesis to respiration and includes a simplified representation of the water dynamics in the plant and soil (i.e. soil moisture, transpiration). However, the DayCENT model (Parton et al., 2001) not only models the carbon and water cylces, but also simulates the Nitrogen, Phospherous and Sulphur dynamics of the plant.

Models can also differ according to time-scales. Some are used to assess historical dynamics, others are used for monitoring, while some are also used for making

predictions. Models that are used for monitoring include those that seek to quantify the current rates of certain processes, such as GPP and NPP and how they might vary temporarily and spatially, such as Ruimy, Saugier & Dedieu (1994), Running et al. (2004) and Yuan et al. (2007). Under the category of monitoring, it could also be argued that models that are used to infer measurements, such as radiative transfer models (e.g. Myneni et al., 2002) which are used to infer vegetation characteristics from reflectance data from a satellite sensor, and partitioning models (Reichstein et al., 2005) that are used to estimate GPP and respiration fluxes from NEE measured at flux towers, should be included. A large focus for many modelling activities is to be able to predict what vegetation characteristics, such as NEE, woody biomass, etc.. will be in the future. For example, Balshi et al., (2009) used the TEM model to estimate decadal total carbon emissions for the North American boreal forests due to wildfires up until 2100. However, making accurate predictions is difficult, and so an increasingly common approach is to use a model to estimate what a certain climate quantity will be under a changing climate. For example, Parton et al. (2007) used the DayCENT model to assess how soil water content, plant production, soil respiration and nutrient mineralisation would change under an elevated atmospheric CO_2 environment. Another example is Ogle & Pacala (2009) who used the ACGCA model to predict forest succession, growth and species diversity for two different species of tree under different scenarios based on varying gap dynamics and by adjusting some the model parameters.

We now summarise 12 different terrestrial ecosystem models, starting with the simplest model and proceeding to the more complex ones. The list of models, along with the number of parameters, number of states and the spatial scale they operate on is given in table 1.2.

THE PROS AND CONS OF SMALL MODELS, WITH EXAMPLES

The main advantage of the small models described here, namely OpTIC, SPA, DALEC, and SiPNET is that they are fast to run. This is due to there being a small number of parameters, process representation and states to estimate. The small models listed here were also all developed to operate at a site level. The main disadvantage of these models is that due to their simplified nature, they lack sufficient process representation. For exam-

Model	No. of parameters	No. of states	Spatial Scale	Temporal Scale
OpTIC	4	2	Local	Daily
DALEC	11 / 17	17 / 22	Local	Daily
SIPNET	25	22	Local	Half-Daily
SPA	28	21	Local	Half-hourly
ACGCA	32 × No. of trees	12 × No. of trees	Local–Regional	Yearly
DayCENT	~ 30	~ 50	Regional	Daily
CASA	35	20 × No. of Grid cells	Regional – Global	Monthly
TEM	> 50	18 × No. of Grid cells	Continental–Global	Monthly
SDGVM	> 50	> 50	Global	Monthly
BETHY	> 50	20 × No. of Grid cells	Global	Monthly
LPJ	> 50	> 50	Global	Monthly
ORCHIDEE	> 50	> 50	Global	Monthly

Table 1.2 A list of twelve terrestrial carbon models, which vary according to the level of complexity, the spatial scale and time-step they operate over, and the particular processes they try to estimate.

ple, while all of them model part of or all of the terrestrial carbon cycle, they do not include any strong representation of the water cycle (except for SPA) or Nitrogen cycle for example, which are linked to the carbon cycle. However, it could be argued that including extra process detail may not be necessary. For example DALEC has been shown to produce accurate estimates of NEE (Williams et al., 2005) despite being relatively simple. Additionally, because small models operate on a smaller spatial scale than larger models, they do not provide estimates of say GPP or NEE at the global or continental scale. However, they are useful for providing for more fine-scale estimates of NEE (for example), compared to larger global models whose estimates tend to be coarser. We now provide a simple description of the simplest four models, listed in the first four rows of table 1.2.

The OpTIC model (Trudinger et al, 2007) was developed as a highly simplified representation of the carbon dynamics in a terrestrial biosphere model in order to compare

different methods of estimating its parameters. Its two states represent carbon stored in living biomass and soil carbon.

The Data Assimilation Linked ECosystem (DALEC) model (Williams et al., 2005) simulates the flow of carbon from photosynthesis through to respiration for evergreen trees, and estimates carbon stored in foliage, roots, litter, woody biomass and soil. GPP is estimated through a sub-model called the Aggregated Canopy Model (ACM) which uses meteorological data such as minimum and maximum daily temperature, solar radiation, and atmospheric CO_2 concentration. The ACM is simplified version of the photosynthesis submodel used in SPA (described on the next page). There is also a feedback which links the foliar carbon pool to GPP, since (put simply) the more leaves (foliage) on the trees, the more photosynthesis takes place. Another version of the model called DALEC-D is also available; this is the same as DALEC except that it is used for deciduous vegetation and so incorporates an extra pool to store labile carbon, which is much like a savings account that stores carbon from the foliar carbon pool during the growing season and then expends it during the following Spring bloom in order to generate the foliage.

The Simplified PhotosyNthesis and Evapo-Transpiration (SiPNET) model (Braswell et al., 2005) operates on a half-daily time-step rather than a daily one as is the case with DALEC. However there are fewer carbon pools in this model compared to DALEC, which include pools of plant wood carbon, plant leaf carbon and soil carbon. Unlike DALEC, SiPNET has a simple soil moisture sub-model; if soil moisture falls below a certain value the plant becomes water-stressed and results in Gross Primary Production being limited. GPP is estimated in SIPNET by first computing its maximum potential value if there were no limitations on resources. This maximum GPP is then multiplied by various scaling factors which describe the limiting factors, such as radiation, water stress, surface temperature, Vapour Pressure Deficit.

Unlike the previous two models, the Soil-Plant-Atmosphere (SPA) model (Williams et al., 1996) simulates ecosystem photosynthesis and water balance at 30 minute time-steps and at a fine spatial scale which include multiple canopies and soil layers. It consists of two components, which are physical and biological. The physical component determines the physical parts of the vegetation, such as the structure of the

canopy and the level of absorption of photosynthetically active radiation and other wavelengths at different layers of the canopy. It also computes the leaf boundary layer conductance and determines soil water availability. The biological component simulates the assimilation of CO_2 from the atmosphere by photosynthesis and transpiration. It also models the factors that affect C uptake and water loss, such as the variation of leaf biochemical parameters with foliar Nitrogen content, irradiance and leaf temperature.

THE PROS AND CONS OF MEDIUM SIZED MODELS, WITH EXAMPLES

We now present two examples of medium size models used to simulate vegetation dynamics the terrestrial biosphere. These models are advantageous in that the combine the efficiency of smaller models, which are fast to run, but contain a closer level of complexity to larger models which include more processes; in this way they can be a good compromise between the smaller and larger models. However, like smaller models, they tend to not provide global estimates of the quantities of interest, such as NEE. Nonetheless they typically can provide regional estimates, which can be useful for estimating regional inventories of C stocks, such as the DayCENT model (described below). The extra level of complexity, compared to the simpler models, also means that more of our understanding of the processes of the plant biology and phenology are able to be incorporated, resulting in the potential improvement of terrestrial C cycling estimates, but also the possibility to estimate the dynamics of other processes and quantities linked to the C cycle, such as the Nitrogen cycle as in the DayCENT model or the allometrics of the tree as in the ACGCA model, both described next.

The DayCENT model (Parton et al., 2001) simulates the terrestrial Carbon, Nitrogen, Phospherous and Sulphur dynamics. Including the representation of the Phospherous and Sulphur dynamics is important as like Nitrogen, if there is insufficient availability of these nutrients in the plant cells, this poses a limitation on the amount C fixed through photosynthesis. Like the simpler models, the model simulates the accumulation and exchange of C within and between different pools, such as foliage, the woody part of the tree and the soil. It also models the respiration of C from the plant directly and from the soil. Unlike the previous models, it not only estimates NEE but it also predicts the emissions of Nitric and Nitrous oxide (N_2O and NO_x). These model

predictions are then used to inform the United States Environmental Protection Agency who compile the National inventory of N_2O emissions from U.S. agricultural soils (Williams et al., 1995).

The Allometrically Constrained Growth and Carbon Allocation (ACGCA) model (Ogle & Pacala, 2009) is a different type of terrestrial carbon model compared the previous ones listed. While all the models listed here predict GPP, NPP, NEE, etc… the ACGCA differs in that it predicts the height, width of individual and cohorts of trees. The focus of this model is therefore to assess the life cycle of trees, as opposed to predicting seasonal trend, which is why the time-step is in years instead of months or a smaller unit of time. The model predicts how different species of trees interact and compete with each other as they grow. In this way, the main focus of this model is to predict the future state of the forests for a particular region (such as South East United States), i.e. what species will exist and in what quantity for a particular location. Like previous models listed here, it also estimates the carbon dynamics for different parts of the tree, such as the foliar biomass, the woody biomass and the root parts of the tree.

THE PROS AND CONS OF LARGE MODELS, WITH EXAMPLES

The main focus of the larger models, listed in the bottom six rows of table 1.2, is to include as many of the processes that affect the terrestrial carbon cycle. All of the large models listed here operate under a monthly time-step, as this enables prediction of intra-annual as well as inter-annual changes of the state of the terrestrial carbon cycle. Most also operate at a spatial scale of 0.5^o–1^o latitude and 0.5^o–1^o longitude grid cells. They are the only type of model (compared to the smaller models listed earlier) that are able make global estimates of the quantities of interest, such as GPP, NEP and NEE. This ability to fully quantify the terrestrial C cycle across the globe means that they are the most important among all the other types of models.

Although the power and computational efficiency of supercomputers have improved dramatically over the past 30 years, most researchers do not have access them. Thus, the obvious downside to these models is that they are very computationally expensive to run. Some users try to overcome this, to a certain extent, by decreasing the spatial resolution the model predicts over, or by making predictions over a smaller spatial

scale, such a region or a continent (Woodward et al., 1995). Turning off parts of the model which are not needed for individual experiments is another way of making these models more accessible from a computational point of view (Kennedy et al., 2008). Another downside to these models is that by including as much complexity on our understanding of the processes as possible, they can become very hard to manage. For example, making a small change to the model may result in other parts of model giving strange output which may be hard to diagnose.

The Carnegie Ames Stanford Approach (CASA) model (Potter et al., 1993, Carvalhais et al, 2008) was created to make global estimates, at the monthly time-step, of seasonal patterns of NPP, biomass and nutrient allocation, litterfall and nitrogen miteralization, and CO_2 production. It includes 13 different vegetation classes, including four classes of broadleaf trees, two classes of Needleleaf trees, two types of shrubs, grasslands, Tundra, bare soil, cultivation and ice. It operates by using driving data of Normalized different vegetation indices (NDVI), solar radiation, temperature, precipitation to estimate NPP by a semi-empirical methods based how NPP varies with intercepted photosynthetically active radiation (IPAR) for different vegetation types – this way of estimating net C photosynthesis assimilation rates is referred to as a light-use efficiency approach. Soil moisture is estimated using a sub-model, which is driven by temperature and precipitation which, along with estimates from the NPP submodel and estimates of litterfall based on NDVI, is then used to estimate carbon and Nitrogen dynamics in the soil. From this, soil respiration rates are simulated which are then used to predict NEP from the NPP predictions (since NEP = NPP – soil respiration). The soil moisture and soil carbon-Nitrogen submodels are also driven by soil texture data. Finally, there is a feedback of Nitrogen uptake from the soil Carbon-Nitrogen submodel to the NPP submodel.

The Terrestrial Ecosystem Model or TEM (Raich et al., 1991; Balshi et al., 2009) was built to predict the major carbon and nitrogen fluxes and pool sizes in terrestrial ecosystems at the continental to global scales. The five major elements of the state vector in TEM are carbon stored in living vegetation, Nitrogen stored in living vegetation, organic carbon in detritus (i.e. dead organic material such as leaf litter) and soils, organic Nitrogen in detritus and soils, and available inorganic soil carbon. The maximum rate of

C assimilation into the plant per unit area was determined by a simple diffusion scheme for each vegetation type. GPP was then estimated by reducing (by a factor less than 1 but greater than 0) this maximum rate of C assimilation according to various limiting factors on photosynthesis, such as irradiance of photosynthetically active radiation (PAR), atmospheric CO_2 concentration, moisture availability, mean air temperature, the relative photosynthetic capacity of the vegetation and nitrogen availability.

The Sheffield Dynamic Global Vegetation Model (SDGVM) is a generalized, global-scale model that predicts vegetation structure and dynamics using inputs consisting of meteorological, CO_2 and soil texture data (Woodward and Lomas, 2004). The physiology and biophysical module gives a representation of the carbon and water fluxes of vegetation (Woodward et al., 1995). The water and nutrient fluxes module simulates the exchange of canopy CO_2 and water vapour along with the nitrogen uptake and partitioning within the canopy. The plant structure and phenology module define the vegetation leaf area index and the phenology of the vegetation (i.e. the timings at which different plant processes occur, e.g. Spring bloom). Nine plant functional types (PFTs) are used in the model, which include two types of Evergreen tree types and two Deciduous types, croplands, C3 and C4 grasses, urban and bare ground. GPP is estimated using the Farquhar model (Farquhar et al., 1980) which simulates the biochemical processes of electron transport, the Calvin cycle and photorespiration. SDGVM can also estimate NBP, by simulating one type of disturbance, fire, and this is done based on the probability of a fire occurring, where the probability is determined from deviations away from the ideal environmental conditions (mainly temperature and precipitation amount) required for a fire to start (Woodward et al., 2001).

The Biosphere Energy-Transfer and Hydrology (BETHY) model (Knorr, 2000) was built in response to discrepancies between NPP estimates from different terrestrial ecosystem models. Part of the discrepancy could be caused by how certain processes are modelled. For example, common approaches to simulate photosynthesis use the Farquhar model (Farquhar et al., 1980) as is used in SDGVM, a light-use efficiency method that is used in CASA above, or by a simple diffusion scheme as is adopted by TEM. In BETHY, photosynthesis can be simulated by either the Farquhar model or a light-use efficiency method. In this way, global estimates of GPP and NPP can be compared using

the two different photosynthesis models, and this can be seen as one way of quantifying the model uncertainty in relation to structure. BETHY is split up into four main submodels which simulate (i) the energy and water balance, (ii) photosynthesis, (iii) phenology, and (iv) the carbon balance. As with the other larger models, it estimates GPP, NPP and NEP at a monthly time-step. It includes 23 vegetation types including four types of evergreen trees and shrubs, four types of deciduous trees and shrubs, four types of C3/C4 grasses, nine crop types, tundra vegetation and swamp vegetation.

The Lund-Potsdam-Jena (LPJ) dynamic global vegetation model (Sitch et al., 2003) includes the main processes of other models, such as a photosynthesis model, carbon, water and nitrogen dynamics in vegetation and soils, and covers a range of different vegetation types or PFTs (Plant Functional Types) as they are commonly known. Unlike other models, except for ACGCA, LPJ also includes representation of vegetation structure, dynamics, competition between PFT populations, and soil biogeochemistry. The Farquhar model is used in LPJ to simulate photosynthesis and estimate GPP. NPP is then predicted by subtracting autotrophic respiration, calculated as the sum of maintenance and growth respiration, from GPP. As with other models, the C stored in NPP is then allocated to different C stores such as woody carbon, foliar carbon and root carbon.

The Top-down Representation of Interactive Foliage and Flora Including Dynamics (TRIFFID) model (Cox, 2001) is a dynamic global vegetation model, which updates the plant distributions and soil carbon stores. TRIFFID is unique in that competition between the five different PFTs is simulated using a Lotka-Volterra approach. The PFTs are: Broadleaf trees, Needleleaf trees, C3 grass, C4 grass and shrubs. Photosynthesis is modelled using the Farquhar model (Farquhar et al., 1980) and scaled up to the canopy level using Sellers et al. (1992). As with other models, NPP is determined by subtracting autotrophic respiration from GPP. The remaining C or NPP is then allocated to the C stores of the tree. Like some of the other models such as SDGVM, TRIFFID also estimates the removal of C from the trees from disturbances such as natural fires.

LAND SURFACE MODELS AND CONCLUDING REMARKS

Land surface models, such as JULES (Clark et al., 2011) are a further level of complexity to the larger models described above. For example, JULES not only models the carbon, nitrogen and water cycle of the terrestrial biosphere, but also provides estimates for the surface temperature and simulates the shortwave and long-wave radiative fluxes, and the sensible and latent heat fluxes between the land surface and the atmosphere. TRIFFID is the vegetation dynamics module of JULES. Another example is JSBACH (Brovkin et al., 2009) which uses BETHY to simulate the evolution of the vegetation.

In this section, an overview has been given for the different modelling options. In chapter 2 a subset of the models listed above are described again but with more precise detail about how different processes, such as photosynthesis or respiration, are simulated. Furthermore, a more critical review is provided in chapter 2 for the different modelling approaches. Then, an argument is made for the reason for the particular model that is to be used in this thesis based on the required outcomes of this work.

1.3 Combining data and models

1.3.1 Non-DA approaches

While observations and models each provide knowledge about the state of the system, they are more useful when used together. Observations on their own may contain gaps, and certain quantities, parameters or processes may be unmeasurable. A model on its own may contain the right mathematics to describe how the system behaves under certain conditions, but we only know if the model is correctly estimating reality if we can compare it with observations. Thus, models and observations complement each other because: (i) the model allows us to fill in missing (temporal) gaps in the data, to estimate quantities and processes where no measurements are made (i.e. spatial gaps), and to estimate and represent unmeasurable quantities and processes; (ii) the observations allow us to determine whether the model, represented by its parameters and processes, accurately represents what we observe the system to be doing, and so enable us to make educated predictions of how the system will behave in the future. In this way, the benefit

of models such as the last six listed in table 1.2 is that they can provide us with global estimates of the quantities that we are interested in, such as GPP, NPP and NEE. The benefit of the observations is that they can be used to check that the model's estimate of a particular quantity, at locations and times when observations are made, are consistent with what is observed.

Confronting the model with the observations can be done in a number of different ways. A common approach is to compare observations with estimates of these quantities from the model. For the JULES model (Clark et al., 2011), which operates globally, this is carried out by comparing model outputs against a number of different measurements (e.g. NBP, evapotranspiration, latent heat, etc…) at particular sites which have different characteristics (e.g. different meteorological conditions, vegetation, terrain, …). If the model is over or under-estimating certain measurements at certain sites and times, it can be interrogated to determine why this is the case. Also, if the model is wrongly estimating the timing of certain ecological events, such as the Spring bloom, this might have implications for other quantities that the model estimates.

In Krinner et al. (2005) the outputs of the ORCHIDEE model, a dynamic global vegetation model, were validated against measurements as a way of ensuring that the model was able to reproduce what was observed. ORCHIDEE simulates the main processes of the terrestrial carbon cycle such as photosynthesis, autotrophic and heterotrophic respiration and fire, as well as the competitive nature of vegetation, such as competition for light. The model also estimates the distribution of vegetation, leaf density and carbon stocks and fluxes and it is these that were compared against observations located at 28 FluxNet sites in Krinner et al. (2005), which varied according to location with every continent and major biomes represented. The authors described the fit of model to the observations as satisfactory. One of the main model outputs compared with observations was LAI, however the LAI data being compared against were inferred from radiance data measured by the MODIS sensor of the TERRA satellite. In relation to this the authors made an interesting point: 'the modelled LAI is generally too low outside the equatorial belt. … The under-estimation is particularly strong in the middle and high latitudes in summer, partly because absolute LAI errors can be larger where absolute LAI values are large.' The authors continue by saying that satellite LAI estimates are subject

to fairly large uncertainties for dense canopies. This may be true, but the authors do not attempt to quantity the uncertainty of the MODIS LAI in a more precise way or at any other spatial region. For the most part, it appears that the authors treated the MODIS LAI as if it was subject to very little error. This lack of quantification of the observational uncertainty occurs in other studies which attempt to quantify the accuracy of global terrestrial carbon models, for example Woodward & Lomas (2004), Morales et al. (2005) and Williams et al. (1996).

1.3.2 DA approaches to combining data with models

An increasingly important way of using observations to improve the performance of models is Data Assimilation (DA). In DA, the observations and model are combined in a statistically optimal way, rather than the observations being compared with model outputs. The major advantage of DA is that uncertainties in the model and observations (expressed as probability distributions) are specified. A key feature is that under certain conditions the uncertainty in the model outputs, after the observations have been assimilated, is always less than the uncertainty in either the observations or model, e.g. Williams et al. (2005), Quaife et al. (2008). These conditions include: (i) the sensitivity of the model output to the observations, i.e. there's little point assimilating observations from the ocean if the model simulates only the terrestrial carbon cycle with no link to the ocean; (ii) the observations must be independent with one another and with the model output. It should be noted that this reduction in uncertainty as a result of DA can happen if we were to merge two observations together. The point is that by assimilating observations into a model, we not only reduce the uncertainty of the target variable, but we also have more confidence in our modelled estimates of other related quantities. For example, Dente et al. (2007) assimilated Leaf area index data into a wheat crop growth model. This had the consequence of reducing the uncertainty of predicted leaf area index; however this was not the goal – the purpose of the study was to improve the predicted wheat yield, as well as quantifying the corresponding uncertainty, which was achieved by the assimilation of LAI.

DA can be used to estimate the states or parameters of a model, or both. In the next subsection, we define formally what is meant by parameters and states and how both can be estimated using DA.

1.3.3 State and parameter estimation

As stated in subsection 1.2.3 the state variables are the quantities of interest, such as GPP, NEE, autotrophic and heterotrophic respiration, foliar carbon, and woody carbon. The states are governed to large extent by parameters which are physical quantities giving information about what the system is doing. A more formal definition of state variables and parameters of a model is given by Bard (1974):

> *We refer to the relations which supposedly describe a certain physical situation, as a model. Typically, a model consists of one or more equations. The quantities appearing in the equations we classify into variables and parameters. The distinction between these is not always clear cut, and it frequently depends on the context in which the variables appear. Usually a model is designed to explain the relationships that exist among quantities which can be measured independently in an experiment; these are the variables of the model. To formulate these relationships, however, one frequently introduces "constants" which stand for inherent properties of nature (or of the materials and equipment used in a given experiment). These are the parameters.*

Chen et al. (2008) define the states and parameters in reference to a dynamic model. They state that a dynamic model can be expressed as one or more discrete time nonlinear stochastic processes (although it can also be linear):

$$X^{k+1} = f(X^k, U^k, \theta) + \varepsilon^k$$

where *k* denotes the time-step; X^k is the vector of state variables, such as carbon flux or storage attributes (e.g. GPP, NEP, NEE, foliar C, etc…); f is the model operator which defines how the states change from one time point to the next; U^k is the set of externally

specified time-dependent driving data, such as meteorological data and soil data; θ is set of parameters; and ε^k represents the model error.

We next give a generic definition of how the states and parameters of a model can be estimated using DA. We do not refer to a specific DA algorithm, but rather define it more generally for now.

STATE ESTIMATION

With the exception of a minority of DA schemes such as Optimal Interpolation (Cressie et al., 1990), most DA methods used for state estimation, such as the Kalman Filter (Kalman, 1960) or the Ensemble Kalman Filter (Evensen, 1994) are derived from Bayes theorem:

$$p(X \mid d) \propto p(X)\, p(d \mid X) \tag{1.1}$$

where,

$p(X)$ is the probability distribution (p.d.f.) of X that represents our prior knowledge or beliefs about the state vector X, i.e. the p.d.f. whose mode would normally represent our best guess of what X should be (say from the model), and where some measure of the spread of the distribution (e.g. the standard deviation) would normally represent the uncertainty on that estimate.

$p(d \mid X)$ is p.d.f of d given X, i.e. the p.d.f. which represents the knowledge gained from the observations including their uncertainty. It is commonly referred to as the likelihood function.

$p(X \mid d)$ is the p.d.f. of the state vector X given that some data d has been observed. This is what we are trying to determine. The mode of this p.d.f. represents the location of the state space with the highest probability and so it represents the optimal estimate of the state of the system given the information provided for by the observations and our prior guess.

Some DA methods used for state estimation assimilate one observation at a time. In other words, Bayes' theorem above is used every time an observation becomes available. These methods, which include the Kalman Filter (Kalman, 1960), the Ensemble Kalman Filter (Evensen, 1994), and the Particle Filter (Van Leeuwen, 2010), are referred to as sequential (Williams et al., 2010). In essense at time t=0 the model is run forward, then

when an observation becomes available at a future time point, the prior p.d.f. and likelihood function from equation 1.1 are determined and $p(X \mid d)$ is determined (up to proportionality) either analytically if this is possible (e.g. Kalman Filter) or it is approximated using DA schemes such as the Ensemble Kalman filter or the Particle filter. In the case of the Kalman Filter, the assimilation of the first observation results in a vertical shift in the trajectory of the elements from the state vector which are linked to the observation(s) being assimilated. The model then runs forward until a next observation appears and the process is repeated. This continues until all the observations have been assimilated. Throughout this process, the model parameters are fixed at specified values determined by independent means, such as from literature or by estimating the parameters using other data. A schematic of this is given in the left panel of figure 1.7.

DA can be used to estimate the states of a model by assimilating all of the observations in one go and are referred to as 'batch' methods for this reason (Williams et al., 2009). For state estimation, a DA scheme such as 4D-Var (Ziehn et al., 2012) is an example of this. It is the weighted sum of the model's estimate of the state of the system and what is observed. In multiple dimensions, it is written as follows:

$$J(\mathbf{x}_0) = J_b(\mathbf{x}_0) + J_o(\mathbf{x}_0) \tag{1.2}$$

where $J_b(\mathbf{x}_0) = (\mathbf{x}_0 - \mathbf{x}_b)^T \mathbf{B}^{-1} (\mathbf{x} - \mathbf{x}_b)$ and $J_o(\mathbf{x}_0) = \sum_{i=0}^{n} (\mathbf{y}_i - H_i[\mathbf{x}_i]) \mathbf{R}_i (\mathbf{y}_i - H_i[\mathbf{x}_i])$.

B represents the background error and **R** denotes the error covariance matrix corresponding to the observations and model. The time points when observations are present are denoted by the subscript *i*, and *n* is the number of data. H_i is the mapping which converts the model state to a something comparable with the observations. The optimal estimate of $\mathbf{x}_0$ is obtained by finding the global minimum of $J(\mathbf{x}_0)$. This is achieved by using a global descent algorithm (Pearson et al., 2008). The right panel of figure 1.7 shows this in a graphical sense. The blue line refers to the model trajectory with no influence from the observation, while the red line refers to the model trajectory after the J_b and J_o components of J from equation (1.2) have been minimised. As with sequential methods for state estimation, the parameters are fixed at pre-specified values if batch methods just as 4D-Var are used to estimate the states of the model.

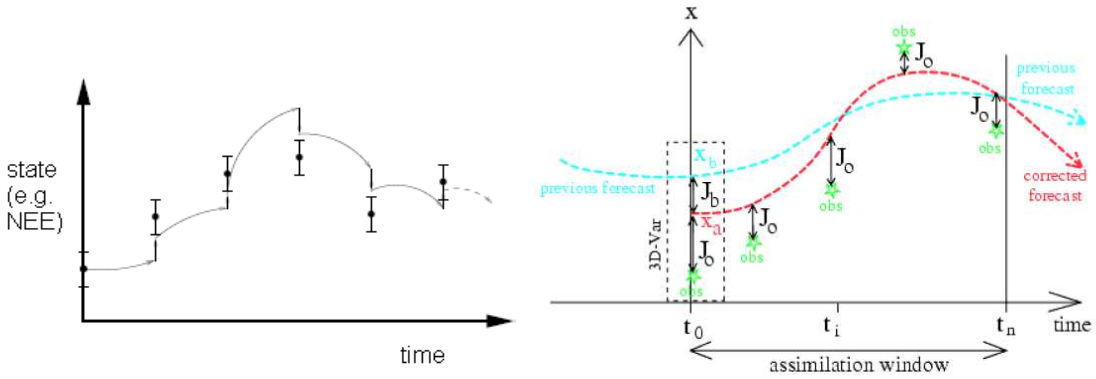


Figure 1.7 Left panel: schematic of the Kalman filter (Pearson, 2008); Right panel: A schematic of the 3D-Var and 4D-Var schemes (Bouttier & Courtier, 1999). The 3D-Var only uses observations at the same time point as the background error covariance matrix, whereas 4D-Var uses future observations.

PARAMETER ESTIMATION

For the majority of DA carried out on large models such as CASA, BETHY, JULES, this has been done to estimate the model parameters, e.g. Richardson et al. (2010), Knorr & Kattge (2005) and Evensen (2009). As with the state estimation problem, the DA schemes can be split into sequential or batch methods. For a sequential DA scheme such as the Ensemble Kalman filter, an approximation of the Kalman filter that that can operate with large and complex models, the parameters are estimated by adding the them to the state vector, which is then referred to as the augmented state vector (Evensen, 2009). This means that the parameters are not treated as being time-invariant, but change for each time-step. For situations where we need an estimate of the parameters, e.g. to make forecasts, the parameter set on the final time-step is used (Hill et al., 2012). These methods are still based on the version of Bayes' theorem as quoted in equation (1.1).

For batch DA schemes used for parameter estimation, an alternative version of Bayes' theorem is used as the basis.

$$p(\theta \mid d) \propto p(\theta)\, p(d \mid \theta) \tag{1.3}$$

This is the same as equation (1.1) except that instead of *X*, used to denote the state vector, θ is used to represent the vector of parameters. Thus, $p(\theta)$ represents the p.d.f. of the prior estimates (or our best guesses) of the parameters – these could be from literature or a previous study. Then, $p(d \mid \theta)$ represents the likelihood function, and finally $p(\theta \mid d)$ is

the p.d.f. of the posterior estimates of the parameters. The Metropolis algorithm and 4D-Var are examples of DA schemes used to estimate parameters in this way. For 4D-Var, the procedure is similar to the state estimation problem except that the parameters are augmented to the $\mathbf{x}_0$ vector from equation 1.2 (Ziehn et al., 2012).

Most of the DA studies carried out on global terrestrial carbon models have been used for parameter estimation, and so we now briefly review some of these studies. MacBean et al. (2010) assimilated soil water data derived from satellite sensor measurements to estimate the 13 parameters of an adapted version of the CASA model in order to predict CH4 fluxes from a peatland site. The purpose of the study was to determine at what level of accuracy the satellite data would need to be and at what frequency in order for the data to be able to constrain the parameters. To address these questions, synthetic data was used. The authors found that even with infrequent measurements with a high uncertainty, the true value of the hydrology related parameters could be found by assimilating synthetic soil moisture data. An extension to this study could be to test the effect of the length of the dataset being assimilated on the calibrated parameters.

An interesting study by Ziehn et al. (2012) estimated 19 parameters of the BETHY model using a 4D-VAR approach and the Metropolis algorithm. For most parameters, the posterior probability distribution functions (pdfs) of the parameters were very similar under both methods. The pdfs of the global NEP annual rate for the 1980s and 1990s were also very similar under both schemes. However, the 4D-VAR scheme was much more computationally efficient than the Metropolis algorithm

Gao et al. (2011) used the Ensemble Kalman filter to assimilate eight streams of data from the Duke Forest between 1996 and 2004 into the Terrestrial Ecosystem (TECO) model, in order to estimate the model parameters. The EnKF was able to constrain all but two parameters, with these two parameters responsible for the slow C processes such as the turnover rate of soil C. As a result of the EnKF only requiring one model iteration, it is very fast to run compared to other DA schemes. However, one question which remains unclear is whether we would get similar parameter estimates if a batch type DA scheme, such as the Metropolis algorithm, was implemented.

In the next section, we outline the research questions that arise from the current knowledge on using DA with the corresponding models.

1.4 Research Priorities and brief outline of thesis

In this chapter, we have described the terrestrial C cycle and outlined the case how key components of it, such as GPP, NPP, NEE, can only be globally estimated using models. However, we have also made the case that these models need to be better constrained using various types of observations, which is necessary to validate the adequacy of the model structure and because these models require better estimates of their parameters. We have also described Data Assimilation, and its potential to not only improve the estimates of the quantities of interest, but also allow us to better quantify the uncertainty in those estimates than models or observations on their own.

Although the theoretical principles behind DA have been around for a while, e.g. the Metropolis algorithm was first formulated in the 1950s (Metropolis et al., 1953), the practical implementations of DA have been more recent. The implementation of DA into terrestrial C models is even more recent, mainly of the past 10 to 15 years, and while this has been happening, there has been less attention to know how to apply DA to different situations. For example there is currently a whole array of different DA methods, but are certain ones better ones for certain situations? Williams et al. (2009) gave a broad but extensive review of how DA has been used with ground based observations. The authors also outlined various challenges that remained for the DA research community using terrestrial ecosystem models, which included:

(1) To explore the effect of assimilating 10+ years of data.
(2) To avoid confounding effects on missing processes in model representation on parameter estimation.
(3) To assimilate more types of data (e.g. pools/stocks of carbon, earch observation data) and to define improved observation operators.
(4) To fully quantify uncertainties arising from estimates of initial conditions, data bias and model structure.

Some of these challenges were one of the main sources of motivation for part of the research done in this PhD, and are discussed to some extent in chapter 2 (subsection 2.3.2). Luo et al. (2011) also stated that not many DA studies have explicitly addressed issues directly pertinent to forecasting, such as forecasting accuracy, sources of uncertainty and usefulness of forecasting under different domains.

In chapter two, greater detail is given on the gaps in our knowledge as well current knowledge of the terrestrial C cycle. These then form the basis of the specific research questions which are addressed in this thesis, and are set out at the end of chapter two. Chapter two also contains a more in depth critical review of the data and models, and an argument for the choice of model used in this thesis.

Chapters three, four and five then address the specific research questions as set out at the end of chapter two. Finally, chapter six concludes by bringing all the findings together and how the results contribute the bigger picture of improving our understanding and quantification of the terrestrial C cycle.

Chapter 2

Methods and Objectives of thesis

2.1 Introduction

In this chapter we provide a technical description of the Data Assimilation (DA) algorithms, the model and the data used to carry out the research in this thesis (chapters 3 – 5). For the DA algorithms, we also provide an in-depth critical review of the most commonly used ones. We then review the literature on where DA has been used not only in terrestrial C modelling but also in other fields. This then forms the basis for the choice of DA algorithms to use in thesis, and helps direct more clearly the broad aims of this thesis and the specific research questions that are to be addressed. Determining what DA algorithms to use will then inform the choice of model to use in this thesis. A description of the techniques and methods used to obtain data used in the thesis is also given. A critical review is given for the ground based measurements but for the satellite inferred data, this is left to the introduction section of chapter five, the only chapter to use satellite data. We now outline the structure of the thesis:

Section 2.2 begins by describing and in some cases derives the DA algorithms considered for this thesis, with a critical review. These include: sequential methods (subsections 2.2.2 to 2.2.7) which includes the Kalman Filter, the Ensemble Kalman Filter, the Ensemble Kalman Smoother, the Particle Filter and Optimal Interpolation; global search algorithms (subsections 2.2.8 to 2.2.9) which include the Metropolis algorithm and the Genetic algorithm; and Gradient Descent algorithms (subsection 2.2.10) which include the 3D-VAR and 4D-VAR schemes.

Section 2.3 moves onto discussing the current knowledge on DA in terrestrial C modelling and other types of modelling. This is then used to identify the gaps in our current knowledge of DA and a broad overview of the sorts of research questions which are to be addressed in this thesis. This will then form the basis (in subsection 2.3.2) for the choice of DA algorithms to be used in this thesis. For the remainder of section 2.3 (subsections 2.3.3 and 2.3.4) the DA algorithms to be used will be tested to ensure that there are no implementation problems. This is mainly for the benefit of chapter four, where the algorithms are compared when estimating the parameters of a model. Many of the DA algorithms also can make slightly different assumptions and so it is also important to identify the implications of these assumptions on parameter estimates. Assuming that there are no implementation issues, we also discuss the expectations of the DA algorithms when estimating parameters.

Sections 2.4 and 2.5 then describe and critically review the model and data used in this thesis. A decision about the choice of model is also made and this is based partly on the DA algorithms to be used and the type of research questions being addressed.

Although the broad research questions will have been addressed earlier in this chapter and at the end of chapter one, in section 2.6 we set out the specific research questions that are considered in this thesis. We discuss the reasons for choosing the particular questions, and we also give an overview of how they will be answered in the remainder of the thesis.

2.2 Principles and Options for Data Assimilation

2.2.1 The Dynamic Linear Model

A dynamic linear model is a linear model which is dependent on time. It is made up of the evolution equation and the observation equation, given below. Equation (2.1a) describes the evolution of the system as described by the model, whereas (2.1b) converts between the model states and the observations. The RHS of both have noise terms because the model is not perfect, nor is the conversion between model and observation space.

The Evolution equation: $\mathbf{x}_{k+1}^t = \mathbf{M}\,\mathbf{x}_k^t + \mathbf{w}_k, \quad \mathbf{w}_k \sim N_{n_x}(\mathbf{0}, \mathbf{Q}_k)$ (2.1a)

The Observation equation: $\mathbf{y}_k = \mathbf{H}\mathbf{x}_k^t + \mathbf{v}_k, \quad \mathbf{v}_k \sim N_{n_y}(\mathbf{0}, \mathbf{R})$ (2.1b)

where $\mathbf{x}_k^t$ is an $n_x \times 1$ dimension vector which is the true state (denoted by the superscript 't') of the system at time step *k*;

$\mathbf{M}$ is the evolution operator represented by a $(n_x \times n_x)$ dimensional matrix. It is based on the science that describes the system, for example a sequence of differential equations. When the model is non-linear, the matrix **M** becomes a function, i.e. the evolution equation would be written as: $\mathbf{x}_{k+1}^t = \mathbf{M}(\mathbf{x}_k^t) + \mathbf{w}_k$;

$\mathbf{Q}_k$ is the $(n_x \times n_x)$ dimensional covariance matrix of the model error (also called representation error);

$\mathbf{y}_k$ is an $(n_y \times 1)$ dimension vector, which contains all the observations of the system at time step *k*;

$\mathbf{H}$ is an $(n_y \times n_x)$ dimensional matrix known as the observation operator, which converts the state of the system at time step *k* to a form which may be compared to the observations at that time step;

$\mathbf{w}_k$ is an $n_x \times 1$ dimension vector containing model noise, which is added at each timestep. It is specified by a zero-mean multivariate normal distribution, with error covariance matrix $\mathbf{Q}_k$;

$\mathbf{v}_k$ is an $n_x \times 1$ dimension vector containing observation noise, which is added at each timestep. It is specified by a zero-mean multivariate normal distribution, with error covariance matrix $\mathbf{R}$.

$\mathbf{R}$ is the $(n_y \times n_y)$ dimensional covariance matrix of the observation or measurement errors and is defined as $\mathbf{R} = E(\mathbf{v}_k \mathbf{v}_k^T)$;

The evolution equation describes how the system evolves through time, while the observation equation describes the relationship between the observations and the state vector.

2.2.2 An Introduction to DA by deriving the 1D Kalman Filter

One of the earliest DA schemes to be used is the Kalman Filter (Kalman, 1960). Its most celebrated application was for the navigation system on board the Apollo Space spacecrafts. It is a sequential DA technique, which means that observations are only assimilated into the model one time-step at a time. Although it is not used in this thesis (for reasons explained later), we derive it because the Ensemble Kalman filter (EnKF), which is used in this thesis, is an approximation of the Kalman filter (KF). For ease of explanation, we start by deriving the KF in the one dimensional case. In this case, $n_x = n_y = 1$ from subsection 2.2.1. In other words the vectors $\mathbf{x}_k^t$, $\mathbf{y}_k$, $\mathbf{w}_k$, $\mathbf{v}_k$, $\mathbf{Q}_k$ and $\mathbf{R}$ are all scalars.

To aid the explanation, we use a simple example. Assume that we wish to know the value of an unknown temperature, denoted by *T*. Suppose that we have an observation of *T* (written T_o) and a best guess or forecast of what *T* might be (e.g. from a model), which we denote as T_f. We represent the observational and model errors by a Normal distribution. Thus:

$$T_o = T^{(t)} + e_o \tag{2.2a}$$

$$T_f = T^{(t)} + e_f \tag{2.2b}$$

where $T^{(t)}$ is the true value of *T*, with e_o and e_f representing the Normally distributed errors in T_o and T_f, where $e_o \sim N(0, s_o^2)$ and $e_f \sim N(0, s_f^2)$. The variances are given by:

$$s_o^2 = E(e_o^2) - [E(e_o)]^2 = E(e_o^2) \tag{2.3a}$$

$$\text{and } s_f^2 = E(e_f^2) - [E(e_f)]^2 = E(e_f^2) \tag{2.3b}$$

We also assume that T_o and T_f are independent, i.e.

$$\text{Cov}(e_o, e_f) = 0 \quad \Rightarrow \quad E(e_o e_f) = 0 \tag{2.3c}$$

The best linear unbiased estimate of *T* is a linear combination of T_o and T_f, denoted as T_a, subject to T_a being an unbiased estimator of *T* and the variance of T_a (denoted s_a^2) being minimised, i.e.

$$T_a = aT_o + bT_f \tag{2.4a}$$

where *a* and *b* are constants to be found subject to E(T_a) = $T^{(t)}$ and s_a^2 being minimised.

(i) If T_a to be an unbiased estimator of T or E(T_a) = $T^{(t)}$, then

$$\mathrm{E}\left(T_a - T^{(t)}\right)=0$$

$$\Rightarrow \quad \mathrm{E}\left(aT_o + bT_f - T^{(t)}\right)=0 \qquad \text{(using equation (2.4a))}$$

$$\Rightarrow \quad \mathrm{E}\left(a(T^{(t)} + e_o) + b(T^{(t)} + e_f) - T^{(t)}\right)=0 \qquad \text{(using equations (2.2a) \& (2.2b))}$$

$$\Rightarrow \quad \mathrm{E}\left(ae_o + be_f + (a+b-1)T^{(t)}\right)=0$$

$$\Rightarrow \quad \mathrm{E}\left(ae_o\right) + \mathrm{E}\left(be_f\right) + \mathrm{E}\left((a+b-1)T^{(t)}\right)=0$$

$$\Rightarrow \quad 0 + 0 + (a+b-1)\,\mathrm{E}\left(T^{(t)}\right)=0$$

$\mathrm{E}\left(T^{(t)}\right)\neq 0$, thus $a+b-1=0$, which implies that $b=1-a$

(ii) The variance of e_a is minimised.

The variance of e_a (s_a^2) is E(e_a^2) – [E(e_a)]² = E(e_a^2) (since E(e_a) = 0). Thus:

$$\mathrm{E}(e_a^2) = \mathrm{E}\left\{(T_a - T^{(t)})^2\right\}$$

$$= \mathrm{E}\left\{(aT_o + bT_f - T^{(t)})^2\right\} \qquad \text{(using equation 2.4a)}$$

$$= \mathrm{E}\left\{(a(T^{(t)} + e_o) + b(T^{(t)} + e_f) - T^{(t)})^2\right\}$$

$$= \mathrm{E}\left\{(a(T^{(t)} + e_o) + b(T^{(t)} + e_f) - T^{(t)})^2\right\} \qquad \text{(using equations 2.2a \& 2.2b)}$$

$$= \mathrm{E}\left\{(ae_o + be_f + (a+b-1)T^{(t)})^2\right\}$$

$$= \mathrm{E}\left\{(ae_o + be_f)^2\right\} \qquad \text{(since } a+b-1=0\text{)}$$

$$= \mathrm{E}\left(a^2e_o^2 + b^2e_f^2 + abe_oe_f\right)$$

$$= a^2\,\mathrm{E}(e_o^2) + b^2\,\mathrm{E}\left(e_f^2\right) + ab\,\mathrm{E}(e_o e_f) \qquad \text{(using equations 2.3a, 2.3b \& 2.3c)}$$

$$= a^2 s_o^2 + b^2 s_f^2$$

$$= a^2 s_o^2 + (1-a)^2 s_f^2 \qquad \text{(since } b=1-a\text{)}$$

i.e. $\mathrm{E}(e_a^2) = a^2 s_o^2 + (1-a)^2 s_f^2$

$\mathrm{E}(e_a^2)$ is minimised when $\frac{d}{da}\mathrm{E}(e_a^2)=0 \Rightarrow 2as_o^2 - 2(1-a)s_f^2=0$. This implies:

$$a = \frac{s_f^2}{s_o^2 + s_f^2} \quad \text{and} \quad b = \frac{s_o^2}{s_o^2 + s_f^2}$$

Thus, equation (2.4a) becomes:

$$T_a = \frac{s_o^2}{s_o^2 + s_f^2} T_f + \frac{s_f^2}{s_o^2 + s_f^2} T_o \tag{2.4b}$$

$$\Rightarrow \quad T_a = \frac{s_o^2 T_f + s_f^2 T_o}{s_o^2 + s_f^2}$$

$$\Rightarrow \quad T_a = \frac{\left(s_o^2 + s_f^2\right) T_f + s_f^2 T_o - s_f^2 T_f}{s_o^2 + s_f^2}$$

$$\Rightarrow \quad T_a = T_f + \frac{s_f^2}{s_o^2 + s_f^2}(T_o - T_f)$$

OR $$\boxed{T_a = T_f + K\,(T_o - T_f)} \tag{2.5}$$

where $K = \frac{s_f^2}{s_o^2 + s_f^2}$ is referred to as the Kalman gain. Using equation (2.5), s_a^2 is given by:

$$s_a^2 = \mathrm{E}(e_a^2) = a^2 s_o^2 + (1-a)^2 s_f^2$$

$$= \left(\frac{s_f^2}{s_o^2 + s_f^2}\right)^2 s_o^2 + \left(\frac{s_o^2}{s_o^2 + s_f^2}\right)^2 s_f^2$$

$$= \frac{s_f^4 s_o^2 + s_o^4 s_f^2}{\left(s_o^2 + s_f^2\right)^2}$$

$$= \frac{s_f^2 s_o^2 (s_f^2 + s_o^2)}{\left(s_o^2 + s_f^2\right)^2}$$

$$\Rightarrow \quad \boxed{s_a^2 = \frac{s_f^2 s_o^2}{s_o^2 + s_f^2}} \tag{2.6}$$

Since $\frac{s_f^2}{\left(s_o^2+s_f^2\right)}$<1 and $\frac{s_o^2}{\left(s_o^2+s_f^2\right)}$<1, s_a^2 will always be less than s_o^2 and s_f^2.

DERIVING THE 1D KALMAN FILTER USING BAYES' THEOREM

The formulae for T_a and s_a^2 (equations 2.5 and 2.6) can also be derived using Bayes

theorem (Bayes, 1753). This theorem is central to most DA schemes. Rather than representing the forecast and observed estimates of T in terms of means and standard deviations, we represent them in terms of their probability distribution functions. Bayes theorem states:

$$p(T \mid d) \propto p(T)\, p(d \mid T) \tag{2.7}$$

where

$p(T \mid d)$ is the probability of T given we have some data d. This is what we are trying to determine.

$p(T)$ is the prior probability of T, i.e. the p.d.f. whose mode would normally represent our best guess of what T should be (say from the model), and where some measure of the spread of the distribution (e.g. the standard deviation) would normally represent the uncertainty on that estimate.

$p(d \mid T)$ is probability of d given T, i.e. the p.d.f. which represents the observational uncertainty. It is commonly referred to as the likelihood function.

To demonstrate Bayes theorem in practice, it is now applied to the Kalman filter example at the start of this subsection to derive the formulae for T_a and s_a^2. Since T_f is our forecast (i.e. our best guess) of T and since the error on our guess (i.e. e_f) follows a N(0, s_f^2) distribution (equation (2.1a), this is equivalent to saying that T is Normally distributed with mean given by T_f and variance s_f^2. In a similar way, the data value (T_o) can be thought of representing the data's estimate of T. Since the error on the data (i.e. e_o) follows a N(0, s_o^2) distribution (equation (2.2a), this is equivalent to saying that the data is Normally distributed with mean given by T_o and variance s_o^2. In other words,

$$p(d \mid T) = \frac{1}{s_o\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(T - T_o)^2}{s_o^2}\right)$$

$$p(T) = \frac{1}{s_f\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(T - T_f)^2}{s_f^2}\right)$$

Thus, using Bayes theorem:

$$p(T \mid d) \propto p(T)\, p(d \mid T)$$

$$= \frac{1}{s_f\sqrt{2\pi}} \times \frac{1}{s_o\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(T-T_f)^2}{s_f^2}\right)\exp\left(-\frac{1}{2}\frac{(T-T_o)^2}{s_o^2}\right)$$

$$p(T \mid d) \propto \exp\left(-\frac{1}{2}\left(\frac{(T-T_f)^2}{s_f^2}+\frac{(T-T_o)^2}{s_o^2}\right)\right) \tag{2.8}$$

Applying the natural logarithm to both sides of equation (2.8), to make the mathematics neater, gives:

$$\ln\left(p(T \mid d)\right) \propto -\frac{1}{2}\left(\frac{(T-T_f)^2}{s_f^2}+\frac{(T-T_o)^2}{s_o^2}\right) \tag{2.9}$$

Expanding the inner brackets, collecting like terms and dividing the top and bottom by $\left(s_f^2+s_o^2\right)$ gives:

$$\ln\left(p(T \mid d)\right) \propto -\frac{1}{2}\left(\frac{T^2 - 2\frac{s_o^2T_f+s_f^2T_o}{s_f^2+s_o^2}T + \frac{s_o^2T_f^2+s_f^2T_o^2}{s_f^2+s_o^2}}{s_f^2s_o^2 \Big/ s_f^2+s_o^2}\right)$$

$$= -\frac{1}{2}\left(\frac{\left(T-\frac{s_o^2T_f+s_f^2T_o}{s_f^2+s_o^2}\right)^2 + A}{s_f^2s_o^2 \Big/ s_f^2+s_o^2}\right), \text{ where } A = \frac{s_o^2T_f^2+s_f^2T_o^2}{s_f^2+s_o^2} - \left(\frac{s_o^2T_f+s_f^2T_o}{s_f^2+s_o^2}\right)^2$$

$$\propto -\frac{1}{2}\left(\frac{\left(T-\frac{s_o^2T_f+s_f^2T_o}{s_f^2+s_o^2}\right)^2}{s_f^2s_o^2 \Big/ s_f^2+s_o^2}\right)$$

i.e. $$\ln\left(p(T \mid d)\right) \propto -\frac{1}{2}\left(\frac{\left(T-\frac{s_o^2T_f+s_f^2T_o}{s_f^2+s_o^2}\right)^2}{s_f^2s_o^2 \Big/ s_f^2+s_o^2}\right) \tag{2.10}$$

Thus up to proportionality, $T \mid d$ has normal distribution with mean equal to $T_a = \frac{s_o^2T_f+s_f^2T_o}{s_o^2+s_f^2}$ which is the same as equation (2.4b), and variance given by $s_a^2 = \frac{s_f^2s_o^2}{s_o^2+s_f^2}$ which is equation (2.6).

2.2.3 Deriving the Kalman Filter in two or more dimensions

We now derive the KF in multiple dimensions, i.e. where n_x>1 and n_y>1 from subsection 2.2.1. We first define $\hat{\mathbf{x}}_k^-$, which represents our estimated state of the system before assimilating the observations. This is the prior estimate of the state. Once we have assimilated our observations we update the estimate, and the posterior estimate of the state is written as $\hat{\mathbf{x}}_k$. The most traditional Bayesian way to derive the posterior (up to proportionality) would be to multiply the pdf (probability density function) of the prior distribution with that of the data likelihood [ie, $p(\hat{\mathbf{x}}_k) = p(\hat{\mathbf{x}}_k^- | \mathbf{y}) \propto p(\mathbf{y}|\hat{\mathbf{x}}_k^-)\ p(\hat{\mathbf{x}}_k^-)$], which is what Mandel (2007) does. However, this is quite complicated. A clearer way of deriving the posterior is if we express it first as a weighted sum of the prior estimate and the observations:

$$\hat{\mathbf{x}}_k = (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\hat{\mathbf{x}}_k^- + \mathbf{K}_k\mathbf{y}_k, \tag{2.11}$$

where $\mathbf{K}_k$ is known as the Kalman gain, which contains information about the error covariance matrices of the model state and the observations ($\mathbf{Q}_k$ and $\mathbf{R}_k$). $\mathbf{H}_k$ converts the prior state ($\hat{\mathbf{x}}_k^-$) into a form comparable with the observations ($\mathbf{y}_k$). Essentially, equation (2.11) is the matrix equivalent of equation (2.4a) where b=1–a. Therefore one way of thinking about it is that, in one dimensional terms, if $\mathbf{K}_k$ is large we give more weighting to the observation (i.e. our posterior estimate of the state is closer to the observation). However, if $\mathbf{K}_k$ is small, we give more weighting to the prior estimate (ie our posterior estimate is closer to the prior estimate). After expanding the bracket and factorising in terms of $\mathbf{K}_k$ we get:

$$\hat{\mathbf{x}}_k = \hat{\mathbf{x}}_k^- + \mathbf{K}_k(\mathbf{y}_k - \mathbf{H}_k\hat{\mathbf{x}}_k^-) \tag{2.12}$$

This is a more commonly used form (equivalent to equation (2.5) in the 1D case), although it is sometimes also written as:

$$\hat{\mathbf{x}}_k = \hat{\mathbf{x}}_k^- + \mathbf{K}_k\mathbf{d}_k \tag{2.13a}$$

$$\mathbf{d}_k = (\mathbf{y}_k - \mathbf{H}_k\hat{\mathbf{x}}_k^-) \tag{2.13b}$$

The above two formulae are called the *update equations.* However, we have still to derive $\mathbf{K}_k$. To do this, we first define a few things. Let $\mathbf{P}_k^-$ and $\mathbf{P}_k$ be the covariance matrices for our prior and posterior estimates of the state ($\hat{\mathbf{x}}_k^-$ and $\hat{\mathbf{x}}_k$) respectively (the multivariate equivalent of s_o^2 and s_f^2 from the 1D KF) . In other words,

$\mathbf{P}_k^- = \mathrm{E}[(\hat{\mathbf{x}}_k^- - \mathbf{x}_k^t)(\hat{\mathbf{x}}_k^- - \mathbf{x}_k^t)^T]$ and $\mathbf{P}_k = \mathrm{E}[(\hat{\mathbf{x}}_k - \mathbf{x}_k^t)(\hat{\mathbf{x}}_k - \mathbf{x}_k^t)^T]$, where $\mathbf{x}_k^t$ is the true state of the system at time *k*. It can be shown that:

$$\mathbf{P}_k^- = \mathbf{M}_{k-1}\mathbf{P}_{k-1}\mathbf{M}_{k-1}^T + \mathbf{Q}_{k-1} \quad \text{and} \quad \mathbf{P}_k = (\mathrm{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{P}_k^-,$$

where **I** is the identity matrix - see appendix A for the proofs. Note that $\mathbf{P}_k$ involves $\mathbf{K}_k$, therefore if we use information in $\mathbf{P}_k$ to optimise $\mathbf{K}_k$, this will be the best course of action. One way of measuring how good the posterior estimate of the states are is to sum the variances of their errors. This is the same thing as summing the elements in the leading diagonal of $\mathbf{P}_k$. Therefore, since we want to optimise $\mathbf{K}_k$, we choose $\mathbf{K}_k$ such that the trace of $\mathbf{P}_k$ is minimised. After some algebraic manipulation, we deduce that:

$$\mathbf{K}_k = \mathbf{P}_k^-\mathbf{H}_k^T(\mathbf{H}_k\mathbf{P}_k^-\mathbf{H}_k^T + \mathbf{R}_k)^{-1}$$

The proof of this can be found in Appendix A. Explanation of how the Kalman filter works is almost complete. A final point needs to be made about the initial state, $\hat{\mathbf{x}}_0$. Before anything can happen, this must be estimated along with its corresponding error covariance matrix $\mathbf{P}_0$. This can be done via expert knowledge, related observational information, or past experiments. More recently some studies, for example Braswell et al. (2005) or Fox et al. (2009), have treated $\hat{\mathbf{x}}_0$ as parameters to be estimated. Note that $\mathbf{P}_0$ here is the same as the background error covariance matrix, **B**, from the variational data assimilation methods (see subsection 2.2.10).

To summarise, we outline the step by step process for implementation of the Kalman filter:

(1) Recall first of all the evolution and observation equations:

a. $\mathbf{x}_{k+1}^t = \mathbf{M}_k\mathbf{x}_k^t + \mathbf{w}_k$, $\quad \mathbf{w}_k \sim N_{n_x}(\mathbf{0}, \mathbf{Q}_k)$

b. $\mathbf{y}_k = \mathbf{H}_k\mathbf{x}_k^t + \mathbf{v}_k$, $\quad \mathbf{v}_k \sim N_{n_{yk}}(\mathbf{0}, \mathbf{R}_k)$

(2) Start with $\hat{\mathbf{x}}_0$ and $\mathbf{P}_0$. (the initial conditions)

(3) Do *k* = 1→ N (i.e. we do a loop from day *k=1* to *k=N*)

(4) Calculate $\hat{\mathbf{x}}_k^- = \mathbf{M}_{k-1}\hat{\mathbf{x}}_{k-1}$ (prior estimate of state)

(5) Calculate $\mathbf{P}_k^- = \mathbf{M}_{k-1}\mathbf{P}_{k-1}\mathbf{M}_{k-1}^T + \mathbf{Q}_{k-1}$ (error covariance matrix of prior estimate)

(6) if {observations, $\mathbf{y}_k$ exist at time k} then

(7) calculate $\mathbf{K}_k = \mathbf{P}_k^-\mathbf{H}_k^T(\mathbf{H}_k\mathbf{P}_k^-\mathbf{H}_k^T + \mathbf{R}_k)^{-1}$ (Kalman gain matrix)

(8) calculate $\hat{\mathbf{x}}_k = \hat{\mathbf{x}}_k^- + \mathbf{K}_k(\mathbf{y}_k - \mathbf{H}_k\hat{\mathbf{x}}_k^-)$ (posterior estimate of state)

(9) calculate $\mathbf{P}_k = \mathbf{S}_k\mathbf{P}_k^-\mathbf{S}_k^T + \mathbf{K}_k\mathbf{R}_k^-\mathbf{K}_k^T$ (error covariance matrix of posterior estimate)

(10) else

(11) $\hat{\mathbf{x}}_k = \hat{\mathbf{x}}_k^-$

(12) $\hat{\mathbf{P}}_k = \hat{\mathbf{P}}_k^-$

(13) end do

Note that:

(1) In many descriptions of the KF, a subscript is not used for the evolution matrix ($\mathbf{M}_k$) since $\mathbf{M}_0 = \mathbf{M}_1 = \mathbf{M}_2 = ... = \mathbf{M}_k = ... = \mathbf{M}_n$, so we just use $\mathbf{M}$ instead. For similar reasons, subscripts are often not needed for $\mathbf{H}_k$, $\mathbf{Q}_k$ and $\mathbf{R}_k$, which are just replaced by **H**, **Q**, and **R** respectively.

(2) The prior estimate of the state is sometimes called the forecast, with a superscript f being used. Also, the posterior estimate of the state is sometimes called the analysis, with a superscript *a* used.

(3) If the KF is used to give a sequential estimate to the standard Least Squares Regression problem, then the solution is not at the global optimal (i.e. for all the point collectively) until the end of the time-series, i.e. until all the observations have been assimilated. It is important to note though that the solution is optimal at each point where an observation is present, since at each point the matrix $\mathbf{K}_k$ is chosen such that the trace of the matrix $\mathbf{P}_k$ is minimised (see top of page 56).

LIMITATIONS OF THE KF

The Kalman Filter (KF) has been successfully applied in many real world problems, most notably as part of the Apollo space missions (Mackenzie, 2003). However, it has two major drawbacks:

(a) When the model is non-linear, a matrix cannot be used to model the evolution of the system, so the evolution equation becomes, $\mathbf{x}_{k+1} = \mathbf{M}(\mathbf{x}_k)$, where $\mathbf{M}(\cdot)$ is an operator. This means our prior estimate of the state at time *k+1*, i.e. $\mathbf{x}_{k+1}$ is not Normal. However, one of the assumptions of the Kalman filter is that the prior and likelihood are Normally distributed. In more practical terms, the formulae for $\mathbf{P}_k^f$ and $\mathbf{K}$ incorporate a matrix $\mathbf{M}$, thus these cannot be updated when $\mathbf{M}$ is non-linear. This can be overcome by using the Extended Kalman Filter (Evensen, 2007) which operates in exactly the same way as the Kalman Filter except that the non-linear model operator $\mathbf{M}(\cdot)$ is linearised, allowing the evolution of the modeled system to be put into matrix form. However, there are problems with the Extended Kalman Filter (see Evensen (2007) for details) so that in practice it is not greatly used.

(b) When the dimension of the state vector is very large (e.g. of the order of millions of elements as in weather forecasting), the dimension of the error covariance matrix, $\mathbf{P}_k^f$, becomes too large. Consequently, storing $\mathbf{P}_k^f$ and propagating it forward in time (i.e. using it in the formula $\mathbf{P}_k^f = \mathbf{M}\mathbf{P}_{k-1}^a\mathbf{M}^T + \mathbf{Q}_k$) becomes computationally very expensive.

2.2.4 The Ensemble Kalman Filter (EnKF)

USING THE EnKF TO ESTIMATE THE MODEL STATES

A widely used alternative to the Kalman Filter, which overcomes the problems of non-linearity and a large state vector, is the Ensemble Kalman Filter (EnKF), which was invented by Geir Evensen (Evensen, 1994). It is a sequential data assimilation algorithm which, like other DA algorithms, is based on Bayes' theorem (Bayes, 1763). It is an adaptation of the Kalman Filter method which uses stochastic methods based on parallel model runs (called ensembles) to estimate $\mathbf{x}_k^f$ and $\mathbf{P}_{k+1}^f$, with the error covariance matrices representing the covariances between the ensembles, rather than between the elements of the state vector. For terrestrial ecosystem models, the EnKF has been successfully used for state and parameter estimation (Williams et al., 2005, Quaife et al., 2007). The formulation of the EnKF is described in detail by Evensen

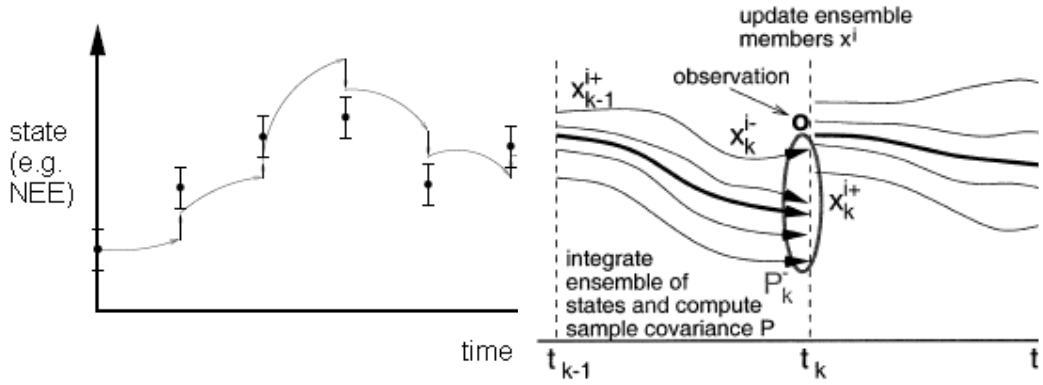


Figure 2.1 Schematic of the Kalman Filter (left panel) and the Ensemble Kalman Filter (right panel), taken from Pearson (2008) and Reichle (2001).

(2003, 2009). An outline of how it works is now given with a schematic of how it works given in the right panel of figure 2.1 with left panel corresponding to the Kalman filter.

STEP 1: Referring to the evolution and observation equations (equations (2.1a) and (2.1b)), we first need estimates of the matrices $\mathbf{Q}_k$ and $\mathbf{R}$, and the initial conditions, $\hat{\mathbf{x}}_0$ and $\mathbf{P}_0$. We define the ensemble state matrix (with dimension n × N) at time *k=0*:

$$\hat{\mathbf{X}}_0^a = \left[\hat{\mathbf{x}}_0 \mid \hat{\mathbf{x}}_0 \mid ... \mid \hat{\mathbf{x}}_0\right] + \left[\phi_1 \mid \phi_2 \mid ... \mid \phi_N\right], \quad \phi_i \sim N_n(\mathbf{0}, \mathbf{P}_0)\ \forall\, i$$

where *n* is the number of elements of the state vector and *N* is the number of ensembles.

STEP 2: At time point *k*, the n × N forecast ensemble matrix, $\hat{\mathbf{X}}_k^f$, is defined as:

$$\hat{\mathbf{X}}_k^f = \mathbf{M}(\hat{\mathbf{X}}_{k-1}^a) + \left[\tau_1 \mid \tau_2 \mid ... \mid \tau_N\right], \quad \tau_i \sim N_n(\mathbf{0}, \mathbf{Q}_k)\ \forall\, i \qquad (2.14)$$

STEP 3: The estimated forecast error covariance matrix, $\tilde{\mathbf{P}}_k^f$, is given by:

$$\tilde{\mathbf{P}}_k^f = \frac{1}{N-1}\left[\hat{\mathbf{X}}_k^f - \overline{\mathbf{X}_k^f}\right]\left[\hat{\mathbf{X}}_k^f - \overline{\mathbf{X}_k^f}\right]^T \qquad (2.15)$$

where $\overline{\mathbf{X}_k^f} = \left[\overline{\mathbf{x}_k^f} \mid \overline{\mathbf{x}_k^f} \mid \cdots \mid \overline{\mathbf{x}_k^f}\right]$, (2.16a)

$$\overline{\mathbf{x}_k^f} = \frac{1}{N}\sum_{j=1}^{N}\left(\hat{\mathbf{x}}_k^f\right)_j, \qquad (2.16b)$$

and $\left(\hat{\mathbf{x}}_k^f\right)_j$ is the *j*th column of the matrix $\hat{\mathbf{X}}_k^f$.

STEP 4: The analysis ensemble estimates of the state are contained in $\hat{\mathbf{X}}_k^a$ with corresponding error covariance matrix represented by $\tilde{\mathbf{P}}_k^a$. These two matrices are calculated using:

$$\hat{\mathbf{X}}_k^a = \hat{\mathbf{X}}_k^f + \tilde{\mathbf{K}}_k(\mathbf{Y}_k - \mathbf{H}\hat{\mathbf{X}}_k^f) \tag{2.17}$$

$$\tilde{\mathbf{P}}_k^a = \frac{1}{N-1}\left[\hat{\mathbf{X}}_k^a - \overline{\mathbf{X}_k^a}\right]\left[\hat{\mathbf{X}}_k^a - \overline{\mathbf{X}_k^a}\right]^T \tag{2.18}$$

where,

$\tilde{\mathbf{K}}_k$ is the Kalman Gain matrix, which contains the information about the uncertainty in the observations and the uncertainty of the model. It is given by $\tilde{\mathbf{K}}_k = \tilde{\mathbf{P}}_k^f \mathbf{H}^T (\mathbf{H}\tilde{\mathbf{P}}_k^f \mathbf{H}^T + \mathbf{R})^{-1}$. See appendix A for derivation.

$\mathbf{Y}_k$ is a matrix created by replicating the vector of observations by the number of ensembles and adding noise: $\mathbf{Y}_k = [\mathbf{y}_k \mid \mathbf{y}_k \mid ... \mid \mathbf{y}_k] + [\varepsilon_1 \mid \varepsilon_2 \mid ... \mid \varepsilon_N]$,

$\varepsilon_i \sim N_n(\mathbf{0}, \mathbf{R}) \; \forall \, i$.

H is the observation operator, a matrix which converts the model state to something comparable with the observations.

The tildas on $\tilde{\mathbf{P}}_k^f$, $\tilde{\mathbf{P}}_k^a$ and $\tilde{\mathbf{K}}_k$ indicate the Ensemble Kalman versions of the $\mathbf{P}_k^f$, $\mathbf{P}_k^a$ and $\mathbf{K}_k$ used in the original Kalman Filter. [Note that from Evensen (2009), an alternative formula for $\tilde{\mathbf{P}}_k^a$ can be used given by $\tilde{\mathbf{P}}_k^a = (\mathrm{I} - \tilde{\mathbf{K}}_k \mathbf{H})\tilde{\mathbf{P}}_k^f$].

STEP 5: Steps 2 to 4 are repeated for all time-steps, i.e. k=2, 3, … For time-steps where there are no observations, $\hat{\mathbf{X}}_k^a = \hat{\mathbf{X}}_k^f$ and $\mathbf{P}_k^a = \mathbf{P}_k^f$.

STEP 6: At any time *k*, the best estimate of the state of the system is given by $\overline{\mathbf{x}_k^a}$, the sample mean of the ensemble of the analysis estimates, given by:

$$\overline{\mathrm{x}_k^a} = \frac{1}{N}\sum_{j=1}^{N}\left(\hat{\mathrm{x}}_k^a\right)_j, \tag{2.19}$$

where $\left(\hat{\mathrm{x}}_k^a\right)_j$ is the *j*th column of the matrix $\hat{\mathrm{X}}_k^a$. This is because the distribution of the ensembles at the analysis step is assumed to be Normal. We check this assumption by plotting the ensembles at each time-step for each element of the state vector.

USING THE EnKF TO ESTIMATE THE MODEL PARAMETERS

The EnKF can be used to estimate parameters whilst simultaneously estimating the model's states. This is done by adding the parameters to the state vector, which is then referred to as the augmented state vector (Evensen, 2009). This means that the parameters are not treated as being time-invariant, but change for each time-step. The evolution of the parameter elements of the augmented state matrix is the identity matrix. In other words the forecast estimate of a parameter *p* at time *t* is the analysis estimate of *p* at the previous time step, *t–1*. In mathematical terms, we would represent this as: $p_t^{(f)} = p_{t-1}^{(a)}$ where the superscripts denote the forecast and analysis estimates of *p*. Evensen (2003) states that the parameter distributions, represented by the spread of the ensembles, on the final time-step should be used since the parameter estimates will be then conditional on all the observations, and will be as globally optimal as the Ensemble Kalman Smoother (Evensen & van Leeuwen, 2000). This is what is done in Fox et al. (2009) when the EnKF was used to estimate the parameters to the DALEC model.

Furthermore, when the EnKF is used to estimate model parameters, the model error in propagating the augmented state vector forward in time is often required to be very small. In Fox et al. (2009) and Spadavecchia et al. (2011) the model error was set to be less than 0.5% relative the augmented state vector. This was also done in chapter four for the EnKF parameteristaion set up. As with Fox et al. (2009) and Spadavecchia et al. (2011) this was done to limit the effect of the model error getting very large, so that the ensembles become very heavily stochastic, when there are large gaps between observations.

LIMITATIONS OF THE EnKF

A consequence of the EnKF, which is also true for the KF, is that the mass balance of the model is likely to be broken. In other words, for an individual ensemble member

(i.e. for one of the columns of $\hat{\mathbf{X}}_k^a$ in equation (2.17)), certain relationships between the model states (such as NEE = $R_a + R_{h1} + R_{h2} - GPP$) may not be true. This has no negative consequences in practical terms, but it is nonetheless limitation of the EnKF. Secondary filtering methods to retain the mass balance (e.g. Pan & Wood, 2006) can be implemented but this reduces the generality of the results and adds a whole extra level of complexity.

Since the EnKF assumes that the model noise is Gaussian, if the model is non-linear any Gaussian uncertainty will be transformed to another distribution once inside the model (equation (2.14)). However at the analysis step (equation (2.17)), this distribution will be forced to be Gaussian even if the model does not want it to be. For this reason, the Particle Filter (PF) is sometimes used instead of the EnKF for state estimation problems because any distribution can be used for the model noise and are allowed to persist at the analysis step, i.e. it is not forced to be a particular distribution. However for models such as DALEC and DALEC-D which are only slightly non-linear, this issue is unlikely to be a problem (Quaife, T., 2011, [personal communication]).

2.2.5 The Ensemble Kalman Smoother (EnKS)

The Ensemble Kalman Smoother (Evensen & van Leeuwen, 2000) is the more general case of the Ensemble Kalman filter where the estimate of the model state at time *t* is conditional not only on data at time *t* but at time points at all previously occurring observations. Moreover, the posterior distribution of the state vectors at the previous time points are updated.

In mathematical terms if $\mathbf{x}_t$ and $\mathbf{d}_t$ represent the modelled and observed state of the system at time *t,* then the Ensemble Kalman filter approximates the conditional probability distribution $p(\mathbf{x}_t|\mathbf{d}_t)$ using Bayes' theorem, i.e. $p(\mathbf{x}_t|\mathbf{d}_t) \propto p(\mathbf{d}_t|\mathbf{x}_t)p(\mathbf{x}_t)$ and representing the probability distributions by a distribution of parallel realizations or ensembles. In contrast, the Ensemble Kalman Smoother seeks to estimate the probability distribution, $p(\mathbf{x}_0,....,\mathbf{x}_{t-1},\mathbf{x}_t|\mathbf{d}_0,\ldots,\mathbf{d}_{t-1},\mathbf{d}_t)$. Using Bayes' theorem, this proportional to:

$$p(\mathbf{x}_0,....,\mathbf{x}_t|\mathbf{d}_0,\ldots,\mathbf{d}_t) \propto p(\mathbf{d}_0,\ldots,\mathbf{d}_t|\mathbf{x}_0,....,\mathbf{x}_t)p(\mathbf{x}_0,....,\mathbf{x}_t) \qquad (2.20)$$

Assuming that the model evolution is a first order Markov process (Evensen & van Leeuwen, 2000), i.e. $p(\mathbf{x}_t \mid \mathbf{x}_0, ..., \mathbf{x}_{t-1}) = p(\mathbf{x}_t \mid \mathbf{x}_{t-1})$, and that the data $\mathbf{d}_0, \ldots, \mathbf{d}_{t-1}$, $\mathbf{d}_t$ are independent, equation (2.20) can be expressed as:

$$p(\mathbf{x}_0, ..., \mathbf{x}_t \mid \mathbf{d}_0, \ldots, \mathbf{d}_t)$$

$$\propto p(\mathbf{x}_0) p(\mathbf{x}_1 \mid \mathbf{x}_0) p(\mathbf{d}_1 \mid \mathbf{x}_0) p(\mathbf{x}_2 \mid \mathbf{x}_1) p(\mathbf{d}_2 \mid \mathbf{x}_1) \ldots p(\mathbf{x}_t \mid \mathbf{x}_{t-1}) p(\mathbf{d}_{t-1} \mid \mathbf{x}_t)$$

As with the EnKF, this equation is evaluated sequentially. (Evensen & van Leeuwen, 2000). Beginning with the probability distribution of the initial state vector $p(\mathbf{x}_0)$, the EnKS can be computed sequentially using the following equations:

$$p(\mathbf{x}_0, \mathbf{x}_1 \mid \mathbf{d}_1) \propto p(\mathbf{x}_0) p(\mathbf{x}_1 \mid \mathbf{x}_0) p(\mathbf{d}_1 \mid \mathbf{x}_0) \tag{2.21}$$

$$p(\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2 \mid \mathbf{d}_1, \mathbf{d}_2) \propto p(\mathbf{x}_0, \mathbf{x}_1 \mid \mathbf{d}_1) p(\mathbf{x}_2 \mid \mathbf{x}_1) p(\mathbf{d}_2 \mid \mathbf{x}_1) \tag{2.22}$$

$$\vdots$$

$$p(\mathbf{x}_0, ..., \mathbf{x}_t \mid \mathbf{d}_0, \ldots, \mathbf{d}_t) \propto p(\mathbf{x}_0, ..., \mathbf{x}_{t-1} \mid \mathbf{d}_0, \ldots, \mathbf{d}_{t-1}) p(\mathbf{x}_t \mid \mathbf{x}_{t-1}) p(\mathbf{d}_{t-1} \mid \mathbf{x}_t) \tag{2.23}$$

Evensen & van Leeuwen (2000) state that at the final time point in the time window the EnKS operates over, the post-assimilation state vector is the same as if corresponding one computed by the EnKF. In this way, the parameter estimates obtained using the EnKF are globally optimal in the same way that the EnKS estimates are also. However, the authors also emphasise that at all other time points the EnKF and EnKS estimates will be different. In particular, the EnKF estimates at all time points except the end one are sub-optimal but they are optimal for the EnKS. The reason for this optimality is that for a time point t_s, where $0 < t_s < t$ and t is the current time window, the EnKS estimates depend on observations both forward (to the current time t) and backwards in time, whereas the EnKF estimates are only affected by those back in time and in particular by the previous time point only. In practical terms, this results in a smoothing effect on the trajectories of the state vector elements.

The main disadvantage to the EnKS compared to the EnKF is that it is more computationally demanding and more complicated to set up. This is especially true given that the parameter estimates produced by the EnKF are globally optimal and the same as those produced by the EnKS (Evensen & van Leeuwen, 2000). The EnKS is appropriate when we require these optimal estimates of the model states for all time points such as in Stroud et al. (2010). However, for large complex models which are computationally expensive to run, where we might be primarily interested in obtaining parameter estimates and uncertainties, the EnKF will be far more computationally

efficient than the EnKS. Even when we are only interested in state estimation, Evensen & van Leeuwen (2000) state that the computational cost may be so high that a trade off may be to apply a lagged EnKS, where model states are only updated for a subset of every spaced time point. When the model is simple, the EnKF may still be preferred – Evensen & van Leeuwen (2000) state that the EnKF is a first guess of the EnKS and this sense can be thought of as an approximation of EnKS. Therefore, if an approximation is all that is needed, then the EnKF may be preferred not only to overcome the additional computational cost, but also the extra time required to include the additional complexity in the programming code.

2.2.6 The Particle Filter (PF)

To derive the particle filter (PF), we start with Bayes' theorem:

$$p_m(\psi \mid d) = \frac{p_d(d \mid \psi)\, p_m(\psi)}{p_d(\psi)}$$

where ψ is the model state at a unspecified time point, and d is the data being assimilated. The ps refer to the pdfs of the model and data (subscripts m and d). Standard Bayesian statistics theory states that $p_d(\psi)$ can be expressed as $\int p_d(d \mid \psi)\, p_m(\psi).d\psi$, i.e.:

$$p_m(\psi \mid d) = \frac{p_d(d \mid \psi)\, p_m(\psi)}{\int p_d(d \mid \psi)\, p_m(\psi).d\psi} \tag{2.24}$$

Since $E[f(\psi)] = \int f(\psi)\, p(\psi).d\psi \approx \frac{1}{N}\sum_i f(\psi)$, equation (2.24) becomes:

$$p_m(\psi \mid d) \approx \frac{p_d(d \mid \psi)\, p_m(\psi)}{\frac{1}{N}\sum_i p_d(d \mid \psi)} \tag{2.25}$$

$p_m(\psi)$ is approximated by making N random draws from it, and these are referred to as particles. Van Leeuwen (2010) states that the model is represented by a sum of delta functions positioned at the model states chosen as particles:

$$p_m(\psi) = \frac{1}{N}\sum_{i=1}^{N} \delta(\psi - \psi_i)$$

Substituting this into equation (2.25) results in:

$$p_m(\psi \mid d) \approx \frac{p_d(d \mid \psi)\left(\frac{1}{N}\sum_{i=1}^{N}\delta(\psi - \psi_i)\right)}{\frac{1}{N}\sum_i p_d(d \mid \psi)}$$

Thus,

$$\boxed{p_m(\psi \mid d) = \sum_{i=1}^{N} \omega_i \delta(\psi - \psi_i)} \tag{2.26}$$

where

$$\omega_i = \frac{p_d(d \mid \psi_i)}{\sum_i p_d(d \mid \psi_i)} \tag{2.27}$$

is referred to as the weight of the *i*th particle. Although any probability can be used for $p_d(d \mid \psi_i)$, the pdf of the observations given the model state ψ_i, it is often treated as being Normal, i.e.

$$p_d(d \mid \psi_i) = A \exp\left[-\frac{\{d - H(\psi_i)\}^2}{2\sigma^2}\right]$$

where $H(\psi_i)$ is the observational operator and σ is the standard deviation of the observational error. Summarised in figure 2.2, the step by step process of the particle filter is then (van Leeuwen, 2009):

(1) Sample N particles ψ_i from the initial model probability density p($\psi^{(0)}$), in which the superscript 0 denotes the time index.

(2) Integrate all particles forward in time up to the measurement time. In probabilistic language we denote this as a sample from p($\psi^{(n)} \mid \psi_i^{(n-1)}$) for each *i* (in other words for each particle ψ_i run the model forward from time *n–1* to time *n* using the non-linear model equations). We then add some random model noise to $\psi^{(n)}$.

(3) Calculate the weights according to (2.27) and attach these weights to each corresponding particle (i.e. equation (2.26)). Note that the particles are not modified only their relative weight is changed.

(4) Increase n by 1 and repeat steps 2 and 3 until all observations up to the present have been processed.

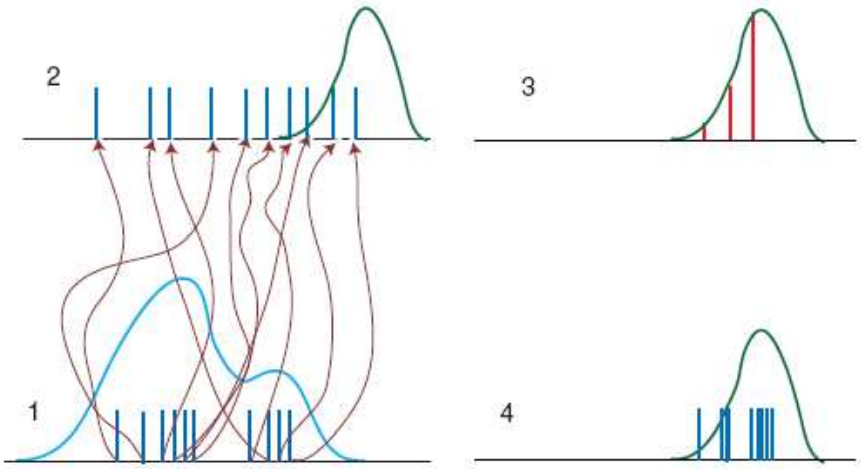


Figure 2.2 The standard particle filter (van Leeuwen, 2010). The prior (blue in the online article) pdf is sampled by a number of particles (10 in this case), indicated by the vertical bars (dark blue). These particles are all propagated forward in time using the full nonlinear equations, indicated by the lines (brown). When observations are present we see the prior particles as vertical bars (blue) again. The pdf of the observations is given by the curve (green).

A limitation with the PF in this form is that the particles which continually remain furthest from the observations will get the smallest weights until these weights become close to zero. Van Leeuwen (2009) found that when applying this method to a low-dimensional model, after only a few analysis steps most of the weight lay with one particle. This is referred as *filter degeneracy*. To overcome this problem, the PF was modified such that particles with very low weights are thrown away, and particles with high weight are resampled (i.e. multiple copies are made); the higher the weight of the particle, the more copies are made, until the total number of particles becomes N again.

2.2.7 Optimal Interpolation

This was one of the first DA schemes to be used by the Met Office to make weather forecasts (Cressie, 1990). It is similar to KF except that the error covariance matrix represents spatial correlation rather than uncertainty in the state vector. Its advantage is that it is simple to implement and it can be quick to run. Its weaknesses are that no uncertainty is specified for the observations (or the model) – the result of this is that when different sets of observations are used on different parts of the model state, biases in the model's estimates can occur.

2.2.8 The Metropolis algorithm

The Metropolis algorithm is a full Markov Chain Monte Carlo (MCMC) algorithm. It uses a numerical least squares approach to find the distribution of model parameters that provides optimal fits between the model and observations, and can avoid local minima. Unlike traditional least squares, which finds a single set of parameters (the optimal set), the Metropolis algorithm finds a distribution of parameter sets which are close to the set that gives the global minimum of the sum of squared differences between the observations and model. The algorithm was first developed by Metropolis et al. (1953), and was adapted for applications to geophysics (Mosegaard & Tarantola, 1995) and later for terrestrial carbon modelling by Knorr & Kattge (2005), Hill & Williams (2009) and Fox et al. (2009). We begin by giving an overview of the basic algorithm.

OVERVIEW OF BASIC FORMULATION OF THE METROPOLIS ALGORITHM

In the subsections that follow this one, we describe how the Metropolis algorithm is applied in a Bayesian setting, but when the algorithm was first developed it was not developed fort this purpose. Metropolis et al. (1953), who first formulated the Metropolis algorithm, used it to determine ideal pressure of a configuration of interacting molecules (simulated using a simple model) at a phase transition, given a specified temperature and a specified number of molecules. The configuration was optimal if the corresponding pressure and pre-determined temperature obeyed the ideal pressure law. The ideal pressure, based on a specific configuration, could not be determined analytically because it required determining the expectation of the Boltzmann distribution, used to simulate the distribution of molecules, which involved an unsolvable integral. It was unsolvable because it contained an unknown normalising constant, and the exponential of potentially complicated Energy function. The Metropolis algorithm is unique in that it allows one to generate samples from the desired distribution without needing to know the value of the normalising constant. Put simply if we wish to know the probability distribution g(x), where g(x) = c*f(x), and c is known but f(x) is known, the Metropolis algorithm allows us to approximate g(x) withoug knowing c. For this reason, it has become a very widely used algorithm with the Metropolis et al. (1953) paper being cited close to 20,000 times.

We now present a basic overview of the algorithm. As above, let g(x) be the probability distribution we require to know, and let f(x) be a probability distribution proportional to g(x). We first randomly select a point x_0, and choose a proposal distribution Q(x*|x) which described the locations of future points in the sample space starting with x_0. A common choice is the Normal distribution, which says that points close to the current sampled point x are more likely to be selected than further away points. The function Q is known as the proposal distribution. Then, for each iteration *i* (starting at i=0) a candidate point x* is chosen from the distribution $Q(x^*|x_i)$. An acceptance ratio $\alpha = f(x^*)/f(x_i) = g(x^*)/g(x_i)$ is then computed and this determines whether the candidate should be accepted or rejected. If $\alpha \geq 1$, then the candidate point x* is more likely than x_i, therefore we automatically accept x* as the new sampled point x_{i+1}. If $\alpha < 1$, then we accept the candidate with probability α. The algorithm continues in this way for i=1,2,3,4, …., until part of the sample space with highest probability is reached.

A key development of the algorithm occurred through Hastings (1970), and we describe this now. The Metropolis algorithm operates by randomly moving around the sample space (full details provided after this subsection); this movement is controlled by a proposal distribution Q(x*|x), where x* is the proposed point in the space and x is the current location in the space. For the Metropolis algorithm described by Metropolis et al. (1953), the proposal distribution needs to be symmetrical. Hastings (1970) extended the algorithm so as the propropsal distribution need not be symmetric. This extended version of the algorithm is thus called the Metropolis-Hastings algorithm, and the traditional Metropolis algorithm is thus a special case.

A parameter of the proposal distribution is the step-size, or some quantity which describes how big a jump to make when jumping to a new part of the sample space. For example, if the Normal distribution is chosen for the proposal, the size of the jumps is controlled by the standard deviation parameter. When the function g(x) we wish to sample from is multi-dimensional, finding the right proposal distribution can be difficult in order for the step-size to be just right in all dimensions. Problems can arise through choosing the wrong one. Gemen & Gemen (1984) overcame issues such as these by proposing Gibbs sampling. This involves proposing a new sampled

point in each dimension, rather than a multi-dimensional sample point, conditioned on the most recent values of all the others.

Further variants of the Metropolis algorithm continue to emerge, for example Černý, V. (1985) proposed adaptive simulated annealing, where the stepsize varies for different iteration steps. Applications of the Metropolis-Hastings algorithm are diverse and include, among others, Medicine (Gilks & Best, 1995), geophysics (Mosegaard & Tarantola, 1995), hydrology (Vrugt et al., 2003) and terrestrial carbon modelling (Knorr and Kattge, 2005).

A detailed description of how the algorithm has been applied in terrestrial carbon modelling is now given, based on Knorr & Kattge (2005), Hill & Williams (2009), and Fox et al. (2009).

THE DYNAMIC MODEL

Let the dynamic model be represented by the formula:

$$\mathbf{m}=\mathbf{M}(\mathbf{p},\mathbf{S}) \tag{2.28}$$

where

M denotes the actual model.

m is the output of the model represented by an (n×1) vector covering the n time-points the model is run for. For chapters three and four which use the Metropolis algorithm, there were at most two types of observations (which were NEE and LAI). To make the explanation more straightforward, we assume that there is only one type of observation, i.e. **m** is a vector. However we provide more explanation when there are two types (i.e. **m** is a two column matrix).

p is the (p×1) vector of model parameters,

S is the (n×s) matrix of site specific inputs, where s is the number of inputs, for example the meteorological observations such as minimum and maximum temperatures at each time-point.

Since the elements of **S** are fixed and known, the RHS of equation (2.28) is written as **M(p)**. Let **d** represent the (l×1) vector of observations ($1 \leq l \leq$ n) over the 1:n time-window, and let **m*** be a sub-vector of **m** where elements of **m** are included in **m*** if observations are present at the same time-point. For example, if n=10 and observations are only present at the 2nd, 3rd and 10th time-points, then **m*** is a 3×1 vector composed of the 2nd, 3rd and 10th elements of **m**. Since **m*** is determined from

m and **m=M(p)**, **m*** is also technically a function of **p**. For ease of the explanation that follows we set **m*=M(p)**.

BAYES THEOREM

The purpose of the Metropolis algorithm is to find a set of parameters **p** where the misfit between **m*** and **d** is minimized. Frequentist techniques, such as least squares, will give one estimate of **p**. However, Bayesian techniques instead give a distribution of estimates of **p**. Bayesian techniques use Bayes's theorem (Bayes, 1763) at the core of their calculations. In a nutshell, this states that if you have an unknown quantity (e.g. a parameter in a model), rather than just using observations to gain an estimate of this unknown (as is the case in frequentist statistics), we utilize other information we might have, which we refer to as the prior. This prior is represented by a probability distribution. Using Bayes' theorem, we can combine this prior knowledge with our observations to obtain a distribution (rather than a single estimate), called the posterior distribution, by the following formula:

$$f(\mathbf{p}) = v\, L(\mathbf{p})\, \rho(\mathbf{p}) \tag{2.29}$$

where v is a normalization constant, $L(\mathbf{p})$ is the likelihood function which expresses the degree of fit between the data predicted from the model and the observed data and $\rho(\mathbf{p})$ is the prior pdf (Mosegaard & Tarantola, 1995). The likelihood function is expressed as:

$$L(\mathbf{p}) = \exp\{-J_m(\mathbf{p})\} \tag{2.30}$$

with

$$J_m(\mathbf{p}) = \tfrac{1}{2}(\mathbf{M(p)} - \mathbf{d})^T\, \mathbf{C}_d^{-1}\, (\mathbf{M(p)} - \mathbf{d})$$

where $\mathbf{C}_d$ is the error covariance matrix of the observations. Since we assume that the observational errors are Normally distributed, when there is only one observation type $\mathbf{C}_d$ is simply the variance of this distribution. When there are two observation types, $\mathbf{C}_d$ is a 2×2 matrix where the diagonal elements are the variances of the Normally distributed errors of the two observation types and the off-diagonals represent any covariance between the two types. In a similar manner the prior pdf, $\rho(\mathbf{p})$, is written as:

$$\rho(\mathbf{p}) = \exp\{-J_p(\mathbf{p})\} \tag{2.31}$$

with

$$J_p(\mathbf{p}) = \tfrac{1}{2}(\mathbf{p} - \mathbf{p_0})^T \, \mathbf{C}_p^{-1} (\mathbf{p} - \mathbf{p_0})$$

where $\mathbf{p_0}$ is the vector of prior values of the parameters (our first guess at what the parameters should be), and $\mathbf{C}_p$ is the error covariance matrix of the priors which represents the uncertainty of these prior beliefs. The above is consistent with what is described in Mosegaard & Tarantola (1995), Knorr & Kattge (2005), Hill & Williams (2009) and Fox et al. (2009).

Throughout this subsection the priors are assumed to be Normally distributed, and this is what was done for the Metropolis algorithm in chapter four. However in some situations, we may have very poor knowledge about the priors; and the uniform distribution is often used (this is what is done in chapter three). At a later point in this subsection we describe how the implementation of the algorithm changes with this change of distribution for the prior.

USING MONTE CARLO TO APPROXIMATE THE POSTERIOR DISTRIBUTION

What is described below is consistent with what is described in Mosegaard & Tarantola (1995), Knorr & Kattge (2005), Hill & Williams (2009) and Fox et al. (2009). The posterior pdf, $f(\mathbf{p})$ (equation 2.29), is usually impossible to derive analytically so instead Markov Chain Monte Carlo techniques are used to approximate $f(\mathbf{p})$ by representing it as a sample of parameter sets $\mathbf{p}^{(1)}$, $\mathbf{p}^{(2)}$, …, $\mathbf{p}^{(N)}$, which have a distribution whose highest density is the maximum of $f(\mathbf{p})$. This sampled distribution can be used to compute the expected value of $f(\mathbf{p})$, using the standard expectation formula given by:

$$\langle f(\mathbf{p}) \rangle = \int \mathbf{p} f(\mathbf{p}) \, . \mathrm{d}\mathbf{p} \cong \frac{1}{N} \sum_{i=1}^{N} \mathbf{p}^{(i)}$$

Since $L(\mathbf{p})$ and $\rho(\mathbf{p})$ are assumed to be Normally distributed, analytically this would guarantee that their product, i.e. the posterior distribution $f(\mathbf{p})$ (equation (2.29), would also be Normal. Using Monte Carlo techniques to approximate $f(\mathbf{p})$ should also result in it being Normal – whether it is or whether it is not, we use the modal value (i.e. the value with the highest probability) as the best estimate of the parameter. If the posterior is Normal, then the mode is approximately equal to the mean. The width of the marginal distribution gives us an idea of the uncertainty we have on each parameter estimate; this is represented by an interval with the 2.5th and 97.5th percentiles as its lower and upper bounds.

LOG-NORMALISED PARAMETERS

We follow the methodology of Knorr & Kattge (2005), Hill & Williams (2009) and Fox et al. (2009) where a differentiation is made between model parameters written as $\{\underline{P}_1^{(m)}, \underline{P}_2^{(m)}, \ldots, \underline{P}_M^{(m)}\}$ and log-normalized parameters, $\mathbf{p}=\{p_1^{(a)}, p_2^{(a)}, \ldots, p_M^{(a)}\}$, the latter of which are used in the Metropolis algorithm, where the superscripts *(m)* and *(a)* denote *model* and *algorithm*, and where M is the number of model parameters (e.g. for DALEC and DALEC-D, M equals 11 and 17 respectively). In order for the model parameters to always be positive (which is important as it is biologically meaningless for them to be negative), we assume that the prior distribution of the *i*th model parameter $\underline{P}_i^{(m)}$ is Lognormal with probability distribution function given by:

$$f(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(\frac{(\ln x - \mu)^2}{2\sigma^2}\right), \quad x > 0$$

A consequence of this is that when we take the natural logarithm, the resulting distribution is Normal. In other words, if $\underline{P}_i^{(m)} \sim Log\text{-}N(\mu, \sigma^2)$, then $\ln(\underline{P}_i^{(m)}) \sim N(\mu, \sigma^2)$. To make the algorithm run more efficiently, we also scale the model parameters so that the means of each of them (i.e. each $p_i^{(a)}$) are the same. Knorr & Kattge used a value of 1, and for consistency this same value is adopted here, thus $p_i^{(a)} \sim N(1, s^2)$. The authors also used equation (2.24) to link $\underline{P}_i^{(m)}$ to $p_i^{(a)}$, and the same formula is adopted for this thesis too, again for consistency.

$$p_i^{(a)} = \ln\left(\frac{\underline{P}_i^{(m)}}{\underline{P}_{i,0}^{(m)}}\right) + 1 \qquad (2.32)$$

where $\underline{P}_{i,0}^{(m)}$ is mode of the log-Normal distribution prescribed to $\underline{P}_i^{(m)}$, which is prescribed beforehand. The Metropolis algorithm can be initialised from random location in the parameter space, but here (as was done in Knorr & Kattge, 2005) they are initiated from the mode of $\underline{P}_i^{(m)}$ in the model parameter space, i.e. $\underline{P}_{i,0}^{(m)}$. Thus, the *i*th parameter used for the algorithm was initialised from 1, i.e. $\underline{P}_i^{(m)} = \underline{P}_{i,0}^{(m)} \Leftrightarrow p_i^{(a)} = 1$ for all.

In Knorr and Kattge (2005), three different runs of the Metropolis algorithm were carried out using different values of *s*, namely 0.125, 0.25 and 0.5. However, the authors of that paper did not say, given the form of equation (2.32), whether these values were appropriate. In the text that follows we describe that, under the

assumptions of $\underline{P}_i^{(m)}$ following a Log-Normal distribution, having the mode of $\underline{P}_{i,0}^{(m)}$ and a variance of $\left(s\times\underline{P}_{i,0}^{(m)}\right)^2$, what the most appropriate value of *s* is. Note that under the form of equation (2.32) any value of for the mean of $p_i^{(a)}$ is acceptable, thus the choice of 1 is arbitrary.

First, we start by rearranging to make $\underline{P}_i^{(m)}$ the subject of equation (2.32), i.e.

$$\underline{P}_i^{(m)} = \underline{P}_{i,0}^{(m)} \exp\left(p_i^{(a)} - 1\right),$$

With the mean and variance of the prior distribution of $p_i^{(a)}$ set to 1 and s^2, this means $p_i^{(a)} \sim N(1,\ s^2)$

$$\Rightarrow \quad \ln\left(\frac{\underline{P}_i^{(m)}}{\underline{P}_{i,0}^{(m)}}\right)+1 \sim N(1,\ s^2) \qquad \text{(using equation 2.32)}$$

$$\Rightarrow \quad \ln\left(\frac{e\underline{P}_i^{(m)}}{\underline{P}_{i,0}^{(m)}}\right) \sim N(1, s^2) \qquad \text{(using ln(ab)=ln(a)+ln(b), and ln(e)=1)}$$

$$\Rightarrow \quad \frac{e\underline{P}_i^{(m)}}{\underline{P}_{i,0}^{(m)}} \sim LogN\left(1, s^2\right)$$

Letting $X=\left(e/\underline{P}_{i,0}^{(m)}\right)\underline{P}_i^{(m)}$, Y=g(X) where g(X) = $\left(\underline{P}_{i,0}^{(m)}/e\right)X=\underline{P}_i^{(m)}$ and noting that g is strictly monotonic, the probability distribution function of Y (i.e. $\underline{P}_i^{(m)}$) can be determined by using the standard formula:

$$f_Y(y) = f_X(g^{-1}(y))\left|\frac{d}{dy}g^{-1}(y)\right|$$

This results in: $\underline{P}_i^{(m)} \sim LogN\left(\ln(\underline{P}_{i,0}^{(m)}),\ s^2\right)$

Thus using the notation from the previous page, $\mu = \ln(\underline{P}_{i,0}^{(m)})$ and $\sigma^2 = s^2$ and the fact that the mode and variance of a Log-Normal distribution is $e^{\mu-\sigma^2}$ and $(e^{\sigma^2}-1)e^{2\mu+\sigma^2}$, this means that mode and variance of the pdf of $\underline{P}_i^{(m)}$ is very close to $\underline{P}_{i,0}^{(m)}$ and $\left(s\times\underline{P}_{i,0}^{(m)}\right)^2$ as long as $s\leq 0.5$.

A final point to note is that although the model parameters are converted to log-Normalised parameters at the start of the Metropolis algorithm, they are converted

back to model parameters before being outputted. In this way, estimates of the model parameters can be compared with other DA methods since it is the model parameters that the parameter estimates are based on.

THE PRIOR DISTRIBUTION OF THE LOG-NORMALISED PARAMETERS

As explained, the log-normalized parameters are all initialized at 1, i.e. $\mathbf{p}$={1,1, …, 1} with corresponding standard deviation fixed at *s*. In terms of the covariances of the parameters for the prior distribution, these are difficult to estimate well, thus for simplicity they are all assumed to be zero. This is consistent with Knorr and Kattge (2005), Hill & Williams (2009) and Fox et al. (2009). Therefore the (i,j)*th* element of the error covariance matrix of the priors is given by:

$$\mathbf{C}_{p_{i,j}} = \begin{cases} s^2 \ if \ \ i = j \\ 0 \quad else \end{cases} \tag{2.33}$$

THE METROPOLIS ALGORITHM

The Metropolis algorithm operates by randomly proposing steps in the parameter space (if the model were to have 2 parameters, the parameter space would be a plane as illustrated in figure 2.3). The sequence of steps in the algorithm is initialized from the mean of the prior parameter distribution, i.e. $\mathbf{p}_i^0$=1 (for the *i*th parameter) which is $\underline{P}_{i,0}^{(m)}$ in model parameter space. It can alternatively be initialized from a random point in the prior parameter space (i.e. for each *i* randomly selecting a number from N(1,0.5)). Subsequent iterations $\mathbf{p}^1$, $\mathbf{p}^2$, … were generated by making a small step $\Delta\mathbf{p}$, i.e. $\mathbf{p}^{j+1}=\mathbf{p}^j+\Delta\mathbf{p}$. Knorr and Kattge (2005) determine $\Delta\mathbf{p}$ by generating a zero mean random vector with error-covariance given by $\mathbf{C}_p$ (equation 2.33). In this thesis $\Delta\mathbf{p}$ was determined in the same way as in Hill & Williams (2009), where each element of the parameter set has a 1/3 chance of increasing by the step-size, 1/3 chance of staying the same and a 1/3 chance of decreasing by the step-size (note: the step-size is chosen so that the algorithm efficiently explores the parameter space; how it is chosen is explained later). Hill & Williams (2009) found from tests that this was a more efficient way of exploring the parameter space than selecting a random step from the Gaussian distribution. We accept the proposed parameter set (i.e. $\mathbf{p}^{j+1}$**)** in a 2-stage acceptance procedure based on the prior and then the observations, as is now

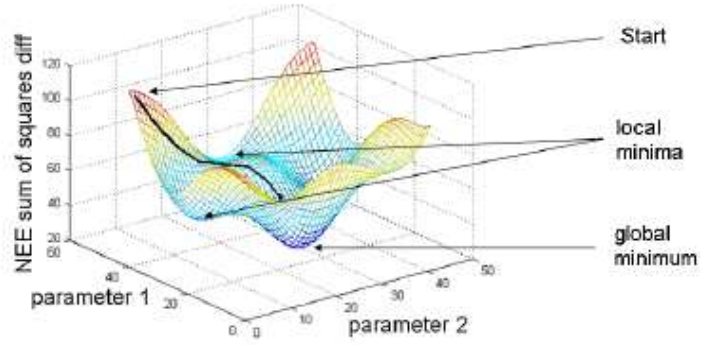


Figure 2.3 A simplified schematic of how the Metropolis algorithm works with 2 model parameters.

explained. Traditionally this 2 stage acceptance procedure is done in one step, however doing it in two stages makes the algorithm more computationally efficient because we do not do the second stage if the first condition is not satisfied.

Condition 1: Acceptance in terms of the prior distribution

The proposed parameter set is accepted in terms of the prior if the sum of squared differences between the proposed and the prior parameter sets (written as $S_p(\mathbf{p}^{j+1})$) is less than that of the previous iteration (i.e. $S_p(\mathbf{p}^{j})$), where the sum of squared differences for a general step *k* is defined as:

$$S_p(\mathbf{p}^{k}) = \sum_{i=1}^{M} \left(p_i^k - p_{0,i}\right)^2 ,$$

with p_i^k and $p_{0,i}$ denoting the *i*th parameter value for the *k*th step and the prior respectively, and *M* representing the number of parameters. This constraint ensures that the proposed parameter steps do not stray too far from the prior parameter set. If the sum of square differences is not reduced, the condition can still be satisfied with probability:

$$\frac{\rho(\mathbf{p}^{j+1})}{\rho(\mathbf{p}^{j})}$$

where $\rho(\mathbf{p}^{j})$ and $\rho(\mathbf{p}^{j+1})$ are the pdfs of the proposed and current parameter sets respectively (equation (2.31)). This is used to avoid the algorithm converging to local

minima. This will never be greater than 1 because if it is expanded using equation (2.31) and in terms of $S(\mathbf{p}^{j})$ and $S(\mathbf{p}^{j+1})$, it becomes:

$$\mathrm{C}=\exp\left(\frac{S(\mathbf{p}^{j})-S(\mathbf{p}^{j+1})}{2s_p^2}\right),$$

where s_p^2 is the variance of the prior distribution. Since in this case $S_p(\mathbf{p}^{j+1}) \geq S_p(\mathbf{p}^{j})$, this implies that C≤1. Moreover, if $S_p(\mathbf{p}^{j+1})$ is only slightly greater than $S_p(\mathbf{p}^{j})$ then the probability of acceptance is much higher than if $S_p(\mathbf{p}^{j+1})$ is significantly greater than $S_p(\mathbf{p}^{j})$.

Condition 2: Acceptance in terms of the observations

If there is one set of observations, say NEE, the proposed parameter set is accepted in terms of the observations if the sum of squared differences between the observations and the NEE model trajectory corresponding to the proposed parameter set (denoted $S(\mathbf{m}^{j+1})$), is less than that of the current parameter set (i.e. $S(\mathbf{m}^{j})$), where $S(\mathrm{m}^{k})=\sum_{l=1}^{L}\left(m_{NEE}^{k}-d_{NEE}\right)^2$, with m_{NEE}^{k} and d_{NEE}^{k} denoting the *l*th modelled and observed NEE for each of the *L* observations. As with the prior acceptance rule, if the sum of squared differences is greater for the proposed step, then the proposed parameter set is accepted according to the observations, with probability:

$$\frac{L(\mathbf{p}^{j+1})}{L(\mathbf{p}^{j})} = \exp\left(\frac{S(\mathbf{m}^{j})-S(\mathbf{m}^{j+1})}{2s_o^2}\right),$$

where s_o^2 = variance of observational error. Similarly to before, this ratio will always be less than or equal to 1 since it is only used when $S(\mathbf{m}^{j+1}) \geq S(\mathbf{m}^{j})$. This is consistent with Knorr & Kattge (2005), Hill & Williams (2009) and Fox et al. (2009). If two sets of observations are used, say NEE and LAI then instead of the above expression we use:

$$\frac{L(\mathbf{p}^{j+1})}{L(\mathbf{p}^{j})} = \exp\left(\frac{S(\mathrm{m}^{j})-S(\mathrm{m}^{j+1})}{2s_{NEE}^2 s_{LAI}^2}\right),$$

where $S(\mathbf{m}^{j}) = s_{LAI}^2\sum_{l=1}^{L}\left(m_{NEE}^{k}-d_{NEE}\right)^2 + s_{NEE}^2\sum_{l=1}^{L}\left(m_{LAI}^{k}-d_{LAI}\right)^2$, s_{LAI}^2 and s_{NEE}^2 are the

variances of the observational random errors. This is what was done in Fox et al. (2009).

If both conditions 1 and 2 are not both satisfied, then the parameter set of the previous iteration is used.

CONVERGENCE OF THE ALGORITHM TO THE POSTERIOR DISTRIBUTION

To determine the convergence of the chains to an area around the global maximum of the posterior probability distribution rather than local maximum, the Gelman criteria (Gelman, 1995) was applied. Convergence of the chains was assessed in exactly the same way in Knorr & Kattge (2005), Hill & Williams (2009) and Fox et al. (2009). The Gelman Criteria describes a reduction factor which computes the weighted ratio of the between-chain and within-chain variances of the J chains (J>1) that are used. As the number of iterations tends to infinity, the reduction factor R tends to 1. Brooks & Gelman (1998) state that an R value of less than 1.2 demonstrates that the chains have converged. The same authors also state that if it takes N iterations for the R values for all parameters to be less than 1.2, the algorithm should be run for N more iterations. The second set of N iterations then represents a sample from the posterior distribution.

After convergence, the distribution of parameter sets is referred to as the posterior distribution. Tests showed that 5,000,000 iterations were sufficient to ensure convergence. The algorithm was then run for a further 5,000,000 in order to build the posterior distribution.

The estimate of each parameter and the bounds of the corresponding 90% posterior interval are calculated by computing the mean, 5^{th} and 95^{th} percentiles from every 200^{th} step to avoid autocorrelations (Knorr & Kattge, 2005) from all the chains after convergence, i.e. from iterations 5,000,001 to 10,000,000. As stated earlier, if the marginal posterior distribution for any parameter is not Normal then the modal value of the marginal distribution is used to represent the best estimate of the parameter.

IMPLEMENTING THE METROPOLIS ALGORITHM

Implementing the Metropolis algorithm, following the procedure of Knorr & Kattge (2005), Hill & Williams (2009) and Fox et al. (2009) involves the following steps for

each of the parallel chains:

(1) Determine the initial parameter set by using the modal value from the prior distribution or by random generation from the prior distribution. Convert the model parameters to Log-Normal parameters using equation (2.32).

(2) Using the initial parameter set, an initial model run is carried out to obtain model predictions, $M^{(0)}$, where $M^{(0)}$ is a matrix with the rows representing the length of the dataset and the columns the number of observation types.

Then for j = 0, 1, 2, …:

(3) A new parameter set is proposed, P_i^{j+1} (where $P_i^0 = \underline{P}_{i,0}$) for each i=1,…,17, by stepping from the current parameter (P_i^j) at the current step j. Stepping occurs in the following way – each parameter has a 1/3rd probability of increasing, 1/3rd probability of decreasing and a 1/3rd probability of staying the at the current value (P_i^j). The step-size is therefore chosen so that the acceptance probability is between 0.2 and 0.8 (Fox et al., 2009, Hill & Williams, 2007). Repeating the algorithm with different step-sizes, it can be shown that as the step-size increases the acceptance rate decreases tending to 0%, and as the step-size decreases the acceptance rate increases tending to 100%. Hill & Williams (2009) state that if the step-size is too large or too small (i.e. corresponding to an acceptance rate of less than 20% or greater than 80% respectively), this will result in inefficient exploration of the parameter space. A step-size too large will mean that the chains will jump around different parts of the parameter space, and a step-size too small will result in a low number of sampled points in the parameter space (Ziehn et al., 2012). In either case, the chains will fail to converge to take excessively long to converge.

(4) The proposed parameter set is accepted in terms of the prior if the sum of squared differences between the proposed and the prior parameter sets (written as $S_p(\mathbf{p}^{j+1})$) is less than that of the previous iteration (i.e. $S_p(\mathbf{p}^j)$). If the sum of square differences is not reduced, the condition can still be satisfied with probability:

$$\exp\left(\frac{S\left(\mathbf{p}^j\right) - S\left(\mathbf{p}^{j+1}\right)}{2s_p^2}\right),$$

where s_p^2 is the variance of the prior distribution.

(5) The model is now run using the proposed parameter set (i.e. P_i^{j+1}) in order to compute the corresponding model predictions, i.e. $M^{(j+1)}$.

(6) The proposed parameter set is accepted in terms of the data if the sum of squared differences between model and observations (i.e. $S(\mathbf{m}^{j+1})$) is less than that of the previous iteration (i.e. $S(\mathbf{m}^{j})$). As with the prior acceptance rule, if the sum of squared differences is greater for the proposed step, then the proposed parameter set is accepted according to the observations with probability:

$$\exp\left(\frac{S(\mathbf{m}^{j}) - S(\mathbf{m}^{j+1})}{2s_o^2}\right),$$

where s_o^2 = variance of observational error.

(7) If the proposed parameter set is accepted based on both the prior and data then the proposed parameter set becomes the current parameter set (i.e. ($P_i^{j} = P_i^{j+1}$) and the model predictions get similarly updated ($\mathbf{m}^{j} = \mathbf{m}^{j+1}$).

(8) For j = 1, 2, …, steps 3 to 7 are repeated until each parameter has converged which is determined by visual inspection (i.e. plotting the trajectories of the parameters)

(9) Assuming all parameters have converged after the *N*th iteration, the parameter sets from the *(N+1)*th to the *2N*th iterations are extracted. To avoid autocorrelations every 200th parameter set in this sequence of *N* is chosen to represent samples from the posterior distribution.

USING A UNIFORM DISTRIBUTION FOR THE PRIOR

Up until this point, a Normal distribution has been adopted for the prior. This is appropriate if we have specific belief about what values the parameters should take and are able to provide an expression for the uncertainty for that belief. If our prior knowledge of the parameters is weak, a uniform distribution may be more appropriate. This uniform distribution means that each value of the parameter in the given range has an equal chance of being the true parameter value. This kind of prior is often called a flat prior because no confidence is placed on a particular value in the

distribution unlike a Gaussian, where the mean is regarded as being the most likely value of the true parameter. The implementation of the algorithm using a uniform prior is almost identical to one using a Normal prior, except that step (4) is missed out, as was done in Fox et al. (2009). For the two chapters where the Metropolis algorithm was used both types of prior distributions were adopted (a Uniform in chapter three and a Normal in chapter four).

CRITICAL REVIEW OF THE METROPOLIS ALGORITHM

The main advantage of the Metropolis algorithm is its ability to approximate the the entire posterior distributions of the parameters, in particular identifying the location of the modal parameter value(s). In this way it is superior to other DA schemes such as the EnKF which has a problem when the posterior parameter distribution is multi-modal or non-symmetrical since it gives the mean as the best estimate of the parameter (Evensen, 2009). It is also superior to gradient descent algorithms such as 3D-Var or 4D-Var which suffer from the corresponding cost function converging to local minimum rather than the global one (Williams et al., 2009). The Metropolis algorithm has been successfully applied in a whole multitude of cases. By 'successfully applied', we mean that the marginal posterior distributions of the parameters have been narrower, indicating a reduction in uncertainty on the parameter estimates, compared to the prior estimates of them. Examples include Knorr & Kattge (2005), Hill & Williams (2009), Fox et al. (2009), Richardson et al. (2010), MacBean et al. (2010) and Ziehn et al. (2012).

A major limitation of the Metropolis algorithm and other MCMC methods is the computation time required if the model is large, particularly if there are lots of parameters to estimate. Ziehn et al. (2012) ran the Metropolis algorithm with 8 chains to estimate the posterior probability distributions for 19 of the parameters from the BETHY model (Knorr, 2000) for 8 million iterations – all the chains were run on a computer cluster in parallel, which allowed up to one million iterations per month and chain, so took eight months in time. Currently therefore, this makes them infeasible to estimate the marginal posterior distributions of all parameters of global terrestrial C models which can easily be in excess of 100. However, they have been widely applied to models less computationally demanding.

2.2.9 Genetic algorithm

We give a brief overview of this algorithm based on Gallagher & Sambridge (1994). The Genetic algorithm is similar to the Metropolis algorithm in that it searches the parameter space to find the global minimum. However, it is a non-Bayesian DA scheme, so no prior can be incorporated into the searching – the search is only guided by the cost function or fitness function as it is more commonly known. It starts with a population of N sets of randomly selected parameters sets called individuals (in the Genetic algorithm used in Fox et al, 2009, N was set to 100). The model is run for each of these parameter configurations and the fitness function is evaluated to assess the fit between the model output and data. The individuals whose corresponding fit is poor (defined by some threshold) are discarded. These are replaced by reproduction or mutation. Reproduction is where two individuals 'mate' to produce a new individual – in practical terms this could be simply that the new parameter set is composed using half of the parameters from each of its 'parents'. There are other ways of reproduction as outlined in Haupt & Haupt (2004). Occasionally, an individual will be created by 'mutation', where the values of the new parameter set will be chosen at random. According to Gallagher & Sambridge (1994), this ensures that the algorithm does not converge to a local minimum. This process is repeated until the fitness values of the individuals in the population converge. As this is a non-Bayesian scheme, the goal is to find the global minimum and the parameter set with this optimal fitness is the estimate of the parameter. No formal regard to given for quantifying the uncertainty for the estimate, and in this way, it is similar to the variational DA approaches whose main goal is to find the parameter set with the optimal fitness solely. This is different to the Metropolis algorithm, which once converged, samples the posterior parameter distribution function. As we require to sample the posterior distribution, and this is something which is not theoretically considered partly because no prior distribution is specified, this is a limitation of the Genetic algorithm for application here. It has also not been as widely used as other global search algorithms such as the Metropolis algorithm.

2.2.10 Variational DA approaches: 3D-VAR and 4D-VAR

PARAMETER ESTIMATON

Variational methods of DA are similar to MCMC methods such as the Metropolis in that they explore the parameter space with the aim of locating the posterior distribution from which it can estimate. However, while the Metropolis algorithm explores the parameter space in a stochastic way variational methods are not stochastic. Instead they attempt to locate the maximum of the posterior distribution in a more computationally efficient way (Santaren et al., 2007). We now derive the 3D-Var and 4D-Var methods based on Ziehn et al. (2012).

The formulation for 3D-Var and 4D-Var is the same as the Metropolis algorithm, i.e. based on Bayes' theorem (Bayes, 1753). Recall that in subsection 2.3.9 the likelihood and prior parameter distribution was given by equations (2.30) and (2.31) which are in their respective full forms:

$$L(\mathbf{p}) = \exp\{-\tfrac{1}{2}(\mathbf{M(p)} - \mathbf{d})^T \ \mathbf{C}_d^{-1}(\mathbf{M(p)} - \mathbf{d})\}$$

$$\rho(\mathbf{p}) = \exp\{-\tfrac{1}{2}(\mathbf{p} - \mathbf{p_0})^T \ \mathbf{C}_p^{-1}(\mathbf{p} - \mathbf{p_0})\}$$

where **M(p)** and **d** are the modelled and observed data, $\mathbf{p_0}$ is the prior parameter set, and $\mathbf{C}_d$ and $\mathbf{C}_p$ are the error covariance matrices of the observations and priors. In a similar way to the Metropolis algorithm the posterior distribution of the parameters is given by $f(\mathbf{p}) = v\,L(\mathbf{p})\,\rho(\mathbf{p})$, where v is a normalization constant. Setting v equal to 1/A, and substituting $L(\mathbf{p})$ and $\rho(\mathbf{p})$ into this formula gives:

$$f(\mathbf{p}) = \frac{1}{A}\exp\{-\tfrac{1}{2}(\mathbf{M(p)} - \mathbf{d})^T \ \mathbf{C}_d^{-1}(\mathbf{M(p)} - \mathbf{d})\} \times \exp\{-\tfrac{1}{2}(\mathbf{p} - \mathbf{p_0})^T \ \mathbf{C}_p^{-1}(\mathbf{p} - \mathbf{p_0})\}$$

This is equivalent to:

$$f(\mathbf{p}) = \frac{1}{A}\exp[-J(\mathbf{p})] \qquad (2.34a)$$

where,

$$J(\mathbf{p}) = \exp\{-\tfrac{1}{2}(\mathbf{M(p)} - \mathbf{d})^T \ \mathbf{C}_d^{-1}(\mathbf{M(p)} - \mathbf{d})\} \times \exp\{-\tfrac{1}{2}(\mathbf{p} - \mathbf{p_0})^T \ \mathbf{C}_p^{-1}(\mathbf{p} - \mathbf{p_0})\} \qquad (2.34b)$$

Thus, determining the location in the parameter space of the modal value of the posterior probability distribution is the same as finding the minimum of J(**p**). The minimum is determined by differentiating J(**p**) w.r.t. **p** to determine its gradient. The exploration in the parameter space is determined by the downward direction of the adjoint, the differentiated J(**p**). As a result of this a lot fewer iterations are typically required compared to global search algorithms like the Metropolis algorithm. As an

example, Zhien et al. (2012) used a variational DA method to estimate 19 of the parameters and associated distributions of the BETHY model. The authors found that the variational method required 1300 forward runs of the model compared to the Metropolis algorithm which was also used and needed 8 million.

STATE ESTIMATION

The description of the state estimation problem is based on Smith et al. (2008). Equation (2.34b) can also be used to estimate the model states with a fixed parameter set. This is referred to as the 3D Variational (or 3D-Var) state estimation method. It is also similar to the KF in one dimension in that it is the weighted sum of the model's estimate of the state of the system and what is observed. In multiple dimensions, it is the same as equation (2.34b) except that in optimising the parameter set **p**, we optimise the state vector at the current time t=0, labelled as $\mathbf{x}_0$, i.e.

$$J(\mathbf{x}_0) = (\mathbf{x}_0 - \mathbf{x}_b)^T \mathbf{B}^{-1} (\mathbf{x}_0 - \mathbf{x}_b) + (\mathbf{y}_0 - H_0[\mathbf{x}_0])^T \mathbf{R}_0^{-1} (\mathbf{y}_0 - H_0[\mathbf{x}_0]) \qquad (2.35a)$$

where **B** and **R** are the background and observational error covariance matrix which specifies the uncertainty in the state of the system given by the model ($\mathbf{x}_0$) and the observations ($\mathbf{y}_0$) at time (t=0). **B** and **R** are the same as $\mathbf{C}_p$ and $\mathbf{C}_d$ from equation (2.34b). H_0 is the mapping which converts the model state to a something comparable with the observations. As with the parameter estimation problem, the optimal estimate of $\mathbf{x}_0$ is obtained by finding the global minimum of $J(\mathbf{x}_0)$, which is achieved by using a global descent algorithm which requires $J(\mathbf{x}_0)$ to be differentiated. The 3D-Var method can be extended so that observations at future times are used to obtain the optimal $\mathbf{x}_0$ – this is referred to as 4D-Var and is written as follows:

$$J(\mathbf{x}_0) = (\mathbf{x}_0 - \mathbf{x}_b)^T \mathbf{B}^{-1} (\mathbf{x}_0 - \mathbf{x}_b) + \sum_{i=0}^{n} (\mathbf{y}_i - H_i[\mathbf{x}_i])^T \mathbf{R}_i^{-1} (\mathbf{y}_i - H_i[\mathbf{x}_i]) \qquad (2.35b)$$

Here, $\mathbf{x}_i$ refers to the model state vector at the *i*th time-step, and is determined by i integrations of the model from t=0. For example, $\mathbf{x}_3 = \mathbf{M}[\mathbf{x}_2] = \mathbf{M}[\mathbf{M}[\mathbf{x}_1]] = \mathbf{M}[\mathbf{M}[\mathbf{M}[\mathbf{x}_0]]]$, where **M** is the model. To determine the optimal $\mathbf{x}_0$ this, like 3D-Var, this also requires $J(\mathbf{x}_0)$ to be differentiated. However given that the RHS of equation (2.35b) involves the model **M,** this also needs to be differentiated. This can be very time consuming if the model is large. 3D-Var and 4D-Var are illustrated in figure 2.4. J_b and J_o refer to the first and second parts (i.e. before and after the plus sign of equations (2.34a) and (2.34b), and $\mathbf{x}_a$ refers to the optimal $\mathbf{x}_0$ (i.e. the value of $\mathbf{x}_0$ when $J(\mathbf{x}_0)$ is at

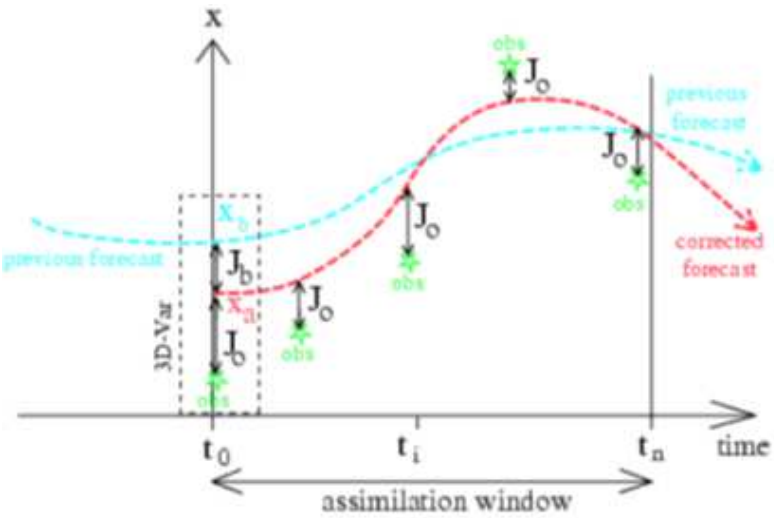


Figure 2.4 A schematic of the 3D-Var and 4D-Var schemes (Bouttier & Courtier, 1999). The 3D-Var only uses observations at the same time point as the background error covariance matrix, whereas 4D-Var uses future observations.

the global minimum).

As with the Ensemble Kalman Filter and Smoother, Smith et al. (2008) state that 3D-VAR and 4D-VAR can also be used to simultaneously estimate the parameters and states of a model. Following the description by Smith et al. (2008), this is done by augmenting the parameters to the state vector, forming an augmented state vector $\mathbf{w}$, such that: $\mathbf{w}=[\mathbf{z}\ \mathbf{p}]^{\mathbf{T}}$, where $\mathbf{w} \in \mathbb{R}^{m+q}$, $\mathbf{p} \in \mathbb{R}^{q}$, and $\mathbf{z} \in \mathbb{R}^{m}$. Treating the parameter vector $\mathbf{p}$ as time-invariant when the model is run forward in time on its own, the evolution of the model can then be written as $\mathbf{z}_{k+1} = \mathrm{M}(\mathbf{p}_k, \mathbf{z}_k)$ and $\mathbf{p}_{k+1} = \mathbf{p}_k$. This can then be written as $\mathbf{w}_{k+1} = \tilde{\mathrm{M}}(\mathbf{p}_k, \mathbf{z}_k)$. The least squares cost function for the augmented state vector is then the same as equation (2.35a) except that $\mathbf{x}_0$ is replaced with $\mathbf{w}$, resulting in $J(\mathbf{w}) = (\mathbf{w} - \mathbf{w}_b)^T \mathbf{B}^{-1} (\mathbf{w} - \mathbf{w}_b) + (\mathbf{y} - H[\mathbf{w}])^T \mathbf{R}^{-1} (\mathbf{y} - H[\mathbf{w}])$, where $\mathbf{B}$ refers to the background error covariance matrix for the augmented system.

CRITICAL REVIEW OF 3D-VAR AND 4D-VAR

The major advantage of Variational DA is its computational efficiency over other schemes such as global search algorithms like the Metropolis algorithm. As stated earlier, Zhien et al. (2012) found that the Metropolis algorithm required 8 million evaluations of the model, whereas the 4D-VAR method only required 1300 evaluations. In computation time, the Metropolis algorithm needed 8 months (run on a computer cluster with the chains running in parallel), whereas the 4D-VAR

approach was several orders of magnitude faster - the authors of Zhien et al. (2012) do not specify the exact amount. For this reason, the variational DA approach has been the DA scheme of choice for very large models such as the Numerical Weather Prediction (NWP) model (Lawless, 2001) and the JULES model (Luke, 2011).

Despite this advantage, variational DA has a few weaknesses. First of all, it is usually only able to find a local minimum of the cost function due to the non-linearity and high dimensionality of the model (Williams et al., 2009). A second problem is that it only focuses on estimating the modal value of the posterior probability density function, rather than that the rest of the distribution thus an estimate of the uncertainty is not normally given or assumptions need to be made (Ziehn et al., 2012). However, Rayner et al. (2005) showed that it is possible to estimate the spread of the posterior distribution by using work by Tarantola (1987) who found that Hessian - the second derivative of the cost function - at the global minimum approximates the inverse of covariance of the parameters. The covariance matrix describes the uncertainty of the parameters under the assumption that they follow a multivariate Gaussian distribution. However, Rayner et al. (2005) points out that the model needs linearising for this to be possible, and if the level of non-linearity in the model is strong the linear approximation may not be good enough to approximate the Hessian, particularly as this is only an approximation itself of the covariance matrix of the parameters. Furthermore, the Gaussian assumption of the posterior distribution may not be true. A final shortcoming of variational DA is where the algorithm is initialised from. Zhien et al. (2012) found that if it was initialised from a point further than 1% away from the modal point of the prior parameter distribution, then the algorithm either did not converge or converged to a non-physical point in the parameter space.

2.3 DA schemes used in thesis and tests

Having given a description and critical review of a broad range of DA schemes, the purpose of this is to provide an overview of the current state of knowledge of DA applied to the improvement our understanding and quantification of the terrestrial C cycle. From this, an outline will be given for the gaps in this current knowledge and which of these gaps are of most urgent need in filling. Arguments for the choice of DA schemes to be used in this thesis will then be presented in light of the type of

research questions to be addressed. The remaining part of section 2.3 consists of preliminary tests on the DA algorithms to be used in this thesis to ensure that they are implemented correctly. This testing of the schemes is also to gain an appreciation for what answers (e.g. parameter estimates) we might expect the DA algorithms to produce given different setups and assumptions. This will be particularly helpful with the discussion section of chapter four.

Looking ahead, the section that follow this one (2.4) will provide an argument for the choice of model to be used in the thesis, based on the type of research questions addressed for this PhD but also based on the DA algorithms that are chosen. Section 2.5 then describes and critically reviews the data required in later chapters. Finally, section 2.6 brings together the chapter by giving a list of specific research questions addressed in this thesis, based on the gaps in our current knowledge of DA in terrestrial C modelling (subsections 2.3.1 and 2.3.2) and the specific DA schemes, models and data used to answer the questions.

2.3.1 Current state of knowledge of DA

At the end of chapter one a broad overview of DA was given with specific descriptions and critical reviews given for nine different DA algorithms covering the three broad groups: global search algorithms, global descent algorithms and sequential algorithms (Williams et al., 2009). In this subsection, we describe how these three different groups of DA algorithms have been applied to constrain models with a focus on terrestrial C models, but other models are also included as well particularly those from oceanography and meteorology. Descriptions are also given of three different DA inter-comparison studies.

APPLICATIONS OF SEQUENTIAL DA ALGORITHMS: THE ENSEMBLE KALMAN FILTER, THE ENSEMBLE KALMAN SMOOTHER AND THE PARTICLE FILTER

Since its inception by Evensen (1994), the Ensemble Kalman filter (EnKF) has become one of the most widely used DA schemes in a whole range of modelling fields. Geir Evensen, the author of Evensen (1994) and an ocean modeler, first applied the EnKF to a multi-layered quasi-Geostrophic model, which put simply simulates the one aspect of the dynamics of large-scale motions of the oceans. Prior to

Evensen (1994), the extended Kalman filter or EKF (Evensen, 1992) had been used as the main alternative to the Kalman Filter for situations where the model was non-linear. Evensen (1994) found that not only did the EnKF give a small error (between the mean and true model states) compared to the EKF, but that unbounded linear instabilities of the evolution of the error statistics that existed with the EKF as a result of the linearization of the model, was not a problem with the EnKF because no linearization of the model was needed. The computational cost of the EnKF for reasonable accuracy was also a fraction of that of the EKF (Evensen, 1994).

The EnKF has been used in a whole range of applications, most notably in an operational context as the Data Assimilation method of choice for the large scale global atmospheric weather forecasting at the Meteorological Service of Canada (Houtekamer & Mitchell, 2005). In Houtekamer & Mitchener (2005), the focus was on what specific set up of the EnKF, e.g. the optimal number of ensembles, specification of the model error, etc…, would result in the smallest error between the mean and true model state vector. Evensen (2009) comments that the biggest advantage of using the EnKF for large, complicated and computationally intensive models (such as those used for weather forecasting is that EnKF) where the state vector can contain 10,000s of elements, is that the computation and propagation of the error-covariance matrix has dimension N×N where N is the number of ensemble members. This is because for the KF and the EKF the dimension of the error-covariance matrix is M×M, where W is the number of elements of state vector with M >>> N for the very large models such as the one used by the Meteorological Service of Canada (Houtekamer & Mitchell, 2005). In this way, Evensen (2009) states 'the uncertainty is represented by a set of model realisations rather an explicit expression for the error covariance matrix.'

Williams et al. (2005) was the first to apply an EnKF to a terrestrial carbon model, namely DALEC. Measurements assimilated into the model included Leaf area index (LAI), foliar and woody litter, gross primary production (GPP), total respiration (R_{tot}) and Net Ecosystem Exchange (NEE), and were collected at a young ponderosa pine stand in central Oregon over a 3-year period. The authors found that the model estimate of total cumulative NEE over the 3 years after assimilating all the data was -419±29 gCm^{-2}, which compared with a model-only estimate of -251±197 gCm^{-2}. The key finding of the study was that the uncertainty (expressed as a standard deviation)

was less when all the data had been assimilated (i.e. 29 gCm^{-2} or 0.47 $gCm^{-2}day^{-1}$) compared to either the model on its own (i.e. 197 gCm^{-2}) or the data on its own (0.5 $gCm^{-2}day^{-1}$). The authors provided as accurate as possible estimates of the uncertainties of the data and the model; they also carried out a simple sensitivity analysis to show that changes in the uncertainties of the model (which are harder to determine than those of the data) did not affect the main finding.

A next major EnKF development occurred in Quaife et al. (2008). Although ground based measurements of NEE and other C fluxes are useful for site based studies, satellite based measurements offer more potential for studies aiming to determine the terrestrial-atmosphere C flux for regional to global scales. In response to this, Quaife et al. (2008) assimilated satellite observations of canopy reflectance into the DALEC model at the same site and over the same 3 year period as the Williams et al. (2005) study. Assimilating this kind of data as opposed to high level satellite products (such as LAI) can be preferable for DA for several reasons, the main ones being: (i) the observational uncertainty is easier to quantify, and (ii) different satellites give different measurements of LAI (for example) because different assumptions are made in the algorithms that convert the satellite's raw data (i.e. canopy reflectance) to LAI. The downside to assimilating canopy reflectance data is that the observational operator (used to convert between the measurements and the model state vector) becomes extremely complex. The modelled NEE after assimilation showed an accurate representation of the ground based measurements of NEE for the study period, however for the wrong reasons: the modelled GPP and total respiration after assimilation tended to be over-estimated when compared to ground measurements. Despite this, these results provide great promise of the ability of the assimilation of satellite measurements such as canopy reflectance into large climate models, such as JULES (Blyth et al., 2010).

Searching the literature on the Ensemble Kalman Smoother (EnKS), application of this appears to significantly less than the EnKF. This may be because a smoothed model trajectory may not be required for a lot of settings. Also, as has been pointed out in subsection 2.2.6, the parameter estimates produced by the EnKF are as globally optimal as those from the EnKS (Evensen & Leeuwen, 2000). The EnKS is more computationally intensive and less straight-forward to implement than the EnKF. Nonetheless, of the fewer studies (compared to the EnKF) which have used

the EnKS, an interesting one is Stroud et al. (2010). The aim of that paper was to produce a time sequence of spatial maps during the development of a large sediment plume within and around Lake Michigan following a major storm in 1998, by combining incomplete satellite data and a model with an EnKS framework. The need for a smoother was important here because a detailed and complete picture of how the plume developed was essential. In other words, knowing the size of sediment built-up for each pixel of the map at very frequent time intervals was essential, in order to help with forecasting and planning in the event of a repeat of such a sediment plume following another major storm. As a result, the EnKS was preferred to the EnKF even at its greater computational cost. The authors of Stroud et al. (2010) found that by applying the EnKS to combine the model and data, the accuracy in the result sediment plume map was 20-30% better than standard approaches for combining the data and model.

As with the EnKS, the applications of the Particle Filter (PF) appear to have been less than the EnKF. A common model that the PF has been used with is the Lorenz model - Ades & van Leeuwen (2012), van Leeuwen (2010). The main utility of the Lorenz model (Lorenz, 1963) is that it is able to simulate the very non-linear and chaotic nature of the atmosphere but with a minimum of three states and a similar number of parameters, meaning that it is very fast to run. Unlike this application of the PF which involved synthetic observations, Quaife et al. (2008) used the PF to assimilate reflectance data from the MODIS sensor of NASA's TERRA satellite into the DALEC model (Williams et al., 2005). This built on previous work (Quaife et al., 2007) where the reflectance data was assimilated into DALEC using the EnKF. Another study (Hill et al., 2011) used the PF to assimilate satellite data of NDVI into a simple model. The data were very coarse, but they applied a method based on knowing the observational uncertainty of making the data more fine scale and assimilating this into the model. This resulted in improved estimates of NEE when compared to independent ground data.

APPLICATIONS OF GLOBAL SEARCH DA ALGORITHMS: THE METROPOLIS ALGORITHM AND THE GENETIC ALGORITHM

Metropolis et al. (1953), who first formulated the Metropolis algorithm, used it to det-

ermine the ideal pressure of a configuration of interacting molecules (simulated using a simple model) at a phase transition, given a specified temperature and a specified number of molecules. The configuration was optimal if the corresponding pressure and pre-determined temperature obeyed the ideal pressure law. The ideal pressure, based on a specific configuration could not be determined analytically because it required determining the expectation of the Boltzmann distribution, used to simulate the distribution of molecules, which involved an unsolvable integral. It was unsolvable because it contained an unknown normalising constant, and the exponential of potentially complicated Energy function.

Since this ground-breaking paper, which has been cited close to 20,000 times the Metropolis algorithm has been applied to a variety of different problems. These include applications to hydrology (Vrugt et al., 2003), Medicine (Gilks & Best, 1995), and geophysics (Mosegaard & Tarantola, 1995), among others. In terrestrial carbon modelling, Knorr and Kattge (2005) used it to estimate the probability distributions of the parameters of the BETHY (Biosphere-Energy-Transfer-Hydrology) model. BETHY simulates the carbon, water and energy exchanges between the biosphere and atmosphere. The prior values of the parameters and the measurements of NEE and Latent Energy (LE) were assumed to be Normally distributed. For the priors, the means were determined from expert knowledge and three sizes of variances were used as there was uncertainty about what these should be. After assimilating only 7 days of half-hourly NEE and LE measurements, the uncertainties of five of the 29 parameters were substantially reduced (compared to the prior distributions) and there was an improvement in the modelled NEE estimates over the 2 years that followed. A limitation with this study was the choice of the distribution to represent the observational errors (i.e. a Normal one), which conflicts with the findings from Hollinger & Richardson (2005) who suggested a double-exponential distribution for the random part of the observational errors, when the data are half-hourly.

A very interesting study was by Richardson et al. (2010). Here the parameters and uncertainties of a version of the DALEC model, operating on a half-daily time-step, were estimated using the Metropolis algorithm. Half daily data used to constrain the parameters included NEE, soil respiration, leaf area index (LAI), litterfall and woody biomass increment, and were collected over a 4 year period. The study showed that assimilating many different types of data, as opposed to just one, improved the

parameter estimates. More specifically, the results showed that assimilating measurements of woody biomass, and to a lesser extent soil respiration, in addition to NEE data gave marked reductions in the uncertainty of the resulting parameter estimates, compared to just assimilating NEE. The authors stated that this made sense because these first two data types provided orthogonal constraints to the NEE data. The authors also found that the assimilation of multiple data types improved the forecasts of the model using the calibrated parameter set.

As explained previously, the Genetic algorithm is similar to the Metropolis algorithm in that it explores the parameter space in a stochastic way, but how this is done is different (see subsection 2.2.9). Despite the Genetic algorithm being applied far less than the Metropolis algorithm, it has been used in a number of settings. One example is Van Wijk (2001) who used it to parameterise a model which simulated the vertical root distribution. The model also carried the assumption that the roots are structured so that the water uptake from the root zone is maximised. The model was calibrated under two different settings: a single plant and two neighbouring plants competing for the same soil water available.

APPLICATIONS OF GLOBAL DESCENT DA ALGORITHMS: 3D-VAR AND 4D-VAR

4D-Var is a DA scheme, which has been used for weather forecasting for almost two decades (Courtier, 1994), and has more recently been used with terrestrial carbon models. In particular, over the past few years it has been applied to the Joint UK Land Environment Simulator (JULES), which is the terrestrial part of the Met Office's Unified climate model (Luke, 2011). It has also been shown to give comparative results to the Metropolis algorithm. Scholze et al. (2011) used a 4D-Var scheme to estimate the probability density functions (pdfs) of the parameters of the BETHY model (where for each pdf the parameter value with the highest probability was chosen as its estimate). They found over 90% of these matched very closely the pdfs produced from using Metropolis algorithm. However the computation time of the 4D-Var approach was significantly less (1 hour) than that of the Metropolis algorithm (8 months). However, the 4D-Var approach required the model to be differentiated which can take a significant amount of time in itself.

INTERCOMPARISON DA STUDIES

The REgional Flux Estimation eXperiment or REFLEX (Fox et al., 2009) compared different DA schemes' estimates of parameters and future states of the same model (DALEC), after assimilating the same 2 years of NEE and LAI in situ data. Eight DA schemes were used, with five being variants of the Metropolis algorithm, one using the Genetic algorithm, one using the EnKF, and one using the Metropolis algorithm and the EnKF. The participants were asked to estimate the parameters and future model states using real and synthetic data, where the synthetic data were created by a running the model forward and adding noise and gaps to the NEE and LAI. The uncertainty reduced for parameters linked to GPP and respiration e.g. the allocation and turnover of foliage and temperature sensitivity), as a result of assimilating the NEE and LAI data. Poorly estimated parameters were those related to the allocation to and turnover of fine root/wood pools, which suggested that assimilating observations of fine roots and woody biomass may help to constrain the parameters connected to these the fine root and woody carbon pools. An interesting feature of the results was that the DA schemes gave different estimates to around half of the parameters, including the different variants of the Metropolis algorithm.

The Optimisation InterComparison or OpTIC study (Trudinger, 2007) also compared different DA schemes when estimating the parameters of a very simple terrestrial carbon model. The schemes included three variants of the Metropolis algorithm, the Kalman filter and extended Kalman filter, 4D-Var, and the Genetic algorithm. The major finding from this study was that the variation in the parameter estimates was not due to the DA scheme, but to the choice of cost function, which defines the mismatch between the model and observations. More specifically, if the cost function is specified by equation (2.3.1) (where $x_1(t_i)$ and $x_2(t_i)$ are the observations at time t_i with $z_1(t_i)$ and $z_2(t_i)$ being the corresponding modelled values), the authors found that if weights w_1 and w_2 are not constant (i.e. vary with time) and vary with noisy observations, this leads to biased parameter estimates.

$$\Theta = \sum_i \left(\frac{(x_1(t_i) - z_1(t_i))^2}{w_1} + \frac{(x_2(t_i) - z_2(t_i))^2}{w_2} \right)$$

A limitation of OpTIC was that the model (which only contained two states and four parameters) was arguably too simplistic. In addition, OpTIC did not use real data, and

so did not evaluate the calibrated model's ability to accurately estimate current or future states of the system.

As part of an EU funded project called 'Carbon Assimilation and Modelling of the European Land Surface' (CAMELS), one of the aims was to produce a Carbon Cycle Data Assimilation Scheme (CCDAS) in order to improve understanding of the current carbon cycle and to provide predictions for future changes. CCDAS has been active for over a decade and has addressed the following questions:

- Where are the current carbon sources and sinks located on the land and do European sinks compare with sinks of other large continental areas?
- How could we make optimal use of existing data sources and the latest models to produce operational estimates of the European land carbon sink?

To address these questions, the BETHY model (Knorr, 2000) has been used consistently in CCDAS, assimilating data from a range of sources (e.g. flux measurements, carbon inventory data, satellite products). The Metropolis algorithm and 4D-Var scheme have been predominantly used in numerous CCDAS studies with the focus primarily on reducing the uncertainties of the parameter estimates. Earlier CCDAS studies primarily used ground data (Knorr & Kattge, 2005), but more recently satellite observations, mainly the fraction of photosynthetically active radiation (faPAR) have been used. The achievements of CCDAS have been to reduce the uncertainties of BETHY's parameters and the resulting estimates of the current and future states of the ecosystem.

EOLDAS (Earth Observation Land Data Assimilation Scheme) is a European Space Agency project (Lewis et al., 2012), with the aim of better using earth observation data in DA. Particular focus is on the use or low-level satellite products, or satellite data which is closer to the raw radiance data. One of the main reasons for this is that it is easier to quantify the observational uncertainty of such products, compared to high-level ones such as Leaf Area Index (LAI) (Quaife et al., 2007). Furthermore, certain climate variables, LAI being a good example, are inferred by the data from a number of different satellites, but many the algorithms used to infer LAI (for example) often carry different assumptions and formulations, resulting in different estimates of LAI (Garrigues et al., 2008). To therefore make full use of satellite derived data, EOLDAS is taking a lead role in testing the utility of assimilating low-level satellite products.

2.3.2 Gaps in our current knowledge of DA and choice of DA algorithms

Despite the extensive research carried out on DA applied to terrestrial C modelling, various issues remain unanswered or unresolved. Among others, the REFLEX project (Fox et al. 2009) raised some fundamental questions about DA:

- Why did the parameter estimates vary among the different DA schemes, even amongst the Metropolis based ones? The choice of cost function may be partly responsible, as suggested by the OpTIC study. However there may be other reasons, such as the initial conditions of the model.
- How do different DA schemes respond if the model suffers from equifinality, where different sets of parameters can result in almost identical or very similar model outputs?
- Are all parameters fixed or should some have different values for different times in the year? If parameters are not fixed, how do DA algorithms estimate these parameters and do they give similar estimates?
- Does the type of prior (e.g. Uniform, Normal) influence DA?

Williams et al. (2009) who reviewed the knowledge gained with DA in the terrestrial C modelling setting, also stated some of the challenges that remain. These included:

(1) To explore the effect of assimilating 10+ years of data?
(2) To avoid confounding effects on missing processes in model representation on parameter estimation.
(3) To assimilate more types of data (e.g. pools/stocks of carbon, earth observation data) and to define improved observation operators.
(4) To fully quantify uncertainties arising from data bias, model structure and estimates of initial conditions.

For (1) this is of particular importance because different DA studies, using terrestrial ecosystem models, have assimilated different numbers of years of observations; for example, the REFLEX study used 2 years (Fox et al., 2009), Quaife (2007) used 3 years, Mo et al. (2008) used 8 years, and Braswell (2005) used 10 years. Since more information is assimilated, longer datasets might be expected to lead to more accurate parameter estimates. Other aspects of the observations might also be important, such as the size of the observational uncertainty and the size and frequency of data gaps, both of which are features of real datasets. In all DA studies these have been kept

fixed. Luo et al. (2011) also stated that not many DA studies have explicitly addressed issues directly pertinent to forecasting, such as forecasting accuracy, sources of uncertainty and usefulness of forecasting under different domains.

A key aspect of CCDAS is the assimilation of satellite observations, such as the fraction of Photosynthetically Active Radiation or fAPAR. A satellite observation closely linked to fAPAR is Leaf Area Index (LAI). LAI is an essential climate variable as it is an indicator of the health of vegetation, and is used to estimate evapotranspiration and the accumulation of carbon from photosynthesis (Sellers, 1997). However, LAI measured from satellites can have unrealistic features, such as excessive temporal noise and unrealistically small prescribed errors (Gao et al., 2008). A key issue, which has not yet been answered, is whether DA is robust enough against these unrealistic features affecting the modelled states after assimilation (Guanter et al., 2012).

Based on the gaps in knowledge of DA applied to terrestrial C modelling the focus of this thesis is, put simply, an exploratory analysis of how DA works in different situations, highlighting where it performs well and is robust and identifying and quantifying those situations where it there are weaknesses and important limitations. This thesis will help indirectly with better representation and quantification of the terrestrial-atmosphere C flux by highlighting how DA should be optimally used. The precise aims and the rationale for these are given in section 2.6.

CHOICE OF DA ALGORITHMS

For now however, a decision is made as to which DA algorithms should be used to address some of these issues. It is decided that two algorithms are to be used, and these are the Metropolis algorithm and the Ensemble Kalman Filter. There are several reasons for this. First of all, both of these schemes have been used very widely used not only for terrestrial C modelling but in a variety of other fields. Although Williams et al. (2009) identified three broad DA groups, namely global search algorithms, global descent algorithms and sequential algorithms, really there are two: batch methods (under which the first two come) and sequential methods. For batch methods, all of the data is used at once and a number of executions of the model are required. For sequential methods on the other hand, each element of the dataset is assimilated one at a time and only one execution of the model is required. The

Metropolis algorithm is preferred over other batch methods because it is the most widely used of all global search algorithms and unlike global descent algorithms it is able to determine the full posterior probability distribution. Although variational approaches such as 4D-Var are more computationally efficient, they typically only converge to the modal point of the posterior probability distribution which may also not be globally optimal. A measure of the spread of the distribution can be approximated but only under assumptions, such as the distribution being Gaussian, which may not be true. Proper quantification of the posterior distribution is essential, and this can only be guaranteed with the algorithms like the Metropolis algorithm. The EnKF is preferred to the EnKS mainly because it is more straightforward to code up, but also when performing parameter estimation it has been previously stated and referenced that the results are equally globally optimal.

In the next three subsections that follow, some preliminary tests are carried out on the Metropolis algorithm and the EnKF. These are carried out to ensure that these two DA schemes are properly working as they should, so as results obtained from using these algorithms in the chapters that follow can be properly interpreted. The next three subsections are organised as follows: in subsection 2.3.3 the preliminary tests are carried out on the two algorithms. Then in subsection 2.3.4, the expectations for the Metropolis algorithm and the EnKF when estimating parameters are discussed, which will be particularly relevant in relation to the discussion section in chapter four.

2.3.3 Basic tests of the algorithms

The purpose of this section is to do some preliminary tests on the Metropolis algorithm and the Ensemble Kalman filter to ensure that they are operating correctly.

For the Metropolis algorithm this was done by assimilating a synthetic dataset and then, based on the examiner's comments of the original thesis that was submitted, initialising the chains from the true parameter set and using very small uncertainty on the observations and the prior distribution of the parameters. A dataset of synthetic observations means that observations that have been created from the output of a model (e.g. a time series of Net Ecosystem Exchange or NEE output) with a noise from a specific probability distribution (e.g. Gaussian) added to the NEE to represent observational uncertainty; this could be based on the observational noise characteristics from a particular site. A certain number of the synthetic observations

in the dataset are also removed to reflect that in real life observations are not present at each time-step, either because they were not measured or inferred from other measurements, or they were removed for other reasons such as known biases (e.g. instrumentation error). For ease of implementation, an existing dataset which is also used in chapter four was used here. For reference, the dataset corresponded to run 1 from table 4.2 in subsection 4.2.8. This dataset is an NEE dataset of one year in length, generated from a forward run of the DALEC-D model (see subsection 2.4.4 for description) and has observational noise and gap characteristics based on of Harvard forest (Munger & Wofsy, 1999). For the purposes of the test carried out in this subsection, this synthetic NEE dataset was recreated with an observational uncertainty based on a zero mean Gaussian distribution with a smaller standard deviation of 0.05 $gC/m^2/day$. The prior distribution for the log-Normalised parameters used in the parameters was also set to 0.05, much smaller than was used for the chapter four runs of the Metropolis algorithm. By initialising the chains from the true parameter set and by using very small uncertainty for the assimilated NEE data and prior distribution for the parameters, we expected the chains to converge almost straight away and in particular to remain or converge to a the true parameter set (the parameter set from the DALEC-D model used to create synthetic data before noise and gaps were added) or within a very close vicinity.

Figure 2.5 shows the values of Rhat, the diagnostic used in the Gelman criteria to assess convergence. We can see from the bottom panel of this figure that the chains for all the parameters had converged by 50,000 iterations, since this was the point by which all the corresponding Rhat values went below 1.2, the vast majority getting below this threshold a lot sooner. From the top panel we can see that the Rhat values became more and more close to the 1 as the number of iterations increased. In figure 2.6 we can see the estimate of the model parameters from this test run of the Metropolis algorithm. The estimate is the modal value of the posterior distribution and the lower and upper limits of the associated error bars represent the 5^{th} and 95^{th} percentiles. The prior mean/modal values of the parameters with the lower and upper limits of the errors bars also representing the 5^{th} and 95^{th} percentiles of this distribution are also given. The red horizontal lines represent the locations of the true parameter values. For 21 out of 23 of these parameters, the posterior parameters are on or very close their true values with the uncertainty intervals containing or very near-

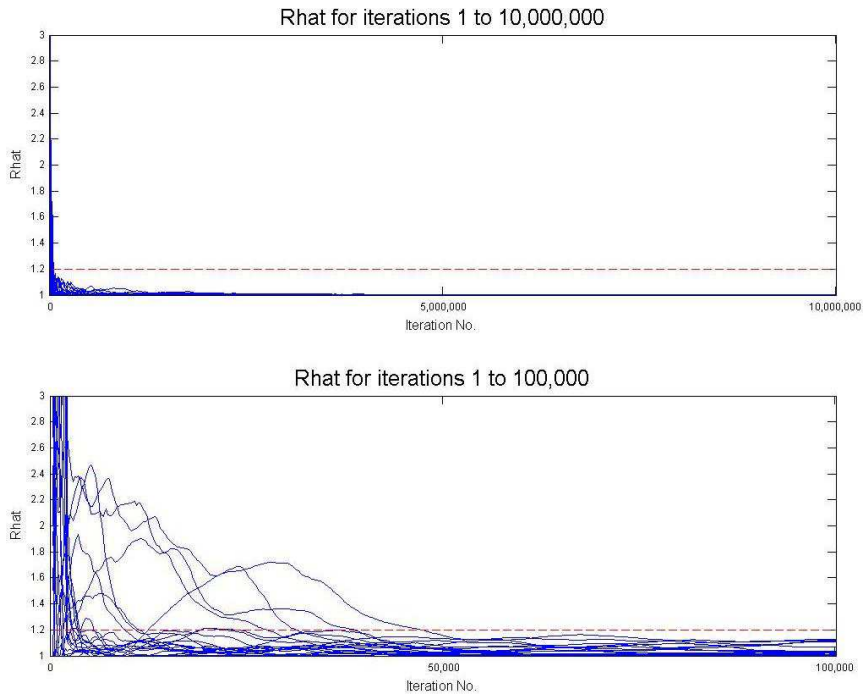


Figure 2.5 The time-series of Rhat values for each of the 23 model parameters to assess convergence of the three chains used in the Metropolis algorithm test run.

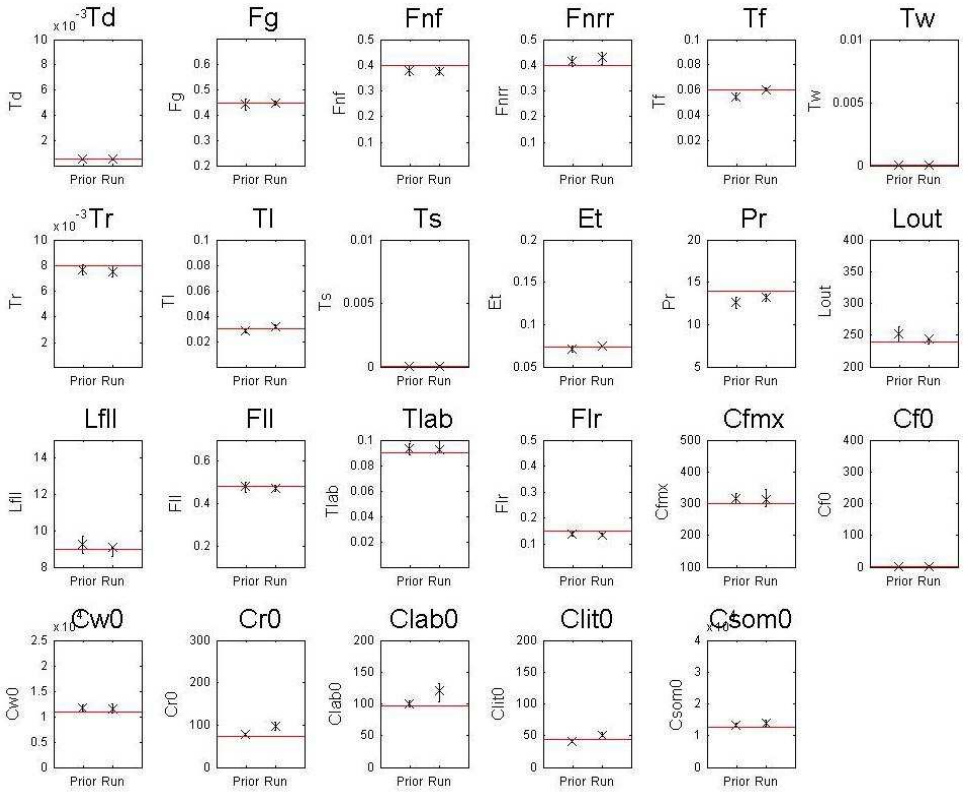


Figure 2.6 The MCMC prior and posterior estimates, with associated uncertainties, of each of the 23 DALEC-D parameters. The red lines denote the true parameter values.

ly containing the true values. In many of these cases, the parameters are closer to the true value compared to the prior estimate. For the remaining two parameters, namely $C_r(0)$ and $C_{lab}(0)$, the posterior estimates are noticeably further away from the true value compared to the prior values. If a parameter or parameters are not constrained by the assimilated data, the posterior estimate of the parameter should revert to its prior value. However correlations exist between the parameters and given that the prior uncertainty is very small it is not zero, and so it is not uncommon that this should happen. Of all the parameters it should happen to, it is almost expected it should occur with the ones that control the initial conditions of the two Carbon pools which are the least removed from the predicted NEE, namely the root Carbon and labile Carbon pools (see subsection 2.4.3 for explanation of model).

Given the results of this test run, the Metropolis algorithm is operating as it should. In figure 2.5, we can see that it has converged very fast as we would expect from starting from the true parameter set and virtually all of the posterior parameter estimates are on or very close to the true parameter values with associated uncertainties that include or very nearly include the true values.

The Ensemble Kalman was next run using the same dataset and setup as the Metropolis algorithm above. As can be seen in figure 2.7, the parameter estimates are on the whole very close to the corresponding true values, as we saw with the Metropolis algorithm. However for 3 of the parameters, namely Lout, Flr and Cfmax, the posterior estimate is either noticeably further away from the truth than the prior value or the posterior uncertainty interval is noticeably wider than that of the prior interval. In one sense this might be surprising given that in chapter 4 we find that the EnKF parameter estimates appeared to be the whole closer to the truth with narrower posterior intervals containing the truth compared to the Metropolis algorithm. However, given the improvement in parameter estimates and uncertainties with dataset length from the chapter 4 results, including from 1 to 2 years, it is likely that if this test run had been carried out using a 2 year dataset the parameter estimates and uncertainties for these three parameters would be significantly improved.

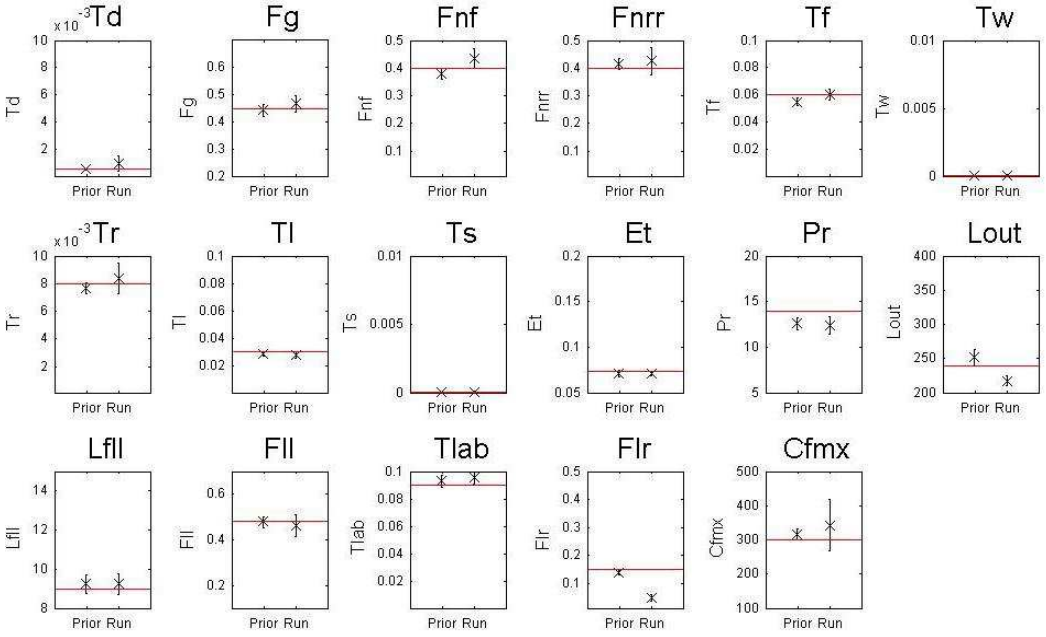


Figure 2.7 EnKF prior and posterior estimates along with associated uncertainties, of each of the 17 DALEC-D parameters. The red lines denote the true parameter values.

2.3.4 Expectations of the EnKF and the Metropolis algorithm for estimating parameters and associated uncertainties

The Metropolis algorithm and the Ensemble Kalman filter are two different types of DA algorithms, with the former being a global search algorithm while the latter is a sequential scheme. In this way, when they are used to estimate parameters and the associated uncertainties, they go about this in different ways, yet in both cases their parameter estimates are globally optimal. This has been shown in their corresponding descriptions in subsections 2.2.4 and 2.2.8 of chapter 2. In chapter 4 of this thesis, the two algorithms are used to estimate the parameters and associated uncertainties of the DALEC-D model. In this section, we describe the slight differences in the assumptions for both schemes when they are implemented, and whether it is likely that these are large enough to result in non-insignificant differences in the resulting parameter estimates. We also discuss more generally when we might expect the two algorithms to give similar results under identical or almost identical setups and when we might expect to see differences.

In chapter 4, where the two algorithms are used to solve the same problem, both are setup in the same way except in two very minor ways. These include (i) a slightly different prior distribution on the parameters and (ii) slightly different model

errors specified. For (i), the Ensemble Kalman filter used a Gaussian prior distribution for the model parameters, while the Metropolis used a log-Normal distribution. However the parameters used in the Metropolis algorithm, referred to as log-Normalised parameters (see subsection 2.2.8), do have a Gaussian prior distribution. Furthermore, the modal value and standard deviation of prior distributions of the model parameters between the two algorithms are virtually identical. For these reasons, we would not expect there to be any differences in the parameter estimates and associated uncertainties due to prior distributions. For (ii), the EnKF has a relative model error specified for each element of the state vector (expressed in terms of a standard deviation) as 0.32% for the parameter elements and 0.1% for the state elements of the augmented state vector. This model noise is added to each ensemble member of the state matrix for each day the model is run forward between observations. The Metropolis algorithm has no model error added onto at the daily time-step. When there are large gaps between the observations in the dataset, the aggregated model noise can become non-insignificant, though not large. In chapter 4, 13 of the 15 runs assimilate datasets with at least 60% of daily data present, and so there are never large enough gaps between the data for the model error to grow anything beyond insignificant, so this is not an issue here for the most part. More importantly, this means that it is unlikely to be a reason for non-insignificant differences between the parameter estimates and associated uncertainties for the two DA schemes.

Evidence for these two slight differences between the two algorithms not resulting in noticeable differences for the parameter estimates can be seen from the REFLEX experiment (Fox et al., 2009) when these exact same two algorithms with exactly the same setups were used to estimate the same parameters to the same model (DALEC-D) used in chapter 4. From figure 2.8, we can see that for 13 of the 17 DALEC-D parameters, the estimates for the two DA schemes are very similar. For each of the remaining 3, there is some difference between the two estimates but not a lot. Interestingly, the lengths of the 5^{th} to 95^{th} percentile intervals are either very good or good agreement for 15 of the 17 parameters. Based on this evidence, it is expected that for chapter 4, the EnKF and Metropolis algorithm will give a similarly good agreement in their parameter estimates and uncertainties when assimilating a dataset

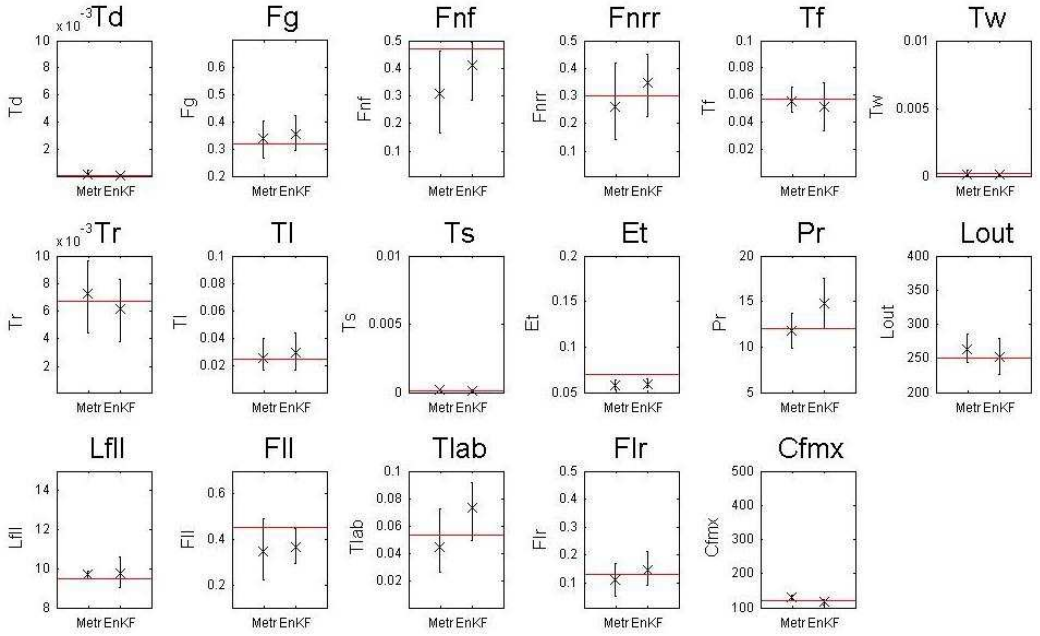


Figure 2.8 The estimates and uncertainties of each of the 17 DALEC-D parameters for the Metropolis algorithm and the Ensemble Kalman Filter used in the REFLEX project (Fox et al., 2009). The red lines denote the true parameter values.

with similar characteristics. In REFLEX, the assimilated NEE dataset was 2 years in length with observational noise and data density equivalent to the central levels used in chapter 4. In other words, the dataset used here is most similar to that of run 2 from chapter 4 (see table 4.2). It is worth noting here that synthetic Leaf Area Index data were also assimilated in REFLEX, however for the 2 year period or 731 individual daily time points the datasetcovered there were only around a dozen LAI data points, thus their influence on the parameter inversions compared to the NEE would have been much less. Therefore this inversion setup used in REFLEX is almost the same as run 2 from chapter 4.

Based on this result from REFLEX, it is reasonable to expect the parameter estimates between the two algorithms to be similar when assimilating shorter datasets, i.e. 5 years or less in the case of the chapter 4 experiments with noise and data densities at their central levels as in REFLEX, i.e. runs 1 to 3 from chapter 4 (see table 4.2). Virtually no terrestrial Carbon DA study has assimilated a dataset 10 years or more, or under varying observational uncertainty and varying data density, so it is difficult to say with confidence whether these would result in differences in the parameter estimates and uncertainties between the EnKF and the Metropolis algorithm. Therefore, we instead explore known factors which might result in

differences and whether we might expect these factors to occur when assimilating longer datasets or when varying the observational uncertainty or data density.

The issue of equifinality has been identified as a potential issue for the Metropolis algorithm by many studies, including Medlyn et al. (2005), Tang & Zhang (2008), Williams et al. (2009) and Luo et al. (2011). This is when different sets of parameters result in equally good fits of the model output (NEE in the case of chapter 4) to the (NEE) data being assimilated. A possible cause for this is the existence of 'redundant' model parameters which have little or virtually no impact on the shape of the modelled NEE's trajectory. As a result, these parameters can be set at other values whilst still resulting in a good fit of the modelled NEE to the NEE data. In the studies listed above, equifinality has been identified to occur when assimilating datasets of 5 years, but the extent to which it could occur when assimilating datasets that are longer or have different noise and gap characteristics is unknown. One could expect equifinality to occur at a similar extent under these differing characteristics of the data, as the different dataset would all contain roughly the same important information such as the seasonal and inter-annual variation, and the peak to peak amplitude.

One way of limiting the effects of equifinality is to assimilate multiple types of observations as in Richardson et al. (2010). In this way, these 'redundant' parameter which control different model processes are better constrained. Medlyn et al. (2005) also states that equifinality can occur as a result of model structure. For example, suppose an inversion algorithm is used to estimate the model parameters to two models with different assumptions about the ecological model – for example, Medlyn et al. (2005) uses two models of soil respiration, one based solely on soil temperature, while the other is based on the turnover rates of two carbon pools. Medlyn et al. (2005) shows that under both hypotheses of the model structure that the modelled output using parameter values estimated by an inversion algorithm is of a similarly good fit to the data. The Ensemble Kalman filter also likely to suffer from equifinality due to model structure, but its impact is likely to be less than the Metropolis algorithm because it is a sequential DA scheme. As a result, the EnKF does not try the find best fit of the modelled and observed NEE for the entire dataset, as batch methods like the Metropolis algorithm do. Instead the EnKF compares the modelled and observed NEE one time-step at a time. As a result, this may cause differences in the parameter estimates between the two DA schemes in chapter 4.

Another plausible reason for potential differences occurring in the parameter estimates between the two DA schemes could be due to the observational error being temporarily correlated (Trudinger et al., 2007). In chapter 4, although the synthetic NEE data are created by adding independently distributed Gaussian noise terms to the 'true' NEE, the resulting residuals between noisy NEE data and the modelled NEE, using a parameter set from the posterior distribution is not guaranteed to be independent in time. However, this temporal independence of the residuals is not guaranteed but it is a requirement of the likelihood function. Breaking this independence assumption of the likelihood may therefore bias the parameter estimates from the Metropolis algorithm, or cause an unrealistic representation of the posterior distribution. Autocorrelation may also exist in the residuals for the EnKF, despite the sequential nature of it and the data being assimilated one time step at a time, and may also cause problems. If this is a problem, one way of overcoming it is to modify the likelihood function to take account of this. However, this may be challenging or not possible for the EnKF due to the restriction of the prior distribution being Gaussian. In chapter 4 (or chapter 3), no modifications to the likelihood function were made primarily because we are interested in seeing the effects of this problem on assimilating datasets with different lengths, noise and gap characteristics.

2.4 Models used

In subsection 1.2.3 of chapter 1, an overview of 12 different models, varying in size and complexity, was presented. In this section, we form arguments for the choice of model to be used in this thesis. Before this, we go into more detail about common approaches to how some of the key ecological processes are represented in the models. To avoid this subsection (2.4.1) being overly lengthy, we refer to a sample of the 12 models outlined in subsection 1.2.3 of chapter 1. The purpose of this is mainly to gain more of an appreciation for how differently models simulate these processes, and to inform which model to adopt for this thesis. In subsection 2.4.2, the arguments are then laid out to which exact model is most appropriate in the context of the aims of this thesis. Finally in section 2.4.3 the model to be used is described in complete detail.

2.4.1 How models simulate the key ecological processes in terrestrial C dynamics

In this subsection, we give a basic overview of how a number of key ecological processes are represented a sample of the 12 models described in brief in subsection 1.2.3. We focus predominantly on 4 models, but the choice is arbitrary and based on gaining an appreciation for how these processes are represented in smaller models versus larger models, where due to the differences in complexity there appears to be the largest gulf. The four models include two small ones, namely DALEC (Williams et al., 2005) and SiPNET (Braswell et al., 2005), and two larger ones, namely SDGVM (Woodward & Lomas, 2004) and LPJ (Sitch et al., 2003).

GROSS PRIMARY PRODUCTION (GPP)

A common approach to modelling the assimilation of carbon into the plant at the leaf level due to photosynthesis is done the theory from Farquhar et al. (1980). This simulates the process of electron transport, the Calvin cycle and photorespiration. Farquhar et al. (1980) models the assimilation of carbon as the minimum of three rate limiting factors. These include: (i) the level of solar radiation and electron transport; (ii) the ratio of the rate of carbon and oxygen fixation by RuBisCO; and (iii) the export rate of photosynthate. The net rate of CO_2 assimilation from Farquhar et al. (1980) is represented by the equation:

$$A_b = V_c\left(1 - \frac{0.5P_o}{\tau P_c}\right) - R_d \qquad (2.36)$$

where A_b is the total carbon assimilated by the leaf (μmol m^{-2} s^{-1}), V_c is the rate of carboxylation, P_o and P_c are internal partial pressures of O_2 and CO_2 respectively, τ is the specificity factor of RuBisCO for CO_2 relative to O_2, and R_d is the rate of respiration in light due to processes other than photorespiration. The three limiting factors to photosynthesis are represented in V_c which is equal to $\min\{W_c, W_j, W_p\}$, where, to be more precise than (i), (ii) and (iii) listed above, W_c, W_j and W_p are the rates of carboxylation limited by RuBisCO, the rate of RuBP regeneration, and triose phosphate utilization, respectively (Woodward et al., 1995).

SDGVM uses this approach by Farquhar et al. (1980) to simulate photosynthesis, whereas LPJ uses the representation by Collatz et al. (1992), which is a modification of the Farquhar model. The main difference is how V_c is estimated – as

stated above, Farquhar et al. (1980) estimates this as the minimum of W_c, W_j and W_p. Collatz et al. (1992), on the other hand, smoothes the minimization of these three rates, resulting in a co-limitation where two of the limiting rates are similar. The Collatz et al. (1992) model has become common as it is a more generalized version for global modelling purposes (Sitch et al., 2003). However, Bonan et al. (2011) has shown that the co-limitation in the representation by Collatz et al. (1992) can have a significant effect on the estimates of GPP.

Water stress can have a large impact on the assimilation of carbon into the plant, and this is incorporated into SDGVM by multiplying stomatal conductance and the RHS of equation (2.36) by a constant β. In LPJ, the effect of water stress on the plant is simulated in a more sophisticated way. If the atmospheric demand for water through transpiration is greater than the supply of water to the plant or the soil, then the stomatal conductance and assimilated carbon estimates are adjusted to be consistent with the transpiration rates.

SiPNET estimates GPP using the idea of limiting factors on carbon assimilation, but does it in a more simplistic way. Using Braswell et al. (2005), it is estimated to be:

$$GPP = GPP_{max} \times D_{temp} \times D_{VPD} \times D_{light} \times D_{water}$$

where GPP_{max} is the maximum daily GPP value, and the remaining terms of the RHS are fractions representing the limiting factors on GPP (equalling 1 when not limiting GPP, less than 1 otherwise) due to temperature, Vapour Pressure Deficit (VPD), the amount of absorbed light, and the water availability. Each of D_{temp}, D_{VPD}, D_{light} and D_{water} are determined using empirical functions. For example, D_{temp} is assumed to be a parabolic function, with a value of 1 when the air temperature (T_{air}) is equal to the optimum temperature (T_{opt}) and a value of 0 when $T_{air} > T_{max}$ or $T_{air} < T_{min}$. It is given by: $D_{temp} = \mathrm{Max}\left(\frac{(T_{max}-T_{air})(T_{air}-T_{min})}{((T_{max}-T_{min})/2)^2}, 0\right)$. GPP_{max} is given by:

$$GPP_{max} = A_{max} \times A_d + R_{f,o},$$

where A_{max} is the maximum photosynthetic rate in the early morning, when this rate is at its peak. A_d is a scaling factor to account for the fact that the maximum rate over an entire day is lower. $R_{f,o}$ is the foliar maintenance respiration rate at an optimum temperature T_{opt} and is given by $R_{f,o} = R_F \times A_{max}$ where R_F is a parameter and A_{max} is the leaf-level maximum net instantaneous CO_2 assimilation rate which can be measured.

In DALEC, GPP is calculated using a sub-model called the Aggregated Canopy Model or ACM (Williams et al., 1997). The ACM is a highly simplified coarse-scale (temporally and spatially) version of the fine-scaled Soil-Plant-Atmosphere (SPA) model developed by Williams et al. (1996). As with many other similarly complex models, SPA simulates the assimilation of C into the leaf using the Farquhar model (Farquhar et al., 1980), leaf-level transpiration using the Penman-Monteith equation (Jones, 1992). The unique feature of SPA is in how it models stomatal conductance which links these two processes. In particular, the rate at which water can be supplied to the canopy is restricted by the plant hydraulics and soil water availability. The rate limits transpiration, causing the stored water to be used conservatively in the morning to delay the onset of stomatal closure in the early afternoon, resulting in the canopy maximising daily C assimilation. The ACM is a semi-emphirical model, in that it is based on the same biological processes to SPA in simulating GPP, but the relations are emphirically based. More specifically, the backbone of the ACM was three assumptions based on the model structure of SPA, namely: (1) plant photosynthetic capacity is linearly related to total foliar N; (2) metabolic and diffusive constraints are colimiting; and (3) incident radiation is hyperbolically related to GPP (Williams et al., 1997). A data-driven approach to the development of the model was also incorporated starting off with these relations. For example, for (1), this is first expressed as:

$$P_N = aNL$$

where N is the foliar N concentration, L is leaf area index, and a is a parameter. This equation was revised to incorporate the effect of temperature on the response of P_N in order to better represent the model behaviour of SPA. It was changed to:

$$P_N = a_1 NL\exp(a_2 T)$$

where T is the average daily temperature.

AUTOTROPHIC RESPIRATION

In both SDGVM and LPJ, autotrophic respiration is modelled by first separating it into maintenance respiration and growth respiration. The first type assists in plant metabolism and maintains plant function, while the second type provides carbon as the raw ingredient for growth. In both models, respiration from maintenance is determined by computing this separately for foliage, sapwood and roots, For example,

for SDGVM, respiration from roots and stems is calculated using $R = r\beta B(e^{0.02t})$, where R is the respiration, r is the proportion of biomass respired, β is a fraction between 0 and 1 which represents the limitation by soil water, B is biomass and t is temperature. After maintenance respiration has been subtracted from GPP, 25% of the remainder is taken as growth respiration, the cost of producing new tissues. What remains is NPP.

SiPNET represents autotrophic respiration as the sum of foliar and wood maintenance respiration. Growth respiration is not modelled separately, but assumed to be included as part of the calculations for maintenance respiration. Both foliar and wood maintenance (R_f and R_w respectively) are modelled using a Q10 function, more specifically:

$$R_f = R_{f,o} Q_{10,v}^{(T_{air} - T_{opt})/10} \quad \text{and} \quad R_m = K_A C_W Q_{10,v}^{T_{air}/10}$$

where $R_{f,o}$ is the foliar respiration rate at the optimum temperature (T_{opt}) for photosynthesis, K_A is a parameter, C_W is the amount of carbon stored in woody biomass and T_{air} is the air temperature.

DALEC, like SiPNET, also does not explicitly model growth respiration, but unlike SiPNET, DALEC does not separate maintenance respiration according to the stems or woody parts of the tree nor does it treat it as temperature sensitive. Instead, DALEC assumes that daily autotrophic respiration is a fraction of daily GPP. However in the deciduous tree version of DALEC, denoted DALEC-D, two extra terms are included when computing autotrophic respiration. They are temperature dependent linear functions of the foliar and labile carbon pools.

HETEROTROPHIC RESPIRATION

In SDGVM, soil respiration of C is controlled by temperature, water supply, the rate of supply and litter and its quality (Woodward & Lomas, 2004). According to Woodward & Lomas (2004), litter quality is calculated by the fixation of carbon and the initial uptake of nitrogen into leaves and wood. Litter supply is also a fraction of NPP. In LPJ, heterotrophic respiration is the sum of the carbon emissions from the litter pools, the intermediate and slow soil pools (Sitch et al., 2003). Most of the respiration comes from the litter pools, where 70% of the decomposed litter goes directly into the atmosphere. As with SDGVM, the litter decomposition rate is

temperature and soil moisture dependent, with both the temperature and soil moisture functions being empirical in nature.

In SIPNET, heterotrophic respiration is linearly dependent on the soil carbon content and varies according the soil temperature using a Q10 relationship and according to soil moisture. DALEC employs a similar temperature sensitive Q10 relationship for the proportion of soil carbon respired heterotrophically.

LEAF AREA INDEX

In SDGVM, Leaf Area Index (LAI) is determined as the minimum value determined by water uptake, transpiration and the allocation of fixed carbon. More specifically, it is $\min(NPP \times (1 - a_s - a_r) \times SLA, NPP \times a_r \times W_U, (P - E)/Vg)$, where NPP is Net Primary Production, a_s and a_r are the allocations of carbon to the shoots and roots, SLA is specific leaf area, W_U is the water uptake by roots, P is precipitation and E is evaporation, V is the vapour pressure deficit and g is the stomatal and boundary layer conductance (Woodward & Lomas, 2004). In LPJ, the LAI of an individual tree is calculated using the formula:

$$LAI_{ind} = \frac{C_f \times SLA}{CA}$$

where C_f is the mass of foliar carbon per unit volume, SLA is a plant functional type (PFT) specific constant, and CA is the stem diameter dependent crown area which is used to represent the fact that LAI varies in different parts of the tree. In SiPNET, LAI is estimated using the formula:

$$LAI = \frac{C_f}{LMA \times C_{frac}}$$

where LMA is leaf mass per area, and C_{frac} is another parameter representing the fractional C content of leaves. In DALEC, LAI is calculated using LAI = C_f /LMA (Williams et al., 2005), where LMA is Leaf Mass per Area and is the reciprocal of SLA (Larcher, 2003).

2.4.2 Choice of model

The choice of model is a two stage process. First of all, we need to decide upon what size and complexity of model is most appropriate for the purpose of this thesis. Then,

of the models available of that size and level of complexity, a decision is made about the precise one to choose.

Common features of large models (e.g. TEM, SDGVM, BETHY, LPJ, ORCHIDEE) are that they simulate the C dynamics of trees, N dynamics, and water cycle. Small models such as DALEC, SiPNET and OpTIC do this, although representation of these processes are more simplified, e.g. for DALEC, N and water availability are fixed at site specific constants. One could think of small models as good approximations or larger ones, containing the essential features of the larger models, but are a lot more computationally efficient. Another reason for their efficiency is that they typically operate over a small spatial scale – obviously if these small models were scaled up spatially, they would take longer. However, for this thesis, we are investigating the advantages and weaknesses of DA, and are not primarily interested in getting global or regional estimates of the quantities of interest – although in the larger context, this is the ultimate goal. Also, as one of the algorithms used in this thesis is the Metropolis algorithm, this is only suited to smaller models operating over a small or coarse spatial scale because of the large numbers of model runs that are required. Many other studies (e.g. Williams et al., 2005, Knorr & Kattge, 2005, and Richardson et al., 2010) have carried out DA on a model at a specific site or sites, and so there is lots of value and potential knowledge gained in doing this.

Of the small models, we choose among four potential ones, namely OpTIC, DALEC, SiPNET and SPA (see table 1.2, subsection 1.2.3) all of which typically operate on the site-level scale. There are other small models available, but these four have been commonly used in DA studies involving terrestrial C modelling (e.g. Williams et al., 2005, Quaife et al., 2007, Trudinger et al., 2007, Brawell et al., 2005, and Williams et al., 1997) and are fairly representative of the range of smaller models. For example, SPA has the most detailed representation of the key processes such as photosynthesis, but the complexity decreases for the others, starting with DALEC, then SiPNET and then OpTIC. Secondly, all four models simulate photosynthesis in different ways, and thirdly two simulate the water cycle, whereas the other two do not. In making the decision of which one of these four models to choose, we first exclude OpTIC. This is because it is too simplistic. It only has two state variables and four parameters. We also exclude SPA, partly because it operates at the half-hourly time-

step and this level of temporal scale is not necessary and could cause the model to be too computationally expensive (due to the excessive number of time-steps required) since we will require the model to be run over multiple years to assess inter-annual and seasonal variability. Furthermore, SPA is probably the most complex of all the small models and this may add excessive computing time. Of the two that remain, namely SiPNET and DALEC, these are different in various ways. We outline some of the main differences and decide upon DALEC for three reasons outlined below.

First of all, they model photosynthesis differently, but arguably DALEC's version is a more accurate approximation of the photosynthesis used in SPA (Williams et al., 1997) where photosynthesis is simulated using the Farquhar model. SiPNET on the other hand, simulates photosynthesis in much more of an empirical way which could be argued is less representative of the actual processes.

Second of all, SiPNET simulates the water cycle whereas DALEC does not. However, this is not necessarily a major advantage to SiPNET. This is because: (i) of all the time periods for the sites used in this thesis are not water-stressed; (ii) although the main effect simulating the water cycle will have on the model will be to influence the estimates of GPP and soil respiration, for GPP the photosynthesis sub-model used in DALEC has been shown to produce good estimates of GPP (Williams et al., 2005) even without simulating water movement. Also, although SiPNET also uses its water estimates to inform its estimates of soil respiration, because the sites are not water stressed this is likely to have only minimal effects on soil respiration estimates. Furthermore, soil respiration is most sensitive to variations in temperature (Lloyd & Taylor, 1994) which both models include.

A third reason DALEC may be preferable to SIPNET because SIPNET operates by solving the differential equations using an approximation. Because of this, the model code is fairly complex. DALEC on the other hand does not solve its equations, and in fact it has very many fewer lines of code to run, which are also much simpler. In this way, DALEC is much more practical and easier to use than SiPNET. DALEC can also be freely downloaded from the internet, whereas for SiPNET permission needs to be granted and from personal experience this took over a year to grant.

In the subsection that follows, the evergreen and deciduous versions of the DALEC model are described in full detail.

2.4.3 The DALEC model

The DALEC model (Data Assimilation Linked Ecosystem Model for Evergreen vegetation) (Williams et al., 2005) is a simple box model that tracks the journey of carbon (C) from photosynthesis through to respiration in Evergreen trees via C pools and respiration fluxes. Figure 2.9 shows it in diagrammatic form. It has 6 carbon pools (represented by the boxes in figure 2.9), namely C_f (carbon stored in foliage, i.e. leaves), C_r (carbon stored in fine roots), C_w (carbon stored in woody biomass, i.e. tree trunks, roots and major branches), C_{lit} (carbon stored in litter), and $C_{som/cwd}$ (carbon stored in soil organic matter and coarse woody debris). There is an additional pseudo-pool for Gross Primary Production (GPP), which represents carbon accumulated from photosynthesis. GPP is estimated on a daily time step by a sub-model called the Aggregated Canopy Model (ACM), which is a simplified version of the Soil-Plant-Atmosphere model (Williams et al., 1997). Williams et al. (2005) states:

> *The ACM is a big-leaf, daily time step model that estimates GPP as a function of LAI, foliar nitrogen, total daily solar radiation, maximum and minimum daily temperature, day length, atmospheric CO_2 concentration, soil–plant water potential, and total soil–plant hydraulic resistance. (Williams et al., 2005, p.95).*

In DALEC, LAI is linearly related to C_f , by the relation, $LAI = C_f/LMA$ (Williams et al., 2005), where LMA (Leaf Mass per Area, m^2/m^2) is a site specific constant. This feedback (represented by the blue arrow in figure 2.9) is present because, put simply, the more leaves you have on the tree the more photosynthesis will take place. Like LMA, foliar nitrogen (gN/m^2) is also a site specific constant and is included in the GPP function because plants need nitrogen to carry out photosynthesis (more specifically they need it to make amino acids, proteins and DNA). It is worth noting that although LMA and foliar nitrogen are treated as constants in reality they vary spatially within a site (Larcher, 2003), however given that DALEC operates on a coarse scale this assumption is valid. Furthermore, the creator of DALEC states that LMA is a parameter which can be determined in the field or from the trait database like TRY (Kattge et al., 2011). The other parameters in DALEC, determined by DA, are those that are poorly constrained by our understanding (Williams, M., 2012, [personal communication]). In the same way, foliar Nitrogen mass is another parameter which is kept fixed for a particular site, whereas if we were on a finer

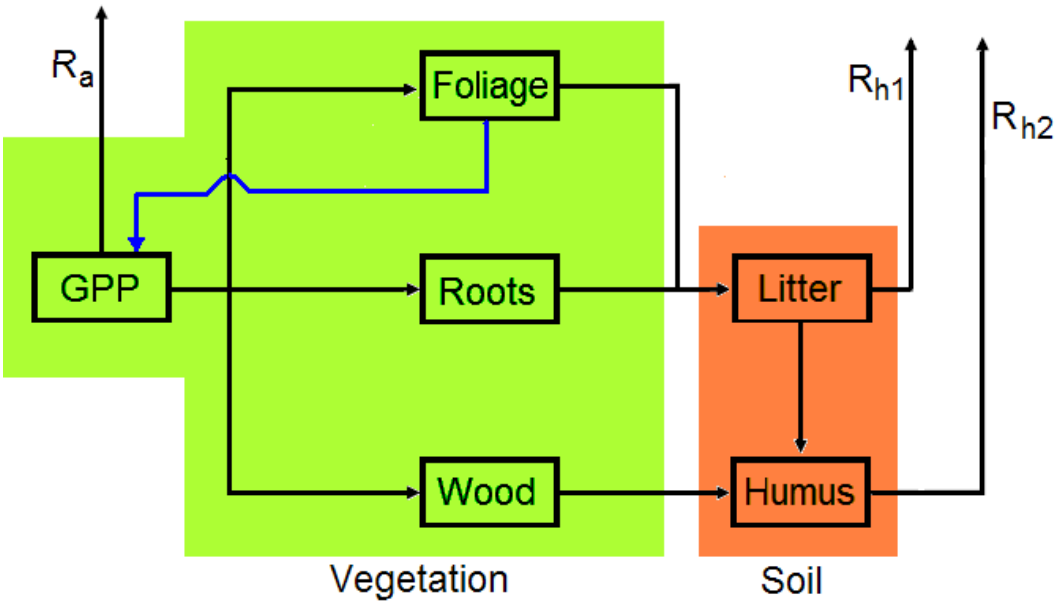


Figure 2.9 The DALEC model for evergreen vegetation, where the green and brown colours refer to vegetation and soil carbon stores.

resolution model then this would need to vary spatially (Williams et al., 1997). This assumption of treating LMA as a site specific constant has been done by a number of other papers, including most notably Williams et al. (2005) where the DALEC model was first described and introduced to the research community. Other papers where it appears include: De Kauwe et al. (2008), Fox et al. (2009), Quaife et al. (2007), Richardson et al. (2010) and Spadavecchia et al. (2011). The assumption of C_f being directly proportional to LAI is another assumption made which is reasonable given the fact that DALEC is a simple model. This relationship is used to provide estimates of LAI from the foliar Carbon pool of the previous time-step when GPP is estimates using the ACM. Since the ACM has been shown to very accurately replicate the SPA model in terms of its GPP estimates but at a much lower cost (Williams et al., 1997), it could be argued that Cf/LAI relation, used in the ACM, is a satisfactory enough approximation. Furthermore, this relation has been used in the first paper to use DALEC (Williams et al., 2005) and all papers to use the model since, such as De Kauwe et al. (2008), Fox et al. (2009), Quaife et al. (2007), Richardson et al. (2010) and Spadavecchia et al. (2011).

Despite the good arguments for treating LMA and foliar nitrogen as constants, there are perhaps stronger arguments for treating as parameters. In particular, if

DALEC is applied to a site or location where these two quantites have not been measured, using an estimate (e.g. from a nearby site or location where it has been measured), this may result in model biases. This is due to work by Chuter (2013) who found that DALEC modelled NEE is most sensitive to LMA and one of the exisiting DALEC parameters, above all the other model parameters. Therefore in hindsight, treating them as parameters even for sites where they have been measured would have been preferable to take account of the uncertainty in their true value arising from erroreous measurements or natural variability in time and space. Therefore, this is a limitation and it is recommended that future uses of the model LMA and foliar nitrogen should be treated as parameters.

The soil-plant water potential (ψ_d) and total soil-plant hydraulic resistance (R_{tot}) are site-invariant constants that are necessary for the estimation of GPP because they control how easy it is for the CO_2 to enter the leaf via the stomata (i.e. the pores of the leaf where CO_2 is absorbed into the plant) – this is referred to as stomatal conductance. The openings and closings of the stomata are controlled by guard cells. In essence if the leaves of the tree have a high stomatal conductance then the guard cells are wide open, and a low value implies they are not as wide open. ψ_d and R_{tot} are also important because they represent the water availability. Thus these two constants can be modified to simulate a drought or very wet conditions. Stomatal conductance is also controlled by the CO_2 levels in the atmosphere, the amount of solar radiation (MJ/m²/day), and the length of the day in hours (calculated using the latitude of the site and the day of the year). The greater the value of each of these the three quantities the greater the amount of photosynthesis occurs during the day. The total daily solar radiation and maximum and minimum daily temperature are provided as driving data for each day the model is run for.

There are a total of 10 fluxes (represented by the black arrows on figure 2.9), which connect the C pools and also provide a way for C to get into the system (via GPP) and out of it (via respiration). They are denoted by the letters A (Allocation), L (Loss), R (Respiration) and D (Decomposition). For example, A_f would be the allocation of GPP (after deducting R_a, autotrophic respiration) to foliage, L_w would be the loss rate of carbon from the C_w pool, etc. Finally, there are 11 parameters which control various aspects of the model, such as the size of the fluxes. These are displayed in table 2.1. They can be split into 4 groups:

(i) The respiration parameters (F_g, T_l, T_s) correspond to how much carbon is respired, which can be autotrophic, (i.e. respired from the leaves and represented as a fraction, F_g, of GPP), or heterotrophic (i.e. the fractions respired from the $C_{som/cwd}$ and C_{lit} pools, i.e. T_l, T_s).

(ii) The NPP allocation parameters, F_{nf}, F_{nrr}, are the fraction of NPP allocated to foliage, and the fraction of the remaining NPP allocated to roots. The NPP left over is allocated to the C_w pool.

(iii) The turnover parameters (T_d, T_f, T_w, T_r) are the fractions of C_{lit}, C_f, C_w, and C_r values which leave these pools. Note that T_d corresponds to the fraction of C going from the C_{lit} pool to the $C_{som/cwd}$ pool.

(iv) The miscellaneous parameters (E_t, P_r): E_t is referred to as the temperature sensitive parameter and is used to calculate the soil temperature (T) from the formula: $T = 0.5 \times \exp(E_t \times T_m)$ where T_m is the maximum minus the minimum daily temperature. This function is equivalent to a curve where an increase of T_m by 10ºC results in T doubling in size. T is used in the calculations of the size of the three C fluxes leaving the C_{lit} and $C_{som/cwd}$ pools by multiplying T by each of the respiration parameters. For example, $R_{h1} = T_l \times T \times C_{lit}$, whereas the other seven fluxes do not rely on T, e.g. the C flux coming out of the C_f pool (referred to as L_f) is calculated by $T_f \times C_f$. P_r is the nitrogen use efficiency parameter, and is used directly in the calculations for the ACM.

The values of the parameters and the initial conditions of the C pools are site specific and need to be estimated which in the past was typically done using expert knowledge or site inventory data, but more recently Data Assimilation (DA) has been used instead. To appreciate the size of these parameters, table 2.1 shows lower and upper bounds that were used as part of the REFLEX project (Fox et al., 2009) which included two different Evergreen sites. Recent work by Chuter (2013) has provided some useful insight into the values of the lower and upper bounds of each parameter based on physical limitations of the model. Chuter (2013) found that for certain values of particular parameters, the model resulted in tipping points where the terrestrial plant ecosystem went from a stable to an unstable environment; in practical terms, this often then leads to the ecosystem dying when in reality this may not happen for other physical reasons which are not represented by the model. This unrealistic feature of the model can be simply overcome by prescribing physical boun-

Symbol	Description	Units	Lower/Upper bounds
T_d	Litter decomposition rate parameter	day^{-1}	1×10^{-6} / 0.01
F_g	Fraction of GPP respired autotrophically	-	0.2 / 0.7
F_{nf}	Fraction of GPP allocated to foliage	-	0.01 / 0.5
F_{nrr}	Fraction of GPP allocated to roots	-	0.01 / 0.5
T_f	Turn over rate of foliage	day^{-1}	1×10^{-4} / 0.1
T_w	Turn over rate of wood	day^{-1}	1×10^{-6} /0.01
T_r	Turn over rate of roots	day^{-1}	1×10^{-5} / 0.1
T_l	Turn over rate of litter	day^{-1}	1×10^{-6} /0.01
T_s	Turnover rate of C_{SOM} pool.	day^{-1}	2.65×10^{-6}
E_t	Exponential temperature dependent rate parameter	-	0.05 / 0.2
P_r	Nitrogen use efficiency parameter	-	5 /20
$C_{f,0}$	Foliar carbon pool	gCm^{-2}	20 / 200 *
$C_{w,0}$	Woody C pool	gCm^{-2}	9200 / 16400 *
$C_{som/cwd}$	Soil Organic matter & coarse woody debris C pool	gCm^{-2}	9700 / 12000 *
$C_{r,0}$	Fine root C pool	gCm^{-2}	20 / 200 *
$C_{lit,0}$	Fresh litter C pool	gCm^{-2}	20 / 200 *

*Table 2.1 The description of the parameters and C pools of the DALEC model. The * refers to the initial values of the five C pools. The lower and upper bounds were those adopted for the REFLEX project (Fox et al., 2009) based on three Evergreen forest sites.*

daries to the model parameters; these boundaries may be further constrained for a particular location by other information such as local site data/knowledge.

As with many ecosystem models, DALEC is represented by a system of ordinary differential equations, and these are given below. The RHS of each of these differential equations is how much each of the C pools change by.

$$\frac{dC_f}{dt} = A_f - L_f, \quad \text{where } A_f = (1-F_g)F_{nf}\times GPP \quad \text{and } L_f = T_f C_f$$

$$\frac{dC_r}{dt} = A_r - L_r, \quad \text{where } A_r = (1-F_g)(1-F_{nf})F_{nrr}\times GPP \quad \text{and } L_r = T_r C_r$$

$$\frac{dC_w}{dt} = A_w - L_w, \quad \text{where } A_w = (1-F_g)(1-F_{nf})(1-F_{nrr})\times GPP \quad \text{and } L_w = T_w C_w$$

$$\frac{dC_{lit}}{dt} = L_f + L_r - R_{h1} + D, \quad \text{where } R_{h1} = T_l C_{lit} T \quad \text{and } D = T_d C_{lit} T$$

$$\frac{dC_{som}}{dt} = L_w + D - R_{h2}, \qquad \text{where } R_{h2} = T_s C_{som} T$$

Several assumptions are made in order to determine the fluxes in the model (Williams et al., 2005):

1. The mass balance of carbon is conserved. This means two things: (a) All C fixed during a day (i.e. GPP) is expended either in autotrophic respiration or allocated to one of the three plant tissue pools, C_f, C_w or C_r; (b) For all C pools except GPP, C going into pool = C going out of pool + growth of pool.
2. Autotrophic respiration is a constant fraction of GPP (Waring et al., 1998).
3. Allocations to vegetation pools are constant functions of NPP.
4. All C losses from the system end up in the atmosphere. In other words, it is assumed that all C exchanges are included in the model. In reality, some C might dissolve out of the litter pool into soil moisture and flow out of the ecosystem, but for simplicity, this is ignored.

In DALEC (as with other models) we refer to the time dependent model variables of interest, such as the 6 C pools and the 10 fluxes, as the model states. We define two additional states, called NEE (Net Ecosystem Exchange) and iNEE. NEE is total respiration (i.e. the sum of autotrophic respiration, R_a, and heterotrophic respiration, R_{h1} and R_{h2}) minus GPP. This is the quantity we are most interested in, since it represents the carbon balance of the forest. iNEE stands for integrated NEE and is the cumulative value of NEE from the first time point to the current time point.

2.4.4 The DALEC-D model

DALEC-D is an extension to DALEC, modelling the carbon dynamics of deciduous forests as opposed to evergreens. From figure 2.10 we can see that DALEC-D contains everything that DALEC does (figure 2.9) but includes an extra C pool, which stores labile carbon (abbreviated as C_{lab}) with connecting fluxes. This difference exists because deciduous trees shed their foliage in the autumn and regain them in spring, so they need this labile carbon pool (a reserve carbon pool) in order for the leaves to come through in spring. Evergreen trees (e.g. pine trees) do not shed their foliage in the winter months hence have no need for a labile carbon pool. C accumulates in the C_{lab} pool during the growing season, when a portion of the C_f pool gets transferred to the C_{lab} pool. The amount of C in the C_{lab} pool remains roughly constant until spring

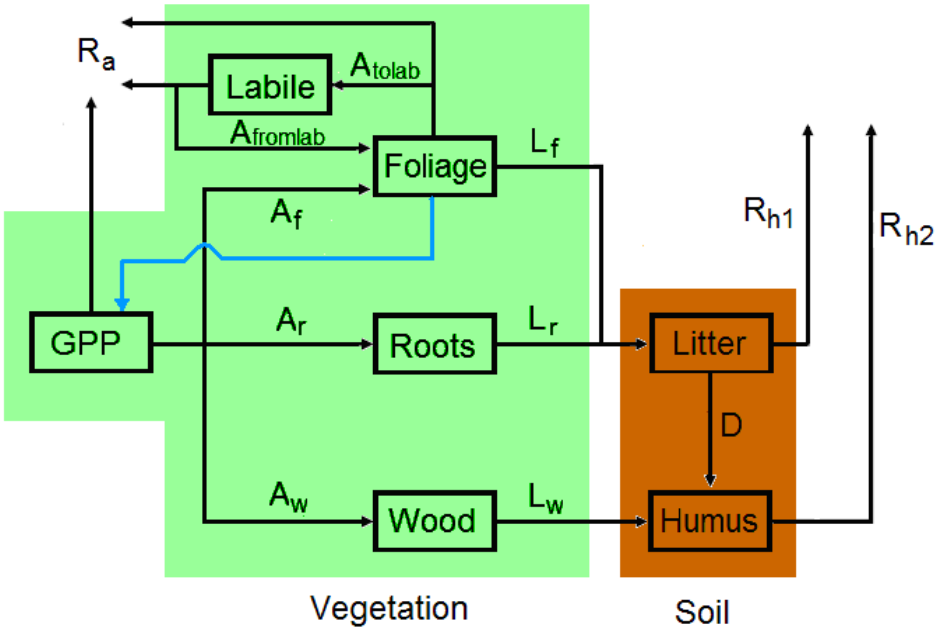


Figure 2.10 The DALEC-D model for deciduous vegetation, where the green and brown colours refer to vegetation and soil carbon stores

when it gets transferred back to the C_f pool in order for the spring bloom to occur. This flow of C between C_{lab} and C_f pools is represented in figure 2.10 by the A_{tolab} and $A_{fromlab}$ fluxes.

In addition to the 11 parameters used DALEC, DALEC-D has 6 others (given in table 2.2 along with the original 11) which are now explained. The timing of when the leaves start to grow in spring (represented by L_{out}) is controlled by a simple growing-degree day (GDD) accumulation function. GDD is a common tool (e.g. used by crop growers) to determine when the spring bloom occurs, as well as other rates of crop development, such as when a crop will reach maturity. The timing of when the leaves fall in autumn is treated as being after the 200th day of the year and only where the minimum daily temperature is less than a certain threshold (represented by L_{fall}). Of all the C that leaves the C_f pool (i.e. $F_{nf}\times C_f$), a fraction (F_{ll}) is transferred to the litter pool (i.e. from figure 2.10, $L_f=F_{nf}\times F_{ll}\times C_f$). T_{lab} is the turnover rate of the C_{lab} pool, which has the same interpretation as the T_f, T_r, T_w, T_l and T_s parameters. During spring, of the amount of C which leaves the C_{lab} pool ($T_{lab}\times C_{lab}$) a proportion (represented by F_{lr}) is respired autotrophically. Finally the timing of the start of summer is determined when either the size of the C_f pool exceeds a maximum threshold (represented by C_{fmax}) or the year-day is at least 200.

Symbol	Description	Units	Lower/Upper bounds
T_d	Litter decomposition rate parameter	day^{-1}	1×10^{-6} / 0.01
F_g	Fraction of GPP respired autotrophically	-	0.2 / 0.7
F_{nf}	Fraction of GPP allocated to foliage	-	0.01 / 0.5
F_{nrr}	Fraction of GPP allocated to roots	-	0.01 / 0.5
T_f	Turn over rate of foliage	day^{-1}	1×10^{-4}/ 0.1
T_w	Turn over rate of wood	day^{-1}	1×10^{-6} /0.01
T_r	Turn over rate of roots	day^{-1}	1×10^{-5} / 0.1
T_l	Turn over rate of litter	day^{-1}	1×10^{-6} /0.01
T_s	Turnover rate of C_{SOM} pool.	day^{-1}	2.65×10^{-6}
E_t	Exponential temperature dependent rate parameter	-	0.05 / 0.2
P_r	Nitrogen use efficiency parameter	-	5 /20
L_{out}	Growing degree days for leaf out	-	200 / 400
L_{fall}	Minimum temperature for leaf fall	^oC	8 / 15
F_{ll}	Fraction of foliar carbon lost transferred to litter	-	0.2 / 0.7
T_{lab}	Turn over rate of labile carbon	day^{-1}	1×10^{-4}/ 0.1
F_{lr}	Fraction of labile carbon respired	-	0.01 / 0.5
C_{fmax}	Maximum C_f	gCm^{-2}	100 / 500
$C_{f,0}$	Foliar carbon pool	gCm^{-2}	20 / 200 *
$C_{w,0}$	Woody C pool	gCm^{-2}	7100 / 12200 *
$C_{som/cwd}$	Soil Organic matter & coarse woody debris C pool	gCm^{-2}	8800 / 17600 *
$C_{r,0}$	Fine root C pool	gCm^{-2}	20 / 200 *
$C_{lit,0}$	Litter C pool	gCm^{-2}	20 / 200 *
$C_{lab,0}$	Labile C pool	gCm^{-2}	20 / 200 *

*Table 2.2 The description of the parameters and C pools of the DALEC model. The * refers to the initial values of the six C pools. The lower and upper bounds were those adopted for the REFLEX project (Fox et al., 2009) based on three Deciduous forest sites. The parameters and C pool used in DALEC-D but not DALEC are highlighted.*

2.5 Data

2.5.1 Introduction

In this thesis, we make use of data measured from ground measured or those inferred from satellite data. As outlined in chapter 1, it is key to have data in order to validate and calibrate models. There are lots of different types of data of different quantities, but there are various challenges or issues that need to be considered when using data, some important ones being: (i) Which is the best data to use and how much should be used? (ii) How do we trust the accuracy of data? (iii) How do we interpret data inferred from satellites?

For (i), one would expect that having more data is better because we have more information to describe what it is going on the real world. For example,

Williams et al. (2005) and Richardson et al. (2010) both calibrate the DALEC model with a whole variety of data including Net Ecosystem Exchange (NEE), woody C, litter C, soil respiration, Gross Primary Production (GPP), Leaf Area Index (LAI). In each case, the each set of authors find that the uncertainty in the modelled estimates of NEE progressively reduces as more types of data are assimilated into the model. Clearly this is expected, because we able to constrain more of processes associated with plant dynamics and growth. In this thesis, only data of NEE (or synthetic data based on the characteristics) and LAI are used and we do not make use of this variety of different data streams because of the specific aims of this work. This is because chapter 3 was motivated by trying to understand some of the results of the REFLEX project (Fox et al., 2009) and part of the same data used there was used here too, and this consisted on synthetic NEE and LAI data. In chapter 4, the aim was to understand how parameter estimates vary with the length, noisiness and density of the dataset being assimilated. We restricted ourselves to NEE data only partly because for the network of FLUXNET sites this is the most continuous type of data available and it is also the main quantity of interest (Baldocchi, 2008). Other data types could have been included as well, but this would have extra complication to the problem, which is a good idea to do for future work, but not for a first step. For chapter 5, the main aim was to understand the limitations of satellite derived data in calibrating model parameters. As with chapter 4, we limited ourselves to one data type namely LAI, because we did not want to over-complicate the problem, but assimilating more types of data would be a good extension to the problem.

For (ii), the accuracy of the data is a very important consideration, because whether data is being used on its own to understand the behaviour of the environment or to validate models, if the data have too high an uncertainty then then it may be too unreliable to use. For this reason, quantifying the data uncertainty is therefore important, and it is critical for data assimilation problems. In the next subsection and in the methodology sections of the chapters that follow, we describe how data uncertainty is quantified. Question (iii) is really a special case of (ii). Characterising and quantifying the uncertainty of data inferred from satellite can be difficult (Demantry et al., 2007) but it is important to be able to quantify it well (De Kauwe et al., 2012). We demonstrate this in chapter 5 when assimilating LAI data inferred from satellite measurements.

In the next subsection we describe how NEE and LAI data, used in this thesis, are measured or obtained, critically reviewing the different methods that could be used to obtain this data and giving arguments for using the particular procedure for this thesis. We also briefly describe how the uncertainty on these data are quantified, with more specific detail is given in the methodology sections of the chapters that follow, chapter 4 for the NEE data, and chapter 5 for the LAI data.

2.5.2 Net Ecosystem Exchange (NEE)

TRADITIONAL WAYS OF MEASURING NEE

Traditional apparatus for measuring NEE include leaf cuvettes (e.g. Field et al., 1982), whole plant chambers (Baldocchi et al., 2010) and soil chambers (Cable et al., 2008). A leaf cuvette is a suitcase size portable analyser which connects to a clamp large enough to cover the area of a leaf so that the leaf is sealed inside. The main quantities it is used to measure are the net photosynthetic assimilation rate and the transpiration rate (Rodeghiero et al., 2007). Chambers, whether whole plant types of soil types, also measure rates of net photosynthetic assimilation of C and transpiration, but the volume of space in which measurements are made is much large. In this way, rather than making measurements of a single leave, one could make measurements of an entire plant (if it was large enough to fit inside the chamber, so a small shrub). As the name suggests, soil chambers are used to make measurements over terrain which is either just soil or has very little vegetation (grass) (Bachmann et al., 2010). While these apparatus are still being used today to make measurements of net photosynthesis and soil respiration rates, e.g. PHACE project (Morgan et al., 2011, Pendall et al., 2004, Parton et al., 2007), they have some major drawbacks. Their physical size limits the space over which C exchange rates can be measured, for example it is very hard or impossible to use a leaf cuvette to make measurements over enough leaves in a reasonable time frame (e.g. an hour) to be able to properly characterise the true spatial variability with a high enough statistical accuracy (Baldochi, 2003). Also, it is very labour intensive to be able to track temporal variability of C exchange between a leaf and the atmosphere (i.e. someone needs to be present with the machine throughout the day) and impossible to do this for the entire plant if it is bigger than a small shrub (Cable et al., 2008). This impossible task becomes even more impossible if we wish to track NEE of the entire plant not only over one day but over consecutive

months, years and decades. Being able to make these types of measurements is critical to be able to validate and calibrate models which quantify how the terrestrial sink of C is changing over large spatial scales (regionally, temporally and globally) and temporally (Baldocchi et al., 2008). This was the motivation to develop new techniques for measuring NEE to be able to make sure continuous and plant wide measurements of NEE.

EDDY-COVARIANCE THEORY FOR MEASURING NEE

Developed in the 1970s and 1980s, the eddy-covariance technique was the first method to provide frequent measurements of the land-atmosphere fluxes of carbon (i.e. NEE) and water vapour all year round (Wofsy et al., 1993), and has remained the only method to do so to the present day. Currently there are almost 200 sites where such measurements are made, in a range of different biomes and terrain (Baldocchi, 2008). Even though it is infeasible for these methods to be used on their own to obtain accurate global estimates of NEE, such the measurements are used to calibrate (i.e. Data Assimilation) and validate global estimates from models. We next describe the theory; one of the most recent summaries of the theory and current knowledge on its application including limitations is Baldochhi (2008). We therefore use Baldochhi (2008) as a basis from which to describe the main principles and equations behind the eddy-covariance method, which are outlined in the paragraphs that follow.

The eddy-covariance technique uses a sonic anemometer to measure the vertical wind speed and an infra-red gas analyser (IRGA) to determine the CO_2 concentration. These two pieces of equipment (shown in figure 2.11) are now described.

Sonic anemometer: Wind (which is the main carrier of the CO_2 fluxes) moves in a vertical direction, but also horizontally in the x and y directions (figure 2.11b). The horizontal movements are mainly due to eddies which form as a result of winds blowing faster at higher levels than at the surface (Baldocchi, 2003). The sonic anemometer converts the circular movements of the wind (the eddies) to a Cartesian coordinate system (figure 2.11b). Assuming that the terrain is flat, we find that the net fluxes of CO_2 in the horizontal direction (i.e. along the x-y plane) become zero, i.e. what goes in one side of each of x- and y- directions come out of the other sides (Baldocchi, 2001). The wind speed is determined by sending an acoustic signal from

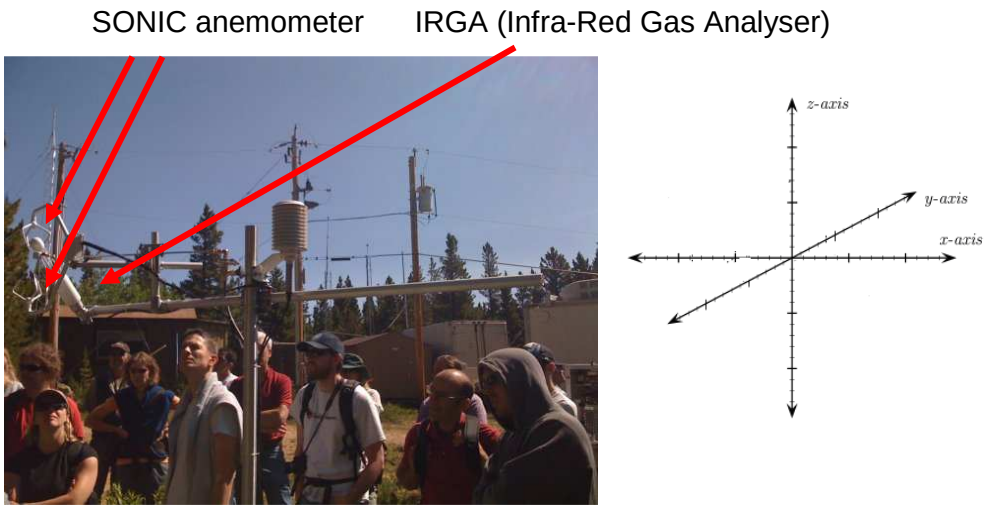


Figure 2.11a (left) A photograph of the IRGA and sonic anemometer (taken by the E. Ryan, author of this thesis). Figure 2.11b (right): The Cartesian coordinate system.

one end of the sonic anemometer to the other end. If the wind is in the same direction as the travelling signal, the signal will move faster than if there was no wind. Since time = distance / speed, we form the equation:

$$t_{out} = \frac{d}{(c + U_{par})} \tag{2.37a}$$

where: t_{out} = time, d = distance, U_{par} = speed of wind, c = speed of sound (U) in the direction of the acoustic signal. This is analogous to a plane travelling faster when in the direction of the wind. The quantities t_{out} and d are easily calculated or already known. When the acoustic signal travels from its destination to its origin this results in a similar equation to equation (2.37a) however this time the wind-speed is deducted from the overall speed because the signal is travelling in the opposite direction:

$$t_{back} = \frac{d}{(c - U_{par})} \tag{2.37b}$$

Solving these simultaneously gives:

$$U_{par} = \frac{1}{2}\left(\frac{d}{t_{out}} - \frac{d}{t_{back}}\right) \quad \text{and} \quad c = \frac{1}{2}\left(\frac{d}{t_{out}} + \frac{d}{t_{back}}\right)$$

Thus we have an expression for the wind speed (U_{par}). This description is consistent with Baldocchi (2007).

IRGA: This measures the CO_2 concentration using a laser located between its three prongs (figure 2.11a). It operates by shining an infrared light through a sample of air

and onto a detector. The energy from the light is absorbed by the CO_2, thus the CO_2 concentration can be inferred from the decrease in the level of energy that reaches the detector (Goulden, 1996).

The covariance part of 'Eddy-covariance': Both the wind speed and CO_2 concentrations are measured at 30Hz, ie 30 measurements / second. For each measurement, the covariance between the vertical wind speed and the CO_2 concentrations is computed and since this is proportional to the vertical flux of CO_2, we are able to work out NEE (Baldocchi et al., 2003). It should be noted that this method does not distinguish between the downward and upward components of the vertical flux (i.e. photosynthesis and C respiration). However in the past decade, it has become possible to approximately partition NEE into its photosynthetic C and respiration constituents by measuring and modelling the C_{12} and C_{13} isotope content of the C flux (Zobitz et al., 2007).

Assumptions: There are four assumptions which are made when measurements of NEE (and other C fluxes) are made using the eddy-covariance method (Baldocchi, 2003):

(1) The terrain is flat.
(2) The underlying vegetation is homogenous and it is situated on flat terrain for an extended distance upwind.
(3) The atmospheric conditions are steady.
(4) The sensors and the data-logging system are able to sense and record the fastest and smallest eddies.

LIMITATIONS OF THE EDDY-COVARIANCE TECHNIQUE

The eddy-covariance technique produces biases when one of the above assumptions is broken, and the resulting bias induced is not accounted for. For example, if the first assumption is not met and the terrain is sloping, then the net horizontal fluxes of CO_2 cannot be assumed to be sum to zero as stated in the text prior to equation (2.37a). Another common problem is the low wind speed at night-time (breaking the third assumption) results in inadequate mixing of the CO_2 leading to an under-estimation of NEE. It some cases, such as the night-time under-estimation, the bias can be

reasonably well quantified in some situations (Aubinet et al., 2005), but for other types of bias (such as non-flat terrain) there is disagreement how to properly account for them (Baldocchi, 2008).

Even if the assumptions are met, there are other problems which could cause errors, unaccounted for, to emerge. For example if the system gets contaminated (by insects or soil getting in for instance) this can cause errors in the eddy-covariance readings. Modern systems contain a piece of hardware called an automatic gain control which gives a measure of how clean the data are, but ultimately it is up to the scientist analysing and processing the raw data to adjust the results to eliminate this bias (Baldocchi et al., 2003). This leads to another problematic issue - there are checks which should be done on the raw data which can be thought of as diagnostics to determine the accuracy of the data. It is unclear if all flux sites do this well.

Ultimately if the data contains biases or is erroneous the data cannot be used, resulting in gaps in the time-series of data, with a typical Fluxnet site having approximately 60% coverage. When the data is made available to modellers, there is also no measure of how accurate the data is. However in the past 6 or 7 years, some tools have emerged to be able to gain some appreciation of data uncertainty (e.g. Hollinger & Richardson, 2005). These are described next.

2.5.3 Quantifying NEE observational errors

Observational error in NEE is defined as the difference between the measured and true values of NEE. This can be split into random and systematic elements (Goulden et al., 1996; Moncrieff et al., 1996). Figure 2.12 gives a graphical explanation of the different between the two. Random errors (figure 2.12a) are due to random natural variations. Systematic errors (figure 2.12b) are a result of a problem with the measurement process (e.g. insects getting inside the sensor) which results in a bias.

RANDOM ERRORS

Hollinger & Richardson (2005) and Richardson et al. (2006) were the first to make significant progress in quantifying the random errors of NEE (and other C flux) measurements. These studies aimed to determine the distribution of random errors by comparing pairs of measurements from two flux towers less than 1km apart from one another (in Hollinger & Richardson, 2005) or from the same flux tower but 24 hours

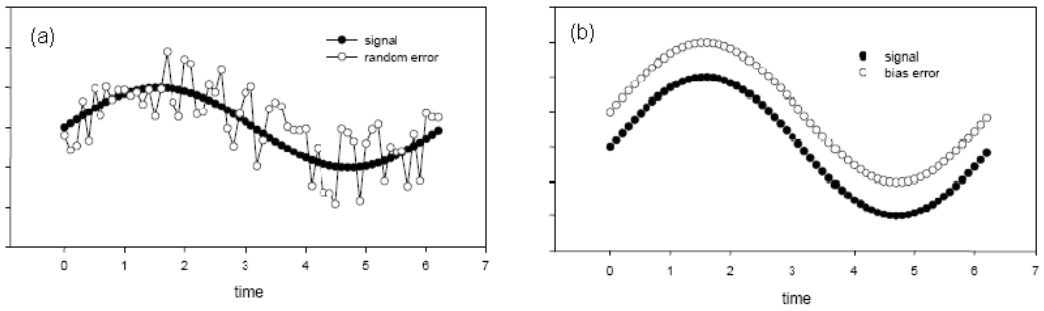


Figure 2.12 A graphical representation of the difference between (a) random and (b) systematic observational errors (Baldocchi, 2012).

apart and under similar meteorological conditions (in Richardson et al., 2006). We outline in the paragraphs that follow the methodology of Richardson et al. (2006) as this is reliant only on NEE from one flux tower, rather than on the strict criteria of two flux towers within 1km of one another.

In mathematical terms we let $x = x^* + \varepsilon$ be a measurement, where the random error ε is an i.i.d. (identically and independently distributed) measurement, distributed with mean zero and variance $\sigma^2(\varepsilon)$ and x^* is the true value, the aim was to estimate $\sigma^2(\varepsilon)$. We now suppose that we have a pair of measurements $x_1 = x^* + \varepsilon_1$ and $x_2 = x^* + \varepsilon_2$, where the random errors, ε_1 and ε_2, are also i.i.d. with mean zero and variance $\sigma^2(\varepsilon)$. The distribution of the random variable $\varepsilon_1 - \varepsilon_2$ has a distribution with expected value and variance:

$$E(\varepsilon_1 - \varepsilon_2) = E(\varepsilon_1) + E(\varepsilon_2)$$

$$\Rightarrow \quad E(\varepsilon_1 - \varepsilon_2) = 0$$

$$\sigma^2(\varepsilon_1 - \varepsilon_2) = \sigma^2(\varepsilon_1) + \sigma^2(\varepsilon_2)$$

$$\Rightarrow \quad \sigma^2(\varepsilon_1 - \varepsilon_2) = 2\sigma^2(\varepsilon)$$

$$\Rightarrow \quad \sigma^2(\varepsilon) = \tfrac{1}{2}\sigma^2(\varepsilon_1 - \varepsilon_2)$$

$$\Rightarrow \quad \sigma^2(\varepsilon) = \sigma^2(\tfrac{1}{\sqrt{2}}(\varepsilon_1 - \varepsilon_2))$$

When tested against multiple sites, the authors found that these random errors $\varepsilon = \frac{1}{\sqrt{2}}(\varepsilon_1 - \varepsilon_2)$ approximately followed a Laplace (or double exponential) distribution which has probability density function:

$$f_\varepsilon(x) = \frac{1}{2b}\exp\left(\frac{-|x-\mu|}{b}\right),$$

where μ and b are referred to as the location and scale parameters. This distribution has a mean and variance given by μ and $2b^2$. It should be noted that this error distribution refers to observations made every half an hour.

SYSTEMATIC ERRORS

Systematic errors result from, for example, the measuring equipment, if incorrectly calibrated. They can also result from one or more of the assumptions of the eddy-covariance method not being met as stated earlier (Baldocchi, 2008). Most of the time, systematic errors can be identified indirectly, e.g. if the measured wind speed is too low (a problem which occurs mainly at night), this can result in an under-estimation of NEE. When systematic errors are identified, a common way of dealing with them is to remove the corresponding NEE measurements since it is very difficult to accurate quantify these biases.

2.5.4 Leaf Area Index (LAI)

LAI (m^2/m^2) is defined as 'the one-sided green leaf area per unit ground area, in broadleaf canopies, and as the projected needle leaf area in coniferous canopies' (Myneni et al., 2002). In this subsection we provide an overview of its importance, how it is measured (from ground and space), and whether these two approaches give the same values. A fuller description and discussion is provided in sections 5.1 and 5.2.

LAI is a key climate variable for estimating the growth of vegetation and for understanding key components of the water cycle and the energy balance (Sellers, 1997). It has a strong influence on: (i) the interception of solar radiation, (ii) photosynthesis, (iii) transpiration, (iv) respiration, and (v) water interception (Breda, 2003). For (i) & (ii), for example, if LAI is high there is more foliage (i.e. leaves, needles) on the trees meaning the total surface leaf area will be larger, resulting in more interception of solar radiation and more photosynthesis. However the more the leaves clump together the less of an impact LAI has on radiation interception and photosynthesis. This is because the clumping of the leaves means that the solar radiation is less likely to be able to penetrate to the entire the surface of each leaf (Breda, 2003).

LAI can be measured from the ground directly or indirectly. Direct methods involve stripping the leaves off a tree or collection of trees and feeding each leaf through a scanner to determine its one-sided surface area (Smith et al., 1991). As a result, such methods give very accurate measurements of LAI. However such methods are very rarely used because they have two major drawbacks: (i) they are destructive, i.e. they involve the removal of the leaves of a tree or trees; (ii) they are very time consuming as a tree can contain many thousands of leaves. As a result indirect methods are normally used. There are many different approaches, but the most common involves estimating LAI by measuring the gap fraction of the canopy using optical techniques (e.g. fish-eye photography) from different zenith angles (Gower et al., 1999). However to do this, we need knowledge of other quantities, which include: (1) the canopy extinction coefficient, defined as the fraction of leaf area that is projected onto a plane perpendicular to θ (Gower et al., 1999); (2) the clumping index, which equals l when the foliage is randomly distributed and less than 1 as it becomes more clumped (Gower et al., 1999); (3) the ratio of Plant Area Index (PAI) to LAI, where PAI is the same LAI except refers to the woody parts of the tree (i.e. the tree trunk and branches) as well as the leafy parts (De Kauwe et al., 2011). A variety of different instruments are available to measure the canopy gap fraction but a common one is the Li-Cor LAI Plant Canopy Analyzer (Gower et al., 1999, Hyer & Goetz, 2004).

Although indirect methods have become the standard ways of measuring LAI because they are not destructive and much quicker than direct methods, measurements can be unreliable if the canopy extinction coefficient, the clumping index or the ratio of PAI to LAI are not estimated well enough (Breda, 2003). However, most validation studies suggest that these indirect methods do give accurate estimates of LAI when compared to direct methods (MacFarlane et al, 2007). Measuring LAI from the ground is useful for site level studies, but if we require global and frequent measurements of LAI, using these indirect methods would become an impossible task since in-situ measurements are only made at a limited number of sites worldwide and typically very few are made annually (e.g. 10–20). The only way therefore to obtain frequent global measurements is to measure LAI from space.

LAI is currently measured from space by a number of satellites which include MODIS (Yang et al., 2006), AVHRR (Holben, 1986), MERIS (Bacour et al., 2006),

and CYCLOPES (Baret et al., 2007). On MODIS for example, global measurements are made every 8 days at a resolution of 1km×1km. Unlike ground based indirect methods of measuring LAI, satellites measure the light reflectance rather than the gap fraction, therefore the algorithm used to convert this raw data to LAI is different to those used to process gap fraction data. Therefore although the algorithms used to estimate LAI in both situations might be similar, the values of retrieved LAI would be different, but perhaps related. Various validation studies have been carried out on satellite derived LAI observations, and suggest the MODIS LAI data offers one of the most realistic representation of the spatial and temporal variability of LAI, e.g. Garrigues et al. (2008), Fang et al. (2012) and De Kauwe et al. (2011).

2.6 Aims and layout of the thesis

Based on the gaps in our current knowledge (subsection 2.3.2), the aim of this this thesis is to gain better insight into how Data Assimilation (DA) should be used. This will help indirectly with improved understanding and quantification of the land-atmosphere exchange of C between by outlining certain areas for how DA should be optimally used. There are two aims to this PhD:

- When using DA to estimate parameters using ground observations:
    a. Understanding the limitations of DA and the conditions it performs best.
    b. Determining likely factors that cause parameter estimates to vary.
- Assessing the impact of assimilating satellite observations of leaf area index to improve the model states, and whether DA is robust against unrealistic features of the satellite data.

In subsections 2.3.2 and 2.4.2, arguments were given for the choice of DA algorithms and model to use based on the aims of this thesis. The algorithms are the Metropolis algorithm and the Ensemble Kalman filter, and the model is the evergreen and deciduous versions of the DALEC model. A brief description of the remaining four chapters that follow is now given.

LAYOUT OF THESIS

Chapter three determined some of the factors that cause parameter estimates from DA

to vary. We assessed whether the differences in the parameter estimates amongst the nine different DA schemes from the REFLEX study (Fox et al., 2009) were due to the different initial conditions of the model and LAI weighting term used in the cost function, which were used for the different algorithms. To test this, a sensitivity analysis was carried out on one of the DA schemes to see how the estimates to the parameters, using the same model and same LAI and NEE data, changed for different initial conditions and LAI weightings. The main finding was that the parameter estimates were most sensitive to changes in the initial conditions of the labile carbon pool in the deciduous version of DALEC.

The results of chapter four were published in a paper in *Global Change Biology* (Hill, Ryan and Williams, 2012). It aimed to determine if the parameter estimates and subsequent forecasts, obtained using the EnKF and the Metropolis algorithm, were sensitive to changes in the length, frequency of gaps and observational error of the dataset being assimilated. It was reported that the dataset length was the most important of the three factors, while varying the data density and observational error had only minimal impact on the estimates and forecasts. The EnKF was very good at estimating parameters and provided reasonably good forecasts. Surprisingly however, while the Metropolis algorithm was in general poor at estimating parameters, it produced very accurate forecasts. The reasons for this are discussed.

In chapter five, MODIS satellite observations of LAI (Leaf Area Index) were assimilated into the evergreen version of DALEC, using the EnKF, in order to estimate NEE, LAI and the carbon pools of DALEC. This chapter reported that assimilating this MODIS LAI, which contains unrealistic features, into DALEC resulted in the modelled NEE and LAI being a worse fit to the observations than if no data had been assimilated. However, if the MODIS LAI is processed, to remove the unrealistic features prior to assimilation, then the fit of the modelled to observed NEE is vastly improved. The improvement was more when the modelled GPP was compared to GPP measurements.

Finally in chapter six, we summarise the findings of this thesis, assessing whether the aims have been met and how this thesis contributes to a gain in knowledge in the wider context. This chapter also outlines the limitations, unanswered questions, and lists several recommendations.

Chapter 3

Assessing the sensitivity of the parameter estimates from a DA scheme to the initial conditions and the cost function

Abstract

This chapter investigates the sensitivity of estimated parameters to the initial values of the labile, litter and fine roots carbon pools of the model and the weighting term of the leaf area index term used in the cost function. This work is motivated by determining what factors might be causing variations in parameter estimates, as stated in the aims section (section 2.6) of this thesis. The choice of factor here, i.e. the initial conditions of the model, was based on the results of the REFLEX study (Fox et al., 2009) where nine DA schemes (six of which were variants of the Metropolis algorithm) gave varying estimates of the model parameters even though the same model (DALEC-D) and the same synthetic observations were used. Here we determine the sensitivity of the parameter estimates using one of the variants of the Metropolis algorithm. The sensitivity of each parameter estimate to each initial conditions of the C pools was determined by averaging over the other initial conditions rather than fixing them at nominal values (such as the central value). Carrying out sensitivity analysis of this kind would typically require thousands of runs of the Metropolis algorithm, but a computationally cheaper approach is to build an emulator, a statistical representation of the input-output relationship of the Metropolis algorithm. To build an emulator

with four inputs requires only 60 runs of the Metropolis algorithm (40 to build it and 20 for validation) and carrying out the sensitivity analysis using the emulator typically requires less than 30 seconds of CPU time. Due to complexities in building an emulator with a multi-dimensional output, it is much easier to build ones with scalar outputs. This meant 17 emulators needed to be built for the 17 parameters.

Before carrying out the sensitivity analysis, it was necessary to ensure that the emulators were accurately representing the Metropolis algorithm. This was done by comparing the emulator's estimates of the parameters with those obtained from the Metropolis algorithm using the inputs from the validation runs. These diagnostics showed that five of the emulators were not predicting the Metropolis algorithm accurately enough. In building the emulators certain assumptions were made, but for two further emulators these were found to be violated, therefore they were also excluded from the sensitivity analysis. The fact that seven of the seventeen DALEC-D parameters were excluded from the results did not matter. This was because of the twelve parameters which varied among the DA schemes in REFLEX, only three were excluded.

For the ten parameters whose emulators passed the diagnostic tests, it was found that eight were sensitive to the initial values of the labile carbon pool. The parameter estimates were also sensitive to the initial values of the litter and fine roots carbon pools and the LAI weighting term but for a fewer number of parameters (five, two and six respectively) and the sensitivity was weaker compared to the initial value of the labile carbon pool. An interesting finding for the labile carbon sensitivity plots was that nine of the ten parameters were very close to the truth when the true initial value of the labile carbon pool was used. However obtaining accurate enough estimates of the initial value of the labile carbon pool is very time consuming and labour intensive. It involves sampling tree tissue and determining the soluble starch and sugars in the laboratory. The process is complex and time consuming. As a result, measurements of labile carbon are rarely available. Hence we tested whether incorporating other information into the assimilation resulted in the parameter estimates being close to the truth even if the initial value of the labile carbon was poorly estimated. This involved carrying out two extra sets of runs of the Metropolis algorithm. For the first, we assumed that the initial values of the litter and fine roots pools were known. For the second, we also increased the number of LAI observations

being assimilated from 10 to 100. For both sets therefore only the initial value of the labile carbon pool was varied. Unfortunately, this extra information made little difference to the sensitivity of the parameters to the initial value of the labile carbon pool.

As a result of the above, we assessed the impact of wrongly estimating this initial condition on the trajectories of the C pools and key model states, using the parameter estimates obtained from the wrongly specified $C_{lab}(0)$ values. We found that over-estimating $C_{lab}(0)$ by a small amount (i.e. $C_{lab}(0) = 110$ gCm^{-2}) compared to the true value (i.e. $C_{lab}(0) = 70$ gCm^{-2}) resulted in the trajectories being closer to the true trajectories, than if $C_{lab}(0)$ had been under-estimated by a similar amount (i.e. $C_{lab}(0) = 35$ gCm^{-2}) or over-estimated by a considerable amount (i.e. $C_{lab}(0) = 170$ gCm^{-2}). Thus, even if the parameter estimates require the $C_{lab}(0)$ value to be accurately estimated, there is scope for $C_{lab}(0)$ being over-estimated by a reasonable amount whilst still being able to estimate the C pools and important model states such as NEE to a reasonable level of accuracy.

An alternative option is to treat the initial conditions as parameters. It was found that even though two of the three initial conditions estimated using DA had posterior estimates far from the truth, the posterior distributions of other parameters and the corresponding trajectories of the C pools were in general consistent with the truth.

3.1 Introduction

3.1.1 Background

The central role of a model is to try to represent and understand reality. For a land-surface vegetation model, like many models, this is often based on mathematical equations which describe our best knowledge of the plant processes and physics of the system. These processes are controlled to a large extent by parameters which are physical quantities giving information about what the system is doing. The Oxford English Dictionary defines a parameter as ‘a quantity whose value is selected for the particular circumstances and in relation to which other variable quantities may be expressed.’ In process models they can be fractions, such as the proportion of GPP

allocated to foliar carbon; they can also represent limits of processes, for example the maximum carboxylation rate at 25˚C (Bethy model, Knorr and Kattge, 2005); or they could be a range of other specific boundaries on how the model should operate, e.g. the canopy light attenuation constant (PnET model, Aber and Federer, 1992). In all models these parameters need to be estimated by some means. In estimating them it is almost always assumed that they are fixed for all time, including the DALEC model which is used in this thesis. However, for large models which estimate the system regionally or globally, the values of the parameters do tend for different places, for example the fraction of GPP allocated to foliage might depend upon the vegetation type. While the assumption of parameters being fixed in time may hold for many parameters, for others they may be quantities which in fact vary in the real world at different times, for example, the turnover rate of foliar carbon might depend upon the season. By fixing these 'varying parameters' we might be imposing an unrealistic assumption on the model. It is therefore important to understand the impact on the model's output (e.g. NEE) by treating the parameters as varying as opposed to being fixed in time. Due to time constraints, this could not be addressed in this PhD.

Observations of the state of the system can also be made, and these can be compared with the model's estimate of the state. Data Assimilation (DA) is a way of combining the model with these observations so that the estimate of the state of the system is better than just using the model or the observations on their own. DA can also be used to estimate parameters by comparing the model's estimate of the state of the system, using a particular configuration of parameters, with observations. This is useful as it provides a way of estimating certain parameters (and processes) that may be unobservable or very difficult to accurately observe in the real world. A corollary is that we do not need the same number or types of observations as the number of parameters and/or processes we are trying to estimate; for example, one could use only Net Ecosystem Exchange (NEE) observations to estimate all the parameters of a model.

3.1.2 Justification for this chapter

One of the aims of this thesis is to determine what causes variations in parameter estimates among DA schemes, as this has been observed in two inter-comparison DA studies, namely OpTIC (Trudinger et al., 2007) and REFLEX (Fox et al., 2009). In

this chapter, it is hypothesized that these differences are not due to theoretical differences in the DA schemes (i.e. they are solving the same problem), but as a result of assumptions that are made in the set-up of the algorithms and the type of data and model used. The REFLEX study (Fox et al., 2009) in particular highlighted several factors that could be causing variations in the parameter estimates and associated uncertainties using a DA. These factors are listed in **(1)** to **(6)** below, but before these are stated and explained, a summary of REFLEX is given.

Nine different DA schemes were employed to estimate the parameter estimates of the evergreen and deciduous versions of the DALEC model using the same set of NEE and Leaf Area Index (LAI) observations. The participants were asked to estimate the parameters and future model states using real and synthetic data, where the synthetic data were created by a running the model forward and adding noise and gaps to the NEE and LAI. For all DA schemes, the uncertainty reduced on parameters linked to GPP and respiration e.g. the allocation and turnover of foliage and temperature sensitivity), as a result of assimilating the NEE and LAI data. Poorly estimated parameters were those related to the allocation to and turnover of fine root/wood pools, which suggested that assimilating observations of fine roots and woody biomass may help to constrain the parameters connected to these the fine root and woody carbon pools. A limitation of REFLEX is that other DA schemes, for example the Particle filter and 4D-Var, were not included. Nonetheless, the results of the study were informative. For 12 out of the 17 parameters the DA schemes gave different estimates; for 7 of these 12 parameters the differences were marked, while for the other 5 the DA schemes gave similar estimates except for only one or two of the schemes, one of which was always the Multiple complex Monte Carlo Markov Chain (MCMC) method. We might expect there to be differences between DA methods that were sequential in nature, e.g. the Ensemble Kalman filter, as opposed to batched methods such as the Metropolis algorithm which use all the data in one go. However six of schemes were variants of the Metropolis algorithm, and so while we might expect there to be slight differences because the methods are mildly different in how they process the data and how they search the parameter space, there are other factors which may also be causing differences in the parameter estimates. These factors, which REFLEX did not address, could be:

(1) Different the initial conditions of the model: In REFLEX, the initial conditions of the small C pools were estimated differently for each of the DA schemes. This may have contributed to the variation in parameter estimates amongst the variants of the Metropolis algorithm. This is particularly true in this instance because the DA was carried out over a 2 year period because the trajectories of the C pools may not have had time to stabilise to a pseudo steady state (Carvalhais et al., 2008). Even if we were assimilating a longer dataset, the assumption of the C pools converging to steady state may not be valid (Braswell et al., 2005), and so the choice of initial conditions of the small C pools may affect the posterior distribution of the parameters from the inversion scheme.

It should be noted here that in some ecological applications of the Metropolis algorithm, the initial conditions of the model are sometimes treated as parameters (e.g. Braswell et al., 2005). The argument to do this is that they are unknowns and so by treating them as parameters, we allow them to be estimated based on the data and some prior distribution. A problem with this however is that the can be no or very little information in the data to give good estimates of the initial conditions from the posterior distributions. Moreover, known issues with the Metropolis algorithm such as parameter equifinality could then result in poor estimates of the initial conditions and the associated uncertainties. As a result of this, it is also common to fix the initial conditions at likely values based on site knowledge (e.g. inventory data from the site). Examples of ecological applications of the Metropolis algorithm where the initial conditions were fixed include Zobitz et al. (2011), Carvalhais et al. (2008), Knorr & Kattge (2005), Richardson et al. (2010) and Williams et al. (2010). In this way, understanding how different choices of the initial conditions of the model affect the parameter estimates of the model is relevant for applications of the Metropolis algorithm where these initial conditions are treated as fixed.

(2) The cost function. This computes the sum of squared difference between the observations and the model output, divided by weighting term. Where there are multiple streams of data, as in REFLEX where there were NEE and LAI data, the weighted sum of squared of the two data types are added together. In

mathematical terms, the cost function involving two data streams is written below where $x_1(t_i)$ and $x_2(t_i)$ are the observations at time t_i, $z_1(t_i)$ and $z_2(t_i)$ are the corresponding modelled values, and w_1 and w_2 are the weights. A common choice for the

$$\Theta = \sum_i \left(\frac{(x_1(t_i) - z_1(t_i))^2}{w_1} + \frac{(x_2(t_i) - z_2(t_i))^2}{w_2} \right) \quad (3.1)$$

weights is the observational errors. For REFLEX, the observational error was given for the NEE data, but not for the LAI data and so this needed to be estimated. The OpTIC experiment (Trudinger et al. (2007)), which also compared different DA schemes when estimating model parameters and using two streams of data, found that the choice of the weights in the cost function did not affect the parameter estimates. However this finding may not be true for REFLEX for two reasons: (i) the DALEC-D model contains non-linearities which the model used in OpTIC did not; (ii) the OpTIC model only had two state variables and 4 parameters to be optimized, whereas DALEC-D has 7 state variables and 17 parameters to be optimized. Given that the values used for the LAI weighting term amongst the variants of the Metropolis algorithm also varied from 0.0005 m^2/m^2 to 0.11 m^2/m^2, it is of value testing whether varying this LAI weighting term would make a difference to the parameter estimates using the DA algorithms from REFLEX. Therefore this is also investigated in this chapter.

(3) The type of prior distribution used: In REFLEX seven of the nine algorithms used a uniform prior distribution with the remaining two adopting a Normal one. We do not investigate the effect of the type of prior distribution on DA predominantly because there was still variation in the estimates for a number of the parameters amongst the DA schemes which adopted a Uniform prior. Also, it is perhaps more clear about what we might expect, in terms of the differences with posterior distributions, whether we used a Uniform or Gaussian prior. In other words, for those parameters which may not be constrained by the observations the default distribution (ignoring correlations between the parameters in the multivariate posterior parameter space) is then the prior. In reality, where there is no or little information in the data to constrain these certain parameters, then the posterior distribution of these

parameters is influenced by the other parameters due to the correlations that exist between them. However if our prior is Gaussian, this will exhibit some 'pull' on the shape of the posterior distribution for these unconstrained or weakly constrained parameters (Knorr & Kattge, 2005). Conversely, if a Uniform prior is assumed then since this has no modal value or values it will have far less influence on these parameters and even no influence if this Uniform prior is uninformative (Van Dongon, 2006). The point of this is that there is some idea how the parameter estimates and posterior distribution might differ if a Uniform versus Gaussian prior distribution was assumed. In contrast, there is less idea of the effects on the parameter estimates and uncertainties to other factors such as the varying initial conditions; therefore it is more critical to investigate this first. Due to time constraints in this PhD, understanding the effect of the type of prior on the shape of the posterior distribution is carried out here, but it is nonetheless in need of investigating in future work.

(4) Parameter equifinality: This is defined as getting very similar modelled trajectories of NEE (say) using different parameter sets (Tang & Zhuang, 2008). Since the Metropolis algorithm operates by finding a set of parameters which gives an optimal fit of the model's output to the observations, parameter equifinality can result in two different sets of parameters giving similarly optimal fits to the observations. This will occur if the model's output is insensitive to certain parameters, which are sometimes called 'redundant parameters'. Chapter four finds that the NEE estimates from DALEC are very insensitive to three parameters (the turnover rates of fine root carbon, litter carbon and labile carbon), and very strongly sensitive to two parameters (the fraction of GPP respired autotrophically and the turnover rate of foliar carbon).

(5) Treating varying parameters as fixed: Instead of all parameter being fixed for all time, some may have different values for different times in the year, and this may also influence equifinality; in other words if it is more meaningful for the value of a parameter to vary with time but we are forcing it to be fixed, then the DA algorithm may not know what value to choose for such parameters and so different sets may result in equally optimal fits of the model to the observations.

(6) Different quantities and qualities of data being assimilated. Although the same dataset was assimilated for each DA scheme, the parameter estimates between the schemes may have varied if the assimilated time-series was longer, had a different number and frequency of gaps, or was more or less noisy. This is considered in chapter four where estimates of the DALEC-D parameters, using the EnKF and the Metropolis algorithm, from assimilating different NEE datasets were compared, and so is not addressed here.

In this chapter, we investigate whether the different initial conditions of DALEC-D and the LAI weighting terms used in the cost function in REFLEX (labelled as (1) and (2) above) caused the parameter estimates to vary amongst the variants of the Metropolis algorithm, as this is unknown. Due to time constraints on this PhD, (3) is not investigated here as discussed on the previous page, nor is (5) for the same reason. (4) and (6) are also not considered here as they are explored in chapter four.

In light of the discussion in the second paragraph of (1) on page 134, as well as investigating the sensitivity of the initial conditions of the small C pools on the parameter estimates from the posterior distribution, the optimisation is also carried out by treating the initial conditions - $C_{lab}(0)$, $C_r(0)$, $C_{lit}(0)$ - as parameters to be estimated, with the LAI observational error set at the most common value used by the REFLEX participants. This will enable us to answer two questions:

(i) Is it better to treat the initial conditions as fixed or as parameters?

(ii) If we treat the initial conditions as parameters, what impact does wrongly estimating them have on the other model parameters and the C pools?

3.1.3 Aims and outline of chapter

The objectives of this chapter are to determine the sensitivity of the DALEC-D model parameters to: (1) the initial conditions of the small C pools and, (2) the LAI weighting term of the cost function, when using DA to estimate the 17 parameters of the model. The major novelty in this chapter is the use or a number of emulators to carry out this sensitivity analysis. The reasons for using an emulator and how it was implemented for the purposes of carrying out a sensitivity analysis described in full in the next section (3.2). In the context of this chapter, an emulator is a statistical model which models the input-output relationship of the Metropolis algorithm, where the inputs are the initial values of three C pools and the LAI weighting term. A key

advantage of using an emulator to do sensitivity analysis is that it requires significantly fewer runs of the Metropolis algorithm than conventional approaches. Details are given in section 3.2 as to how the emulator was built, how the diagnostic checks were carried out to ensure the emulator accurately represented the Metropolis algorithm, and how the sensitivity analysis was performed. In the methods section, we also describe the setup of two follow-on experiments, carried out in response to the findings of the main sensitivity analysis results. At the end of the methods, we finally briefly describe another inversion carried out using the Metropolis algorithm but treating the initial conditions of the small C pools as parameters instead of fixed, in light of the last paragraph of the previous subsection.

In Section 3.3 the results of the sensitivity analysis are shown. Prior to this but also in the same section, plots showing convergence of the Metropolis algorithm are shown, as are tables and plots corresponding to the diagnostic checks of the emulators. Also in section 3.3, the results to the two extra sets of follow-on experiments are also shown, as well as the inversion results from treating the initial conditions as parameters. Discussions of the results are given in section 3.4, followed by the conclusions in section 3.5.

3.2 Methodology

3.2.1 Introduction

This section outlines the sensitivity analysis used to determine how the initial conditions and the weighting of LAI in the cost function influence the parameter estimates. Sensitivity analysis is the study of how the outputs of a function are sensitive to changes in its inputs; for this chapter, the function is the Metropolis algorithm, the inputs are the initial conditions and LAI weighting, and the outputs are the parameter estimates.

The sensitivity analysis could be applied to any one of the six variants of the Metropolis algorithm used in the REFLEX study. This is because all of them operate under the same theoretical principles, i.e. they all search the parameter space, avoid local solutions using the Metropolis rule (Metropolis et al., 1953), and all sample the posterior distribution enabling the statistics of this distribution to be characterised.

They differ in certain minor features of how they are implemented, for example how big a step to make between iterations when searching the parameter space. Despite these differences, all will lead to the same posterior distribution, although the sampling may be slightly different due to the stochastic nature of the Metropolis algorithm so the statistics of the distribution may not be exactly the same (this would apply for repeated application of the same variant of the algorithm). Therefore, the choice of what variant of the Metropolis algorithm is arbitrary and therefore just came down to personal preference. It was decided that the variant by Hill & Williams (2009) would be most appropriate. This was mainly because it is well documented and much easier to code up than other versions. Hereafter all references to 'the Metropolis algorithm' in this chapter refer to this variant.

Sensitivity analysis commonly involves using Monte Carlo methods, but these typically require thousands of executions of the algorithm (Saltelli et al., 2000). Since the Metropolis algorithm takes 1 hour to run using the University of Sheffield Iceberg Computer cluster and with 2 years of assimilated NEE and LAI data, Monte Carlo methods are completely infeasible as they would require thousands of hours of computation time. An alternative approach, which was adopted for this chapter and which requires only 40 runs of the algorithm, involves using an emulator. An emulator is a statistical representation of the input-output relationship of an algorithm (O'Hagan, 2004). For this chapter, the inputs refer to the initial conditions of the three small carbon pools and the LAI weighting in the cost function, and outputs refer to each of the 17 parameter estimates. After carrying out diagnostic checks to ensure the emulator properly represents the input-output relationship of the Metropolis algorithm, sensitivity analysis was carried out in a probabilistic framework by computing the variance of the expected value of the output given each input. The building of the emulator and the sensitivity analyses were carried out using a freely available program called the Gaussian Emulation Machine for Sensitivity Analysis (GEM-SA). Before giving the theoretical basis behind the emulator and the sensitivity analysis, details of the implantation of the model and the Metropolis algorithm specific to this chapter are first described.

3.2.2 The DALEC-D model

The DALEC-D model is a simple box model that tracks the journey of C in Deciduous

trees from photosynthesis through to respiration via various C pools and connecting fluxes, working on a daily time-step. Subsection 2.4.4 describes the model in detail. In REFLEX the 17 DALEC-D parameters were estimated for four sites, two of which were deciduous (namely Hesse in France and Brasschaat in Belgium) and two of which were evergreen (namely Loobos and Vielsalm in the Netherlands). A description of the 17 parameters is given in table 3.1. For the Evergreen sites, the model used was the same as DALEC-D (called DALEC) but did not contain the labile carbon (C_{lab}) pool or the connecting fluxes, because evergreen forests retain their foliage all year round and so there is no need to have a store of labile carbon for the spring bloom. There was no preference as to whether the sensitivity analysis should be carried out on an evergreen or deciduous site, since the variation in parameter estimates from the different variants of the Metropolis algorithm occurred at both types of sites. A deciduous site was chosen mainly because there were more parameters to estimate in DALEC-D (17) than DALEC (11) which might reveal more interesting findings in the results.

Leaf Area Index (LAI) was determined by: $LAI = C_f/LMA$ where LMA, Leaf Mass per Area, is a site specific constant. For the Brasschaat site used in this chapter (subsection 3.2.3), LMA = 22 gCm^{-2}. The reasons for the linear relationship between LMA and LAI and the reasons why LMA and foliar nitrogen are not treated as parameters is discussed in subsection 2.4.3, and so are not repeated here.

3.2.3 The site and observations

CHOICE OF SITE

For the NEE and LAI datasets from the deciduous sites used in REFLEX, real data were taken from Hesse whilst synthetic data were taken from Brasschaat. For this chapter, it was decided for the sake of simplicity that the NEE and LAI datasets from one of the sites should be used. The synthetic dataset for Brasschaat was used because whilst giving a good approximation to the real NEE and LAI data, the true values of the parameters were known. The site is situated in the De Inslag forest in Brasschaat (51˚18'N, 4˚31'E), 20km NNE of Antwerp in Belgium. The topography is almost flat and is at an elevation of 16m. The forest is classified as a Deciduous broadleaf forest with dominant tree species Quercus robur. In 1999, there were 375 trees per hectare with a mean tree diameter at breast height of 30cm. The forest has a temperate climate

Symbol	Description	Units	True value	Range (low/high)
T_d (p_1)	Litter decomposition rate parameter	day^{-1}	3.61×10^{-6}	$1x10^{-6}$/0.01
F_g (p_2)	Fraction of GPP respired autotrophically	-	0.32	0.2/0.7
F_{nf} (p_3)	Fraction of GPP allocated to foliage	-	0.47	0.01/0.5
F_{nrr} (p_4)	Fraction of GPP allocated to roots	-	0.3	0.01/0.5
T_r (p_5)	Turn over rate of foliage	day^{-1}	5.67×10^{-2}	$1x10^{-4}$/0.1
T_w (p_6)	Turn over rate of wood	day^{-1}	1.70×10^{-4}	$1x10^{-6}$/0.01
T_r (p_7)	Turn over rate of roots	day^{-1}	6.73×10^{-3}	$1x10^{-4}$/0.01
T_l (p_8)	Turn over rate of litter	day^{-1}	2.44×10^{-2}	$1x10^{-5}$/0.1
T_s (p_9)	Turnover rate of C_{SOM} pool.	day^{-1}	1.30×10^{-4}	$1x10^{-6}$/0.01
E_t (p_{10})	Exponential temperature dependent rate parameter	-	0.0693	0.05/0.2
P_r (p_{11})	Nitrogen use efficiency parameter	-	12	5/20
L_{out} (p_{12})	Growing degree days for leaf out	°C day	250	200/400
L_{fall} (p_{13})	Minimum temperature for leaf fall	°C	9.5	8/15
F_{ll} (p_{14})	Fraction of carbon loss in transfer to litter	-	0.453	0.2/0.7
T_{lab} (p_{15})	Turn over rate of labile carbon	day^{-1}	5.33×10^{-2}	$1x10^{-4}$/0.1
F_{lr} (p_{16})	Fraction of labile carbon respired	-	0.129	0.01/0.5
C_{fmax} (p_{17})	Maximum C_f value	gCm^{-2}	120	100/500
C_f	Foliar carbon pool	gCm^{-2}	0	N/A
C_w	Wood carbon pool	gCm^{-2}	8900	N/A
C_r	Fine root carbon pool	gCm^{-2}	100	20/200
C_{lab}	Labile carbon pool	gCm^{-2}	70	20/200
C_{lit}	Fresh foliar and fine root litter carbon pool	gCm^{-2}	75	20/200
C_{SOM}	Soil organic matter & coarse woody debris pool	gCm^{-2}	9900	N/A

Table 3.1 The description of the parameters (first 17 rows) and C pools (remaining 6 rows) of the DALEC-D model. The range column shows the lower and upper bounds.

with a mean air temperature of 10˚C and an annual precipitation of 750mm.

HOW THE SYNTHETIC DATA WERE CREATED

The synthetic data used in REFLEX were created by running the DALEC-D model forwards in time, with appropriately chosen initial conditions and parameters. To each of these time series, observational error and gaps were added, based on the characteristics of the eddy-covariance NEE and in situ LAI observations from the Brasschaat site. For REFLEX, the data were taken from 2000 and 2001, which consisted of 231 NEE observations and 10 LAI observations. The values of the parameters and the initial values of the C pools used for the forward run (referred to as the 'true' parameter set and the 'true' initial conditions) were chosen under two criteria: (i) based on expert knowledge of an experienced ecologist they appeared realistic for the site; (ii) the resulting modelled NEE and LAI were good approximations to the real NEE and LAI observations. The initial values of the large

pools (C_w and C_{som}) were fixed to the values used in REFLEX, while the initial value of C_f was set to zero, since the first day of the study was in the middle of winter (1st January) when there are no leaves on the trees.

3.2.4 The Metropolis Algorithm

The Metropolis algorithm is a full Markov Chain Monte Carlo (MCMC) algorithm. A complete description of it is given in subsection 2.2.9 of chapter 2. In this subsection, we describe the particular setup of the algorithm that was used for this chapter, and the rationale behind this.

A key aspect of the formulation of the Metropolis algorithm is the prior distribution. Based on knowledge to the parameters (e.g. from expert belief or previous studies), prior values are given to each of the parameters. The uncertainty about these priors is represented by a probability distribution, which is predominantly a Normal or a Uniform one; this is because the Normal distribution is the most commonly used probability distribution in Statistics for representing the distribution of natural phenomenon, while the Uniform is appropriate if there is little prior knowledge since every value between a specified lower and upper bound has an equal chance of being the correct value. For most of the Metropolis-based approaches used in REFLEX a Uniform distribution was adopted and is also used here, mainly because the differences in the REFLEX results occurred even for these variants of the Metropolis algorithm that adopted a Uniform prior. Lower and upper bounds for the parameters were given in REFLEX (right column of table 3.1). For example, for the first parameter, $T_{d,}$ has lower and upper bounds given by $1x10^{-6}$ and 0.01, so the prior distribution of T_d is $U(1x10^{-6}, 0.01)$. This means that for each parameter each value in the given range has an equal probability of being the true value. The initial value for the *i*th parameter p_i ($\underline{P}_i^o$), where $1 \leq i \leq 17$, is determined by generating a random number from $U(l_i,u_i)$ where l_i and u_i are the lower and upper bounds for p_i.

In Knorr and Kattge (2005) and Hill & Williams (2009) the prior values were converted to log-normalised parameters (P_i) and scaled so that each of their prior values was 1. This ensured that the parameters remain positive; however, this is only necessary for a prior distribution such as the Normal that does not have lower and upper bounds. When the prior follows a Uniform distribution, this transformation is

unnecessary and makes no difference to the parameter estimates. For the variant of the Metropolis algorithm used in this chapter this transformation was adopted as it makes it more fluent from a programming point of view when changing between a Uniform and Normal prior distribution. We adopt it here also for consistency.

The algorithm decides whether to move to the proposed parameter set or stay at the current parameter set. This is determined by first computing the cost function, a weighted sum of squared difference between the observations and the corresponding model predictions at the current (written $J^{(k)}$) and the proposed sets (written $J^{(k+1)}$), and is given as:

$$J^{(k)} = \frac{1}{2}\sum_{i=1}^{731}\left[\frac{\left(m^{k}_{NEE,i} - obs_{NEE,i}\right)^2}{w_{NEE}} + \frac{\left(m^{k}_{LAI,i} - obs_{LAI,i}\right)^2}{w_{LAI}}\right]$$

where $m^{k}_{NEE,i}$ and $m^{k}_{LAI,i}$ are the modelled NEE at the *k*th iterative step for the *i*th day, when an NEE observation and/or an LAI observation are present. Also, w_{NEE} and w_{LAI} are the weights; a common choice for these are the variances of the Normally distributed observational errors (denoted s^2_{NEE} and s^2_{LAI}), since the smaller s^2_{NEE} or s^2_{LAI} is, the more of an influence on the cost function. In $w_{NEE} = s^2_{NEE}$ was set to 0.5gCm^{-2} for consistency with what was used in REFLEX. For the LAI observations, the REFLEX participants were told that s^2_{LAI} should be calculated based on a coefficient of variance of between 0.1 and 0.2. In other words,

$$s^2_{LAI} = (\text{mean of LAI observations} \times C)^2$$

where C is coefficient of variance, and the mean of the LAI observations is 2.236 m^2/m^2. For four of the six variants of Metropolis algorithms used in REFLEX, a coefficient of variance of either 0.1 or 0.15 was used, and so using the above formula this results in s^2_{LAI} being equal to 0.05 and 0.11 respectively. For the other two, one used a value of 0.04 whilst the other used a value of 0.0005 (based on coefficient of variance of 1%). The rationale for the participant who used 0.0005 was because there were a lot fewer LAI observations than NEE ones so more weighting should be given to these LAI measurements. Since there is uncertainty about whether the different specifications of the s^2_{LAI} influenced the parameter estimates, s^2_{LAI} was included as an

input for the sensitivity analysis. It is worth noting here that it is arguably more truthful to treat the LAI observational error as time dependent, e.g. a fraction of the LAI observation itself (Williams et al., 2005). This is because the error on LAI in general scales with the size of the LAI measurement (Williams et al., 2005). In this chapter the LAI observational error is treated as being fixed, mainly because this is what was done in all of the variants of the Metropolis algorithm in REFLEX and so this may have influenced the differences in the parameter estimates and uncertainties. Furthermore, there were only 10 LAI observations assimilated in REFLEX, compared to 231 NEE ones; therefore the effect on the results of assuming a constant LAI error, as opposed to a time-varying one, is likely to be minimal.

An assumption of the likelihood function, where the variance term is constant, is that the residuals between each data point and corresponding modelled value are not autocorrelated. Given that there were only 10 LAI data points over a possible 731 time points and that the LAI data were in general way spaced between each other, it is fair to assume that the LAI residuals are independent between each other. For the NEE residuals, the simplest way to check for their independence is to construct a lag-plot (Upton & Cook, 2004). This is what the correlation between pairs of residuals for each lag is calculated and plotted. At lag equal to 0, there is an autocorrelation of 1 (because the residuals at a particular time are perfectly correlated with themselves), but for lags greater than 0, the correlation coefficient should be close to zero if there is no autocorrelation.

As described in chapter 2, the convergence of the Metropolis algorithm is determined formally using the Gelman criteria. A plot of the trajectories of the Gelman criteria statistic, Rhat, is given for each parameter at the start of the results subsection.

3.2.5 Terminology

For the remainder of the methods, we describe how an emulator is used to represent the input-output relationship of the Metropolis algorithm. Emulators are usually used to model this input-output link of a simulator. Strictly speaking the Metropolis algorithm is not a simulator, however it very closely approximates one. This is because if we were to repeat a run of the Metropolis algorithm, the same posterior distribution is being sampled. Therefore, for purpose of describing how the emulator

is built, we write 'simulator' in place of 'Metropolis algorithm' for the remainder of the methods. We also adopt the following terminology:

- $f(\cdot)$ represents the simulator.
- **x** is a $p\times1$ vector representing the simulator inputs which, for this chapter, are the initial values of the foliar, fine root and labile Carbon pools of DALEC, and the variance of the Normally distributed LAI observational errors (s_{LAI}^2), i.e. $p=4$.
- y = f(**x**) is a scalar representing the simulator output, which is Metropolis algorithm's estimate one of the DALEC-D parameters. GEM-SA can only build emulators which have one output. Therefore we are required to build 17 emulators, one for each of the parameters.

3.2.6 Using a Bayesian emulator to carry out sensitivity analysis

In this subsection we outline how the sensitivity was carried out using an emulator. First we describe why carrying out sensitivity analysis (SA) without using an emulator is either not possible or infeasible. Methods such as variance decomposition (e.g. ANOVA) are not possible for most applications of SA since they require the simulator inputs to be independent. When we do not assume independence, in the case of this chapter, a large number of runs of the algorithm (of the order of thousands) are required to perform SA using standard techniques (Saltelli, 2008). O'Hagan (2004) states, 'Even for a simulator that takes just one second to run, a comprehensive variance-based sensitivity analysis may require millions of simulator runs, and just one million runs will take 11.5 days of continuous CPU time.' As a result, building an emulator, which is a statistical model of the input-output relationship of the simulator, has become an attractive alternative because it is computationally much cheaper to run than the simulator. Therefore as long as the emulator is an accurate representation, carrying out SA using the emulator will be very quick; for this chapter GEM-SA took around 30 seconds to perform SA.

Many types of emulators could be adopted. Probably the simplest is the basic linear model in which both the input and output are one-dimensional and they are related by $y=\alpha x+\varepsilon$ where ε is the residual term and β is a parameter to be estimated from the training data, i.e. from the executions of the algorithm required to

build the emulator. Alternatively, higher order polynomials could be used, for example, $y = \alpha x + \beta x^2 + \varepsilon$. However this has a major limitation – it treats the training data as though they were observations. This is wrong because real data contain random error and so are not a true representation of the real world, whereas the training data are the true representation of the simulator. Therefore, the emulator must have the property of returning each of the training outputs when the corresponding training inputs are used. Statistical models of this form are called interpolators. The most basic interpolators are Lagrange polynomials (Berrut & Trefethen, 2004) and are of the form:

$$P_n(x) = \sum_{i=1}^{n} \ell_i(x) y_i$$

$\ell_i(x) = L_i(x)/L(x_i)$, $L_i(x) = (x - x_1)\ldots(x_i - x_{i-1})(x - x_{i+1})\ldots(x - x_n)$, and (x_i, y_i) $1 \le i \le n$ are the training data. However a major problem with this is that if the distance between the training data points is approximately the same then for large n, the interpolator is ill conditioned. This means that slightly different training data would lead to a very different interpolant (Trefethen, 1991, Henrici, 1982). As a result, 7th or 8th polynomials at most must be used.

An Artificial Neural Network (ANN) is another type of emulator (Abrahart & See, 2007). They operate by assigning many simple but different functions to represent the different processes of the model. For example, if x and y are the input and output of an ANN represented by f, then f is a function of other functions, *i*.e. $y = f(x) = h(g_1(x), g_2(x), \ldots)$ where $h, g_1, g_2, \ldots$ are the other functions. However, Gosset (2003) and Han et al. (2007) suggest that neural networks can be complex to build requiring a huge number of runs of the simulator. There are other types of emulators, but like an ANN, these also need hundreds and usually thousands of runs of the simulator.

Of the range of emulators available, those that adopt a Bayesian framework are regarded as being much more computationally efficient as they often require less than a hundred runs of the simulator. Currin et al. (1988, 1991) was first to use Bayesian techniques for emulators using a Gaussian process prior, and since then many others have done the same, e.g. Welch et al. (1992), O'Hagan (1994), Mrawira et al. (1999), O'Hagan and Kennedy (2001), Oakley and O'Hagan (2002), Rougier et al. (2008). The advantage of a Bayesian approach is that prior knowledge about what we believe

the emulator should be like can be expressed. Then, with information from the training runs, we are able to derive the posterior distribution.

The most favoured type of prior distribution to use in the emulator to represent the simulator's output is the Normal distribution. This distribution is the most commonly used in Statistics and contains some mathematical properties which make it easy and efficient to use in creating an emulator. As a reminder, we have adopted 'simulator' in place of the 'Metropolis algorithm' for the purpose of describing how an emulator is built (see subsection 3.2.4). Since we are treating the simulator's output as a function of its inputs, i.e. $y = f(\mathbf{x})$, we use a Gaussian Process (GP) for the prior. A GP is an extension to the multivariate Normal distribution, where the mean and variance are expressed as functions. For a more formal definition of a GP, if we first assume that $\mathbf{g} \subseteq \Re^n$ is a random variable which follows a Normal distribution, then $\mathbf{g} \sim N(\mathbf{m}, \mathbf{V})$, where $\mathbf{m} \subseteq \Re^n$ and $\mathbf{V} \subseteq \Re^{n \times n}$ are the mean and covariance. If **g** is now a function of an independent variable, say time *t*, i.e. **g**(t), and **g**(0), **g**(1), **g**(2), … each follow a Normal distribution, $\mathbf{x}(\mathrm{i}) \sim N(\mathbf{m}(\mathrm{i}), \mathbf{V}(\mathrm{i}))$, for i=0,1,2, …, then we say that **g**(t) follows a Gaussian Process, written as:

$$\mathbf{g}(\mathrm{t}) \sim GP(\mathbf{m}(\mathrm{t}), \mathbf{V}(\mathrm{t}, \mathrm{t}'))$$

where the mean and covariance, i.e. **m**(t) and **V**(t,t'), are now functions and t' is value of the independent variable not equal to t (Bastos & O'Hagan, 2008). A full definition of the Gaussian Process is given in the next subsection.

3.2.7 Building the Bayesian emulator

The theoretical basis behind how an emulator is built using GEM-SA given below is based on Bastos and O'Hagan (2008). Full details can be found Haylock & O'Hagan (1996). It is important to note that emulators had to be built for each of the 17 parameters because, while GEM-SA can build an emulator with multiple inputs, the output must be scalar. Recent developments of Gaussian process emulators have included multidimensional outputs, e.g. Fricker et al. (2010), Rougier (2008) and Conti & O'Hagan (2010). However these are much more complicated and at the time of writing the statistical theory behind their operation has yet to be fully formulated. Although very recent developments have seen an R package for a multivariate emulator be written (Hankin, 2012).

THE PRIOR DISTRIBUTION

The prior distribution of the simulator output is represented by a Gaussian process model, and takes the form:

$$f(\mathbf{x}) \mid \boldsymbol{\beta}, \sigma^2, \mathbf{b} \sim GP(m_0(\mathbf{x}), v_0(\mathbf{x},\mathbf{x}')) \tag{3.1}$$

where $f(\mathbf{x}) \mid \boldsymbol{\beta}, \sigma^2, \mathbf{b}$ represents the output of the simulator given that $\boldsymbol{\beta}$, σ^2 and $\mathbf{b}$ (defined below) are can be estimated using the training data, and the $m_0(\cdot)$ and $v_0(\cdot,\cdot)$ represent the mean and covariance functions given by:

$$m_0(\mathbf{x}) = h(\mathbf{x})^T \boldsymbol{\beta} \tag{3.2}$$

$$v_0(\mathbf{x},\mathbf{x}') = \sigma^2 c(\mathbf{x},\mathbf{x}';\mathbf{b}) \tag{3.3}$$

In equation (3.2), $\mathbf{x}$ and $\mathbf{x}'$ are two different $p\times 1$ input vectors. The function $h(\cdot): \Re^p \rightarrow \Re^{p+1}$ is a known function of the inputs, represented by a $(p+1)\times 1$ vector of functions. We choose $h(\cdot)$ according to the form we think it should take. A common choice, and what is done in GEM-SA is $h(\mathbf{x})^T = [1\ \mathbf{x}^T]$, e.g. for p = 4 and $\mathbf{x}_1$, **…,** $\mathbf{x}_4$ are the elements of $\mathbf{x}$, then $h(\mathbf{x})^T = [1\ \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4]$. In other words, our prior expectation is that the relationship between the inputs and the output is linear. The vector $\boldsymbol{\beta}$ is a vector of coefficients while σ^2 is a scale parameter, estimated using equations (3.5) and (3.6).

For equation (3.3), $c(\mathbf{x},\mathbf{x}';\mathbf{b}) = \sigma^2 \exp\{-(\mathbf{x} - \mathbf{x}')^T \mathbf{B}(\mathbf{x} - \mathbf{x}')\}$. Here $\mathbf{B}$ is known as the roughness matrix, a p×p matrix, with zeros in the off-diagonal elements and diagonal elements represented by the roughness parameters b_1, b_2, …, b_p (where the roughness vector $\mathbf{b}$ is given by $\mathbf{b}^T = [b_1, b_2, \ldots, b_p]$). These give an indication of whether the input-output relationship for each input variable, given the training data, should be linear, as specified by $h(\mathbf{x})$. The linearity is sometimes called the smoothness of the simulator. Low values of the roughness parameters indicate that a simulator is smooth for a particular input variable (i.e. linearity is appropriate) whereas higher values, particularly ones approaching the scaled maximum of 99, suggest the opposite and building an emulator using GEM-SA may not be appropriate.

$\boldsymbol{\beta}$ and σ^2 are estimated by combining a prior distribution for each with the training data using Bayes theorem. Weak prior distributions are assumed for $\boldsymbol{\beta}$, σ^2 and $\mathbf{b}$, i.e. our prior knowledge of these three quantities is weak, so probability

distributions are chosen which do not favour a particular value strongly. In GEM-SA, these are represented by $p(\boldsymbol{\beta}, \sigma^2) \propto \sigma^{-2}$ and $b_i \sim \text{Exp}(0.01)$ for independent b_1, b_2, …, b_p, where $p(\boldsymbol{\beta}, \sigma^2)$ is the joint probability distribution function of $\boldsymbol{\beta}$ and σ^2, and $\text{Exp}(\cdot)$ is the Exponential distribution. Different notation is sometimes used in the literature. For example, the roughness parameters can be written as $\frac{1}{\psi_1}$, $\frac{1}{\psi_2}$, …, $\frac{1}{\psi_p}$, where ψ_1, …, ψ_p are known as the correlation parameters. For diagonal **B**, c(**x**,**x'**;**b**) can also be expressed as c(**x**,**x'**;**b**) = $\exp\{-\sum_{k=1}^{p} b_k (x_k - x_k')^2\}$, where **x**=[x_1, …, x_p] and **x'**= [x_1', …, x_p'].

THE POSTERIOR DISTRIBUTION

The posterior distribution of $f(\cdot)$ is the distribution derived combining the prior distribution and the information about the emulator from the training runs using Bayes' theorem (see subsection 2.3.3 for definition). We begin by specifying the distribution of the outputs of the training data. Let the elements of $\mathbf{y}^{\mathrm{T}}$ = =[y_1, …, y_n] be obtained from *n* runs of the simulator $f(\cdot)$ where $y_i = f(\mathbf{x}_i)$ for i = 1, …, n, and the $\mathbf{x}_i$'s have been chosen according to some suitable design (discussed in subsection 3.2.3) so as to sufficiently sample the input space. Using equation (3.1) the joint distribution of **y** conditional on $\boldsymbol{\beta}, \sigma^2$ and **b** is given by the multivariate Normal distribution:

$$\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2, \mathbf{b} \sim N_n(\mathbf{H}\boldsymbol{\beta}, \sigma^2 \mathbf{A}), \tag{3.4}$$

where $\mathbf{H}^T$=[$h(\mathbf{x}_1)$ $h(\mathbf{x}_2)$ … $h(\mathbf{x}_n)$] and **A** is an $n \times n$ matrix with $\mathbf{A}_{i,j}$ = c($\mathbf{x}_i$,$\mathbf{x}_j$;**b**). The RHS of equation (3.4) is a multivariate Normal distribution, and not a Gaussian process as in equation (3.1), because the mean and covariance terms are now not functions. Using standard techniques for conditioning in multivariate normal distributions, and Bayes' theorem for the posterior distribution of (β , σ^2) it can be shown that, after integrating out these hyperparameters, the posterior distribution for f(**x**) unconditional on β , σ^2 is:

$$f(\mathbf{x}) \mid \mathbf{y}, \mathbf{b} \sim t_{n-q}(m_1(\mathbf{x}), v_1(\mathbf{x},\mathbf{x}')) \tag{3.5}$$

where $m_1(\mathbf{x})$ and $v_1(\mathbf{x},\mathbf{x'})$ represent the respective posterior mean and correlation functions given by:

$$m_1(\mathbf{x}) = h(\mathbf{x})^T\hat{\boldsymbol{\beta}} + t(\mathbf{x})^T\mathbf{A}^{-1}(\mathbf{y} - \mathbf{H}\hat{\boldsymbol{\beta}}),$$

$$v_1(\mathbf{x},\mathbf{x'}) = \hat{\sigma}^2[\, c(\mathbf{x},\mathbf{x'};\mathbf{b}) - t(\mathbf{x})^T\mathbf{A}^{-1}t(\mathbf{x'}) + (h(\mathbf{x}) - t(\mathbf{x})^T\mathbf{A}^{-1}\mathbf{H})(\mathbf{H}^T\mathbf{A}^{-1}\mathbf{H})^{-1}(h(\mathbf{x'}) - t(\mathbf{x'})^T\mathbf{A}^{-1}\mathbf{H})^T\,],$$

with $\hat{\beta}$ and $\hat{\sigma}^2$ given by :

$$\hat{\boldsymbol{\beta}} = (\mathbf{H}^T\mathbf{A}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{A}^{-1}\mathbf{y}$$

$$\hat{\sigma}^2 = (\mathbf{y}^T(\mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{H}(\mathbf{H}^T\mathbf{A}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{A}^{-1}) / (n - q - 2)$$

The expectation of equation (3.5), labelled $\hat{y}$, is the emulator's estimate of the simulator's output, *y*. If the emulator is represented as $\hat{f}(\cdot)$, then $\hat{y} = \hat{f}(\mathbf{x})$, where **x** is a vector of inputs.

COMPUTING THE ROUGHNESS PARAMETERS

One modelling issue is that the posterior distribution of $f(\cdot)$ is conditional on the roughness parameters $\mathbf{b} = [b_1, b_2, \ldots, b_p]$. Therefore these roughness parameters need to be estimated. The posterior distribution of **b** is computed using Bayes' theorem, i.e.

$$\begin{aligned} p(\mathbf{b} \mid \mathbf{y}) &= p(\mathbf{b})\, p(\mathbf{y} \mid \mathbf{b}) \\ &\propto p(\mathbf{b}) \int\int p(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2, \mathbf{b}) \times p(\boldsymbol{\beta}, \sigma^2).d\boldsymbol{\beta}.d\sigma^2 \\ &\propto p(\mathbf{b})|\mathbf{A}|^{\frac{1}{2}}|\mathbf{H}^T\mathbf{A}^{-1}\mathbf{H}|(\hat{\sigma}^2)^{-\frac{n-q}{2}} \end{aligned} \qquad (3.6)$$

A completely Bayesian approach would integrate out **b** from the product of equations (3.5) and (3.6) to leave the posterior distribution $p(f(\mathbf{x})|\mathbf{y})$. A major problem with this is that (3.6) is extremely complex and so determining $p(f(\mathbf{x})|\mathbf{y})$ would be extremely computationally demanding, even if done numerically, resulting in six levels of integrals (Kennedy & O'Hagan, 2001). Instead a plausible estimate of **b** is determined, using the methodology of Kennedy & O'Hagan (2001). Therefore, the emulator is what is given in equation (3.5) but with the estimate of **b** used in the formulae for **A**, $\hat{\beta}$ and $\hat{\sigma}^2$.

3.2.8 Assumptions and diagnostics of emulators built using GEM-SA

In building the emulator given in equation (3.5), various assumptions are made. For example, in the prior we assume a particular form for the mean and covariance functions (equations 3.2 and 3.3). If these assumptions are not valid or other aspects of the emulator are poorly estimated, such as the coefficients in the $\boldsymbol{\beta}$ vector, then the emulator may not perform well in making predictions. These predictions can be checked in two different ways, but before these are explained we first explain a diagnostic check which can be carried out independent of any predictions.

In equation (3.3), it is assumed that the simulator's output responds with a similar level of smoothness at all points in the input space. In reality, the simulator may be more sensitive to changes in some parts of the input than others. As a result, the credible intervals of the emulator predictions may be too wide in regions of low responsiveness or too narrow where the response is more dynamic (Bastos & O'Hagan, 2008). This can be checked by inspecting the roughness parameters. Small roughness parameters suggest that the smoothness assumption is valid; if the parameters are large, and particularly if they are approaching the scaled maximum of 99, then this assumption may not be appropriate (O'Hagan, 2005). For practical purposes, we require most of the roughness parameters to be less than 10 and all to be less than 25, for this assumption is valid.

It is essential to check that the emulator is an accurate representation of the Metropolis algorithm. This can be achieved in two ways.

(1) A cross validation root mean squared standardised error (RMSSE) is computed. This means that alternative emulators are built, with each one leaving one of the training inputs out. The output when training data are missing is then compared with the estimated output from the emulator, and the RMSSE of all the differences is computed. The RMSSE values are also standardised as the outputs of the different emulators have different units and are of different sizes; the standardisation in GEM-SA is scaled so that values close to 1 indicate good performance of the emulator.

(2) Additional runs (often referred to as 'prediction runs') of the Metropolis algorithm (usually about half the number of training runs) are carried out and differences between the resulting and predicted outputs (from the emulator) are

computed. In this chapter we show that 40 training runs are required, thus 20 prediction runs are carried out. The differences are then standardised so that they are all on the same scale for different parameters, by computing:

$$D_i(\mathbf{y}^*) = \frac{y_i^* - E\left[f(x_i^*) \mid \mathbf{y}\right]}{\sqrt{V\left[f(x_i^*) \mid \mathbf{y}\right]}}$$

where i refers to the ith prediction run ($1 \le i \le 20$), $\mathbf{y}_i^*$ and $E\left[f(x_i^*) \mid \mathbf{y}\right]$ are the parameter estimates using the Metropolis algorithm and the emulator (built using training outputs given by $\mathbf{y}$) respectively, and $V\left[f(x_i^*) \mid \mathbf{y}\right]$ is the variance of the emulator's estimate. We then plot $D_i(\mathbf{y}^*)$ against the index i for i=1, …, 20.

Thus, for the emulator to be an accurate representation of the simulator, we require: (i) most of the $D_i(\mathbf{y}^*)$'s are in within the bounds of a 95% tolerance interval, i.e. (-1.96, 1.96); (ii) the RMSSE values to be close to 1, i.e. between 0.8 and 1.2 is adequate.

3.2.9 Latin Hypercube sampling of the input space

In order to build an emulator, consideration needs to be given to how best to choose the training inputs. The design needs to be space filling, i.e. one in which most or all regions of the input space are represented. One option is to sample the input space using a stratified random sample, but this can require a large number of executions of the model when the dimension of the input space is higher than two. An alternative and popular design is a Latin Hypercube design (LHD).

LATIN HYPERCUBE SAMPLING

In Latin hypercube sampling a stratified random sample is applied to each variable (or in each dimension of the input space), then the order is randomly permuted, and one of the permutations is randomly chosen. McKay (1979) summarises the general definition of an LHD. In k dimensions:

> *'Let X_1, …, X_N be a sample of points from the input space. We divide the range of the input variables X_k ($1 \le k \le N$) into N equally spaced strata, and sample once from each stratum. Let this sample be X_{kj}, j = 1,…, N. These form the X_k component, k = 1,…, K, in X_i, i = 1, …, N. The components of the various X_k's are matched at random.'*

An example is now presented to aid understanding. For simplicity we restrict ourselves to two input variables, X1 and X2, and suppose we require n=3 training inputs. Let the known ranges of X1 and X2 be [2,8] and [150, 300] respectively. Since n=3, the domain of each variable is split up into three equally sized sections. One value is randomly selected from each section, for each variable. Denote the two sets of selected points by 1,2,3 and *a,b,c*. Permuting the elements of the two sets with each other gives:

(1,a)	(1,*a*)	(1,*b*)	(1,*b*)	(1,*c*)	(1,*c*)
(2,b)	(2,*c*)	(2,*a*)	(2,*c*)	(2,a)	(2,*b*)
(3,c)	(3,*b*)	(3,*c*)	(3,*a*)	(3,*b*)	(3,*a*)

One of the permutations is now randomly selected, except for the one on the furthest left (for reasons explained after). Suppose the fourth group from the left is chosen, then the design can be represented as shown in figure 3.1a. In essence the three input parts of the required training data are (1,*c*), (2,*a*), (3,*b*). The design on the furthest left is ignored best if 1,2,3 and a,b,c are paired in ascending order, this gives no information about how simulator outputs behave when the two inputs vary in the opposite direction. Figure 3.1b shows this graphically – i.e. if we have a (1,a), (2,b), (3,c) design then the three forward diagonal squares would be used leaving important regions (i.e. the top right and the bottom left) empty.

The key advantage of an LHD is that it ensures that all parts of the domain of each input dimension are well represented, while only requiring a small number of sampled points. It can also work with a high number of dimensions. The concept of an LHD was first proposed by McKay et al. (1979) and has since become a popular design for computer experiments, particularly those which are expensive to run (for example, see Butler (2001), Santner (2003) and Wang (2003)).

MAXIMIN LATIN HYPERCUBE SAMPLING

A *maximin* Latin Hypercube Design (MmLHD) is a particular class of LHD which has been shown to sample the input space better than standard LHD. In the example above there were five possible sets of pairings for the two inputs, and an LHD consists of randomly selecting one of these five sets. The MmLHD chooses the set in an optimal way by maximising the distances between the three input points. We refer to figure 3.1a and 3.1b to understand this. The design in figure 3.2b was disallowed

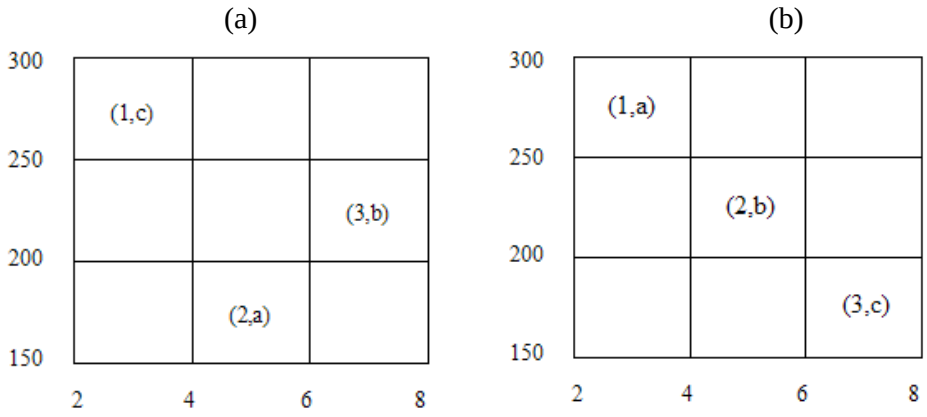


Figure 3.1 A diagrammatic representation of an example set of three points chosen using a Latin Hypercube Design for two inputs.

because there were large gaps in the input space being ignored, whereas the design in figure 3.1a was better because the points were more spread out in the input space. Another way of viewing this is if you were to sum the distances between the points for the bad design (figure 3.1b), this would be less than the sum from the design in figure 3.1a. Morris & Mitchell (1995) provide full details of the MmLHD, and the appendix displays the R script used to generate the inputs using this design.

DECIDING ON THE NUMBER OF TRAINING RUNS

If the simulator is not too computationally expensive and the input dimension is not excessive, then Loepky et al. (2008), on the basis on experiment, suggest the number of training runs should be at least 10 times the input dimension. This rule of thumb is consistent with a number of other papers where Gaussian Process emulators were built using an appropriate space filling design, for example, Bastos & O'Hagan (2008), Gosling (2006), Rougier et al. (2008), and Kennedy et al. (2006). In this chapter there are 4 input variables, therefore 40 training runs of the Metropolis algorithm were carried out when building each of the 17 emulators.

3.2.10 Sensitivity analysis using the emulator

Sensitivity analysis (SA) is concerned with quantifying how changes in the output of a function can be attributed to changes in each input variable x_j (j = 1,…,p). For this chapter, p=4, the function is each of the 17 emulators. Thus, labelling the *i*th output and emulator as $\hat{y}_i$ and $\hat{f}_i(\cdot)$ $(1 \leq i \leq 17)$, the function on which we wish to

perform sensitivity analysis can be expressed as $\hat{y}_i = \hat{f}_i(x_1, x_2, x_3, x_4)$. To explore how *y* changes with x_1, for example, one option is to fix the other inputs at nominal values (e.g. their central values) whilst varying x_1. However, this may lead to biased results as the sensitivity of *y* to changes in x_1 may vary depending on the chosen fixed values of x_2, x_3 and x_4. A better option would be to average over the other inputs (Santner, 2003), i.e. the sensitivity of $\hat{y}_i$ to changes in x_1 would be determined by computing the following expectation for multiple values of x_1:

$$E[\hat{y}_i \mid x_1] = \int \int \int \hat{f}_i(x_1, x_2, x_3, x_4)\, p(x_2, x_3, x_4).dx_2 dx_3 dx_4, \qquad (3.7)$$

where $p(x_2, x_3, x_4)$ is the probability distribution of x_2, x_3 and x_4. This is what is done in GEM-SA.

THE SENSITIVITY INDEX AND TOTAL EFFECTS

The sensitivity of $\hat{y}_i$ to changes in x_j $(1 \le j \le 17)$, is calculated in GEM-SA using the methods set out in Santner (2003) and Oakley & O'Hagan (2004). This sensitivity can be quantified by specifying the percentage of variance in the output which is apportioned to each input variable. If $V_j = Var_{x_j}(E[\hat{y}_i \mid x_j])$ denotes the variance of $E[\hat{y}_i \mid x_j]$, we define the *first order sensitivity index* (S_i) for this input as:

$$S_j = \frac{V_j}{V},$$

where *V* is the total variance of the output. In other words S_j measures the proportion of the total variance *V* due to the main effect x_j, i.e. excluding all interactions with other inputs. For $j_1 < j_2$, $S_{j_1 j_2}$ is the *second order sensitivity index*, written:

$$S_{j_1 j_2} = \frac{V_{j_1 j_2}}{V},$$

with $V_{j_1 j_2} = Var_{x_{j_1} x_{j_2}}(E[y \mid x_{j_1 j_2}]) - V_{j_1} - V_{j_2}$, measures the interaction effect due to inputs x_{j_1} and x_{j_2}; in other words it measures the proportion of the total variance above that of their main effects. This can be extended to higher orders. In general for

$s=1,\ldots,p$ and $1\le i_1<\ldots<i_s<\ldots\le i_p$, the *s-th order sensitivity index* is defined by:

$$S_{j_1,\ldots,j_s}=\frac{V_{j_1,\ldots,j_s}}{V}$$

where $V_{j_1,\ldots,j_s}=Var_{x_{ji_1,\ldots,j_s}}(E[y\mid x_{j_1,\ldots,j_s}])\ -\ V_{j_1}\ -\ V_{j_2}-\ldots-\ V_{j_s}$. By construction, the sensitivity indices satisfy (Santner, 2003):

$$\sum_{j=1}^{p}S_j+\sum_{1\le j_1<j_2\le p}^{p}S_{j_1j_2}+\ldots+S_{1,2,\ldots,p}=1$$

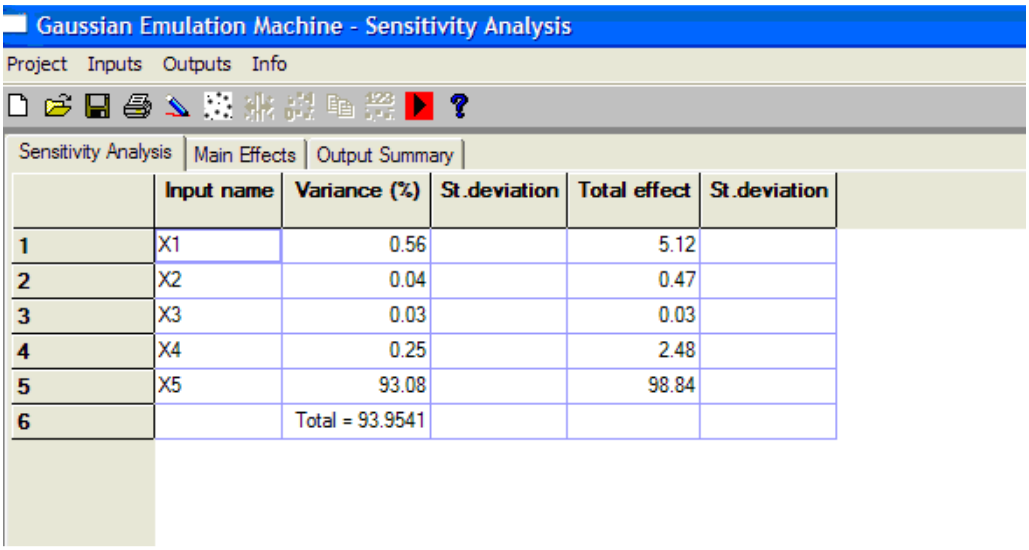
In GEM-SA, sensitivity indices for all orders are taken into account but only the first and second order orders are displayed. The total sensitivity of *y* to input x_j is defined as the sum of all orders of sensitivity indices involving that input, i.e.

$$T_j=S_j+\sum_{j_1>j_2}S_{j_1j_2}+\sum_{j_1<j_2}S_{j_2j_1}+\ldots+S_{1,2,\ldots,p}.$$

HOW SENSITIVITY ANALYSIS IS SUMMARISED IN GEM-SA

When an emulator is built in GEM-SA it produces tabular and graphical output for Sensitivity Analysis. The tabular form gives, for each x_i, the first order (and second order if specified) sensitivity indices and the value of T_i. To aid understanding, table 3.2 shows this tabulated form using a made up example which involves five inputs labelled X1, X2, X3, X4 and X5. For this example we can see that the variance of the output can be mainly attributed to changes in the X5 input. Note that the first order sensitivity indices sum to 93.95% which is close to 100%, so the contribution to the variance of second and higher order interactions (not shown) will be very small.

The graphical output consists of a plot of $E[y\mid x_i]$ against x_i for each of the inputs. This consists of a band of lines, rather than a single line, since $\hat{f}_i(\cdot)$ is uncertain; the thickness of the band is an indicator of the emulator uncertainty. If the band is approximately horizontal, the output is not sensitive to changes in that input. Figure 3.2 shows the graphical output from the example that yielded Table 3.2. We can see that the function is insensitive to the first four inputs but is significantly affected by X5, which corresponds to the significantly higher percentage of the variance for X5 compared to the other inputs in table 3.2. We can also see that all of



	Input name	Variance (%)	St.deviation	Total effect	St.deviation
1	X1	0.56		5.12	
2	X2	0.04		0.47	
3	X3	0.03		0.03	
4	X4	0.25		2.48	
5	X5	93.08		98.84	
6		Total = 93.9541			

Table 3.2 The tabulated form of the sensitivity analysis carried out in GEM-SA using a mock example with 5 inputs.

the sensitivity of the output due to X5 is for the higher values of X5.

In this chapter, only the graphical form is shown, because we are only interested in how $E[\hat{y}_i \mid x_j]$ varies for different x_j. The tabular form shows the variance of $E[\hat{y}_i \mid x_j]$, relative to the total variance V (see previous page), which essentially gives the same information as the graphical form but in a numerical format. In certain situations the tabular form can be useful, but for this chapter it is not.

CODE UNCERTAINTY

An extra source of uncertainty when carrying out SA on an emulator results from the fact that if SA had been carried out using the Metropolis algorithm directly, the values of S_i and T_i would differ from those given by the emulator, because the emulator is only an approximation of the Metropolis algorithm. This uncertainty is called *code uncertainty* and will be negligible as long as the emulator accurately represents the input-output relationship of the Metropolis algorithm (O'Hagan, 2004).

3.2.11 Follow-on experiments

As well as the main sensitivity analysis results, the results section (section 3.3) shows additional sensitivity results. These were carried out in response to the main results in order for the main results to be meaningful. The complete rationale behind these addit-

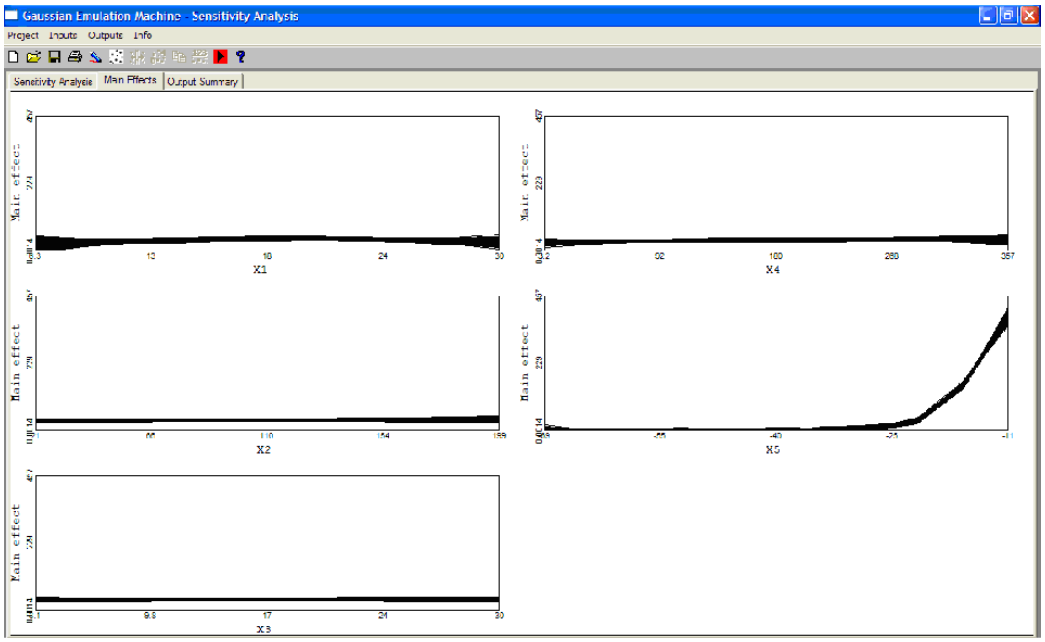


Figure 3.2 The graphical form of the sensitivity analysis carried out in GEM-SA using a mock setup with 5 inputs.

ional experiments is given in the discussion section (3.4), as it is premature to explain this here. However, a description of these extra experiments is given here. For the first of the extra experiments, ten further runs of the Metropolis algorithm are using the same data and setup as the main experiment except carried out with three of the inputs – $C_r(0)$, $C_{lit}(0)$ and the LAI observational error – kept fixed. The other input $C_{lab}(0)$, was varied using a stratified random sample (which is the same thing as a Latin Hypercube design in one dimension. The second of the extra experiments was the same as the first except that an extra 90 LAI data points were used. As explained above, full rationale for these extra experiments are given in the discussion section.

3.2.12 Application of the Metropolis algorithm treating the initial conditions of the small C pools as parameters

To complement the main set of experiments, a further run of the Metropolis algorithm is performed using the same data and exactly the same setup as the runs from the main sensitivity analysis, except that $C_{lab}(0)$, $C_r(0)$, $C_{lit}(0)$ are treated as parameters, with the LAI observational error set at the most common value used by the REFLEX participants. As stated at the end of subsection 3.1.2, this will enable us to answer two questions:

(i) Is it better to treat the initial conditions as fixed or as parameters?

(ii) If we treat the initial conditions as parameters, what impact does wrongly estimating them have on the other model parameters and the C pools?

3.3 Results

In this section, we first assess the convergence of the Metropolis algorithm runs before examining the diagnostics of the emulators which check that the 17 constructed emulators accurately represent the input-output relationship of the Metropolis algorithm, where the inputs are the initial conditions of C_{lit}, C_{lab} and C_r, and the LAI observation error, and the output is the estimate of each of the 17 parameters in DALEC-D (for convenience we name each emulator after the DALEC-D parameter represented as its output). The main sensitivity analysis results are then displayed. Finally the sensitivity analysis results to the additional experiments, as described in subsections 3.2.10 and 3.2.11, are given.

3.3.1 Checking convergence of the Metropolis algorithm

As stated at the end of subsection 3.2.4, the convergence of the Metropolis algorithm was assessed by monitoring the Gelman criteria statistic Rhat for each of the iterations (Gelman, 1995). The Rhat trajectories for each parameter for 4 of the 40 randomly selected runs of the Metropolis algorithm used to construct the emulators is given in figure 3.3. The remaining 36 runs had very similar plots. For each subplot, there are 17 blue lines which represent the Rhat trajectories of each of the 17 DALEC model parameters. For each of the four runs represented by the subplots in figure 3.3, which was also true for the other 36 runs,we can see that all of the parameters have converged by approximately 2.5 iterations, where convergence is determined by the Rhat values going less 1.2 (the dashed red line) (Gelman, 1995). Moreover, in each of the runs most or all of the parameters have converged very early on by around 1 million iterations. The trajectories of the chains for each of the parameters for a randomly selected run can be found in Appendix B.

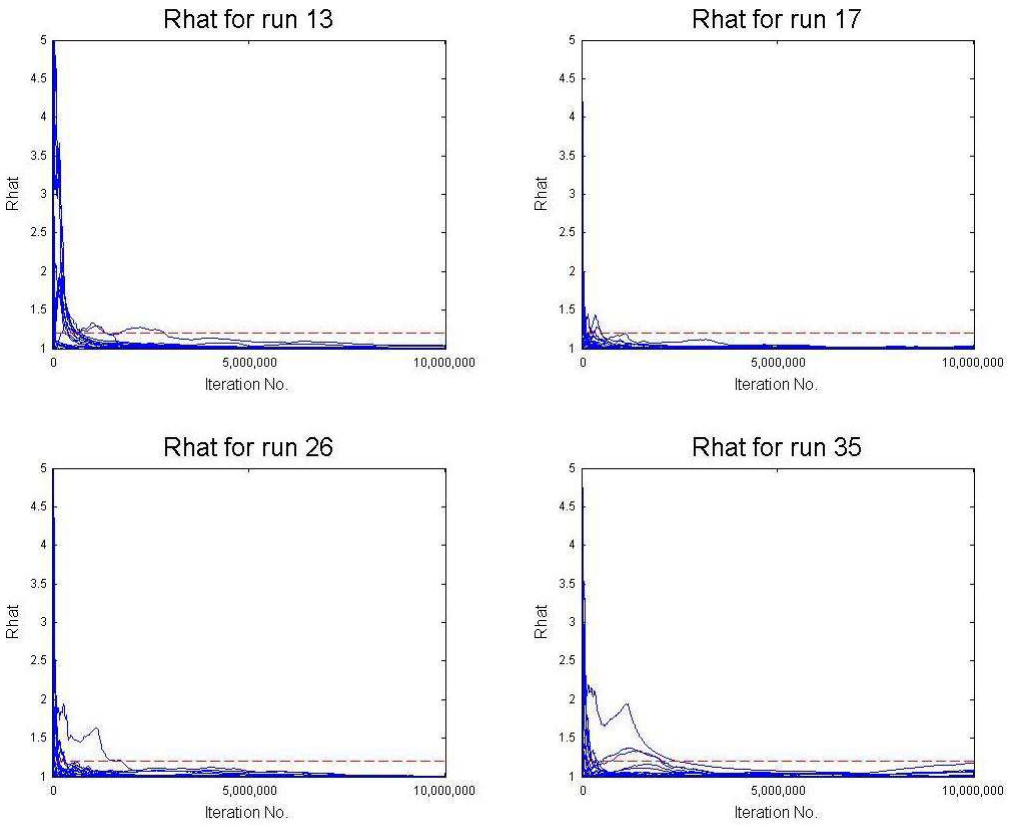


Figure 3.3 The trajectory of Gelman criteria statistic (Rhat) values for each of the 17 parameters for 4 of the 40 runs of the Metropolis algorithm used to the built the emulators for the main sensitivity analysis results.

3.3.2 Emulator Diagnostics

A description of the diagnostics carried out is given in defined in subsection 3.2.8. First we check that the assumption of the smoothness of the simulator outputs to its inputs (equation 3.3) is valid by outputting the roughness parameters (table 3.3) and observing whether any of them are large. As stated in in subsection 3.2.7, small roughness parameters suggest that the smoothness assumption is valid, whereas if the parameters are large, and particularly if they are approaching the scaled maximum of 99, then this assumption may not be appropriate (O'Hagan, 2005). For practical purposes, we require most of the roughness parameters to be less than 10 and all to be less than 25, for this assumption is valid. From table 3.3 we can see that for two of the emulators, corresponding to the T_w (turnover of woody carbon) and F_{lr} (fraction of labile carbon respired) parameters, the value of one of the roughness parameters is greater than 50. Hence the smoothness between the simulator's inputs and outputs can-

Parameter	C_{roots}	C_{litter}	C_{labile}	LAI obs. error
T_d	6.59	0.00	15.08	2.72
F_g	4.42	1.26	5.37	2.45
F_{nf}	0.31	0.10	5.42	1.81
F_{nrr}	1.37	3.29	6.49	0.59
T_r	1.07	6.02	3.75	0.73
T_w	62.35	3.02	1.74	0.00
T_r	0.73	2.91	7.22	1.01
T_l	0.88	4.90	0.78	0.14
T_s	5.36	2.43	1.08	1.13
E_t	0.00	4.81	4.23	1.11
P_r	1.41	0.62	7.08	1.09
L_{out}	0.39	0.86	12.23	3.40
L_{fall}	1.34	3.76	4.27	0.44
F_{ll}	0.42	0.98	10.43	1.44
T_{lab}	0.15	0.65	12.07	0.36
F_{lr}	1.58	49.03	11.19	0.00
C_{fmax}	0.11	0.39	21.48	1.06

Cross-validation RMSSE
0.98
1.03
1.18
1.09
1.03
0.89
1.09
0.99
1.14
1.05
1.19
0.94
1.24
0.83
0.98

Table 3.3 (left) shows the values of the four roughness parameters for each of the 17 emulators constructed. Two emulators (T_w and F_{lr}) had high values of one roughness parameter in each case (illustrated by shading). Table 3.4 (right) shows the cross-validation root mean squared standardised errors (RMSSE) for each of the 17 emulators, which are all close to 1.

not be assumed here and so the sensitivity analysis results for these two emulators are not shown.

Table 3.4 shows the cross validation root mean squared standardised errors (RMSSE), as described in (1) in subsection 3.2.6, for the remaining 15 emulators. All of them have values close to 1 indicating that the emulators are all performing well.

Figure 3.4 shows the plots of the standardised prediction errors, as described in (2) in subsection 3.2.8, of the Metropolis algorithm's outputs versus the outputs predicted from the emulator for two of the parameters. We require the prediction errors to be in the range (-1.96, 1.96), (i.e. between the 2.5^{th} and 97.5^{th} percentiles of a N(0,1) distribution), though if only a few of the errors are outside this range these can be ignored (Bastos & O'Hagan, 2008). For twelve of the parameters (e.g. the fraction of NPP allocated to fine roots, F_{nrr}, left panel of figure 3.4) the prediction errors meet this requirement. For the remaining five parameters (e.g. the turnover of litter carbon, T_l, right panel of figure 3.4) too many of the prediction errors are outside the interval (-1.96, 1.96), suggesting that the emulator is incorrectly estimating the parameter estimates from the Metropolis algorithm for too many regions of the input space.

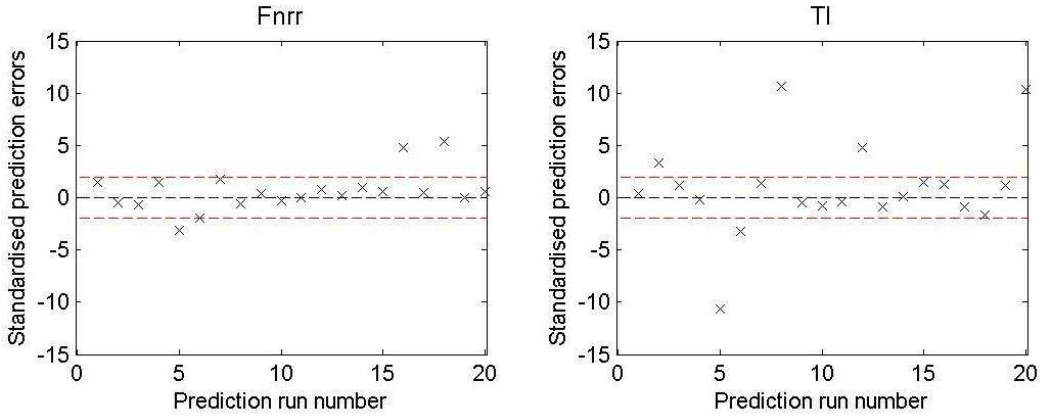


Figure 3.4 Plots of the standardised errors of the prediction runs. For 12 of the 17 parameters (e.g. F_{nrr}, left panel) the vast majority of the errors are within the interval (-1.96, 1.96), the bounds of which are represented by the red dashed lines. For the remaining 5 parameters (e.g. T_l, right panel) more than a few of the errors are outside the interval (-1.96, 1.96).

In summary, the diagnostic checks show that ten of the seventeen emulators accurately represent the input-output relationship of the Metropolis algorithm. The remaining seven emulators did not model the input-output relationship of the Metropolis algorithm good enough. For two of these seven, corresponding to the T_l (turnover of litter carbon) and F_{lr} (fraction of labile carbon respired) parameters, there is insufficient evidence that the assumption of the mean function of the prior being linear is valid; this is essential because it is like saying that a linear model (i.e. $y=ax+b$) to represent data which has curvature between the points when plotted is more appropriate than say a quadratic model, i.e. $y=ax^2+bx+c$. For the other five (corresponding to the T_f, T_r, T_l, E_t and L_{fll} parameters), too many of the prediction errors are too high. Therefore these seven poorly performing emulators were excluded from the results since the sensitivity of the corresponding parameters to changes in the four inputs based on the emulator are likely to be very different than if the sensitivity analysis had been performed using just the Metropolis algorithm.

The fact that seven of the seventeen DALEC-D parameters were excluded from the results did not matter because of the twelve parameters which varied among the DA schemes in REFLEX, only three were excluded. Thus using these remaining ten emulators still gives us great insight into why the estimates of these twelve parameters in REFLEX varied so much among the different DA schemes.

3.3.3 Sensitivity analysis

The sensitivity analysis aims to answer the following questions:

- Which region of the input space are the parameter estimates are sensitive to changes in the inputs?
- Which region of the input space yield the parameter estimates close to the truth?

The results described below show that changes in the C_{lab} initial conditions have the biggest impact on the parameter estimates, with the C_{lit} initial conditions and the LAI observational error having some impact, and the C_r initial conditions having little effect. We abbreviate the initial conditions of the three C pools as $C_{lab}(0)$, $C_{lit}(0)$ and $C_r(0)$.

INITIAL VALUE OF THE LABILE C POOL

Of the ten parameters whose emulators passed the diagnostics (subsection 3.3.1), the estimates of eight were sensitive to $C_{lab}(0)$ (figure 3.5 and Appendix B). For six of these eight parameters (T_d, F_g, F_{nrr}, P_r, F_{ll}, L_{out}), the estimates were most sensitive to small $C_{lab}(0)$ values but were insensitive for larger values (e.g. P_r and L_{out} see figure 3.5a and 3.5c; plots of T_d, F_g, F_{nrr} and F_{ll} are shown in figure B3 of appendix B). For the remaining two parameters (F_{nf} and T_{lab}), the estimates were most sensitive to larger $C_{lab}(0)$ values (figure 3.5b and 3.5d).

The true value of each parameter, sometimes referred to simply as the truth, is given by the red dashed horizontal line in figure 3.6 and Appendix B. We know the true values of the parameters because the assimilated data used was synthetic; see subsection 3.2.3 for details. The parameter estimates were closer to the truth for larger values of $C_{lab}(0)$ for four of the eight parameters (T_d, F_{nrr}, P_r, F_{ll}) and closer to the truth for smaller values of the $C_{lab}(0)$ for two of the parameters (F_g and F_{nf}). For the remaining two parameters (L_{out} and T_{lab}), there is no specific range of $C_{lab}(0)$ over which the parameter estimates are closest to the truth. Therefore, if $C_{lab}(0)$ is either under or over-estimated, we will get at least some parameters correct, but not all. However, for all parameters except T_{lab}, the estimates were close to the truth when the true $C_{lab}(0)$ value (vertical dashed green line) was used. For example for P_r, F_{nf}, and L_{out} see figure 3.5(a)-(c). This highlights the importance of accurately estimating the initial conditions of this particular carbon pool.

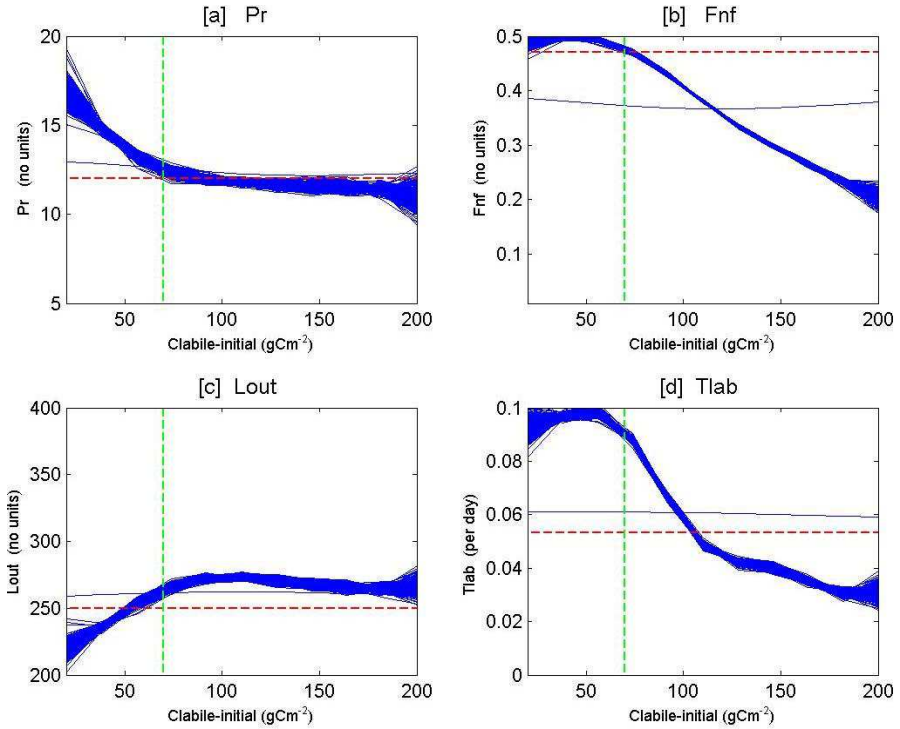


Figure 3.5 The expected value (blue lines) of the (a) P_r, (b) F_{nf}, (c) Lout and (d) T_{lab} parameters for different initial values of the C_{lab} pool. The red horizontal and green vertical dashed lines indicate the true values of the parameter and the true initial value of the C_{lab} pool respectively.

The two parameters, T_s and C_{fmax} whose estimates were insensitive to changes in $C_{lab}(0)$, this is indicated in figures B4(e) and (f) (appendix B) by the fact that the band of blue lines is almost horizontal.

INITIAL VALUE OF THE LITTER C POOL

The parameter estimates were sensitive to $C_{lit}(0)$ for five of the ten parameters, namely F_g, F_{nrr}, T_s, P_r, F_{ll} (figure 3.6 and figure B5 from appendix B). However this sensitivity was, for four of the five parameters (all but F_{nrr}, figure 3.6c), much weaker than the sensitivity to $C_{lab}(0)$; i.e. the gradient of the blue band of lines for the sensitivity plots (e.g. figure 3.6a-b and figure B5 from the appendix B) are much less than the blue band of lines in the $C_{lab}(0)$ sensitivity plots. For the remaining five parameters their estimates were insensitive to changes in $C_{lit}(0)$.

For P_r (figure 3.6a) and F_{ll} their estimates were insensitive for smaller $C_{lit}(0)$ values. For the other three parameters, F_g, T_s and F_{nrr} (e.g. figure 3.6b-c) the estimates were insensitive and for larger $C_{lit}(0)$ values. For F_g, P_r and F_{ll}, when $C_{lit}(0)$ was less than 100gC/m² this resulted in parameter estimates closer to the truth (i.e. closer to the

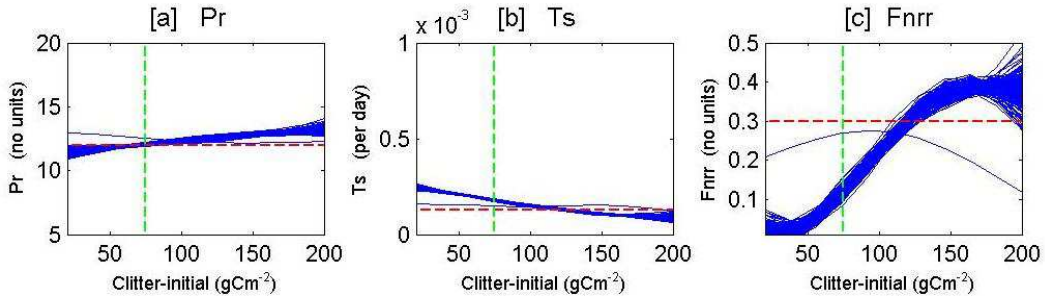


Figure 3.6 The expected value (blue line) of the (a) Pr, (b) Ts, and (c) Fnrr parameters for different initial values of the Clitter pool. The red horizontal and green vertical dashed lines indicate the true values of the parameter and the true initial value of the Clitter pool respectively.

vertical green dashed line). However when the input was greater than 100gC/m² the estimates of T_s and F_{nrr} were closer to the truth.

For the five parameters which were insensitive to changes in $C_{lit}(0)$, the sensitivity plots are not shown but were very similar to figures B5 (e) and (f) (appendix B), i.e. the band of blue lines is horizontal.

INITIAL VALUE OF THE FINE ROOTS C POOL

Only two of the ten parameter estimates (T_d and F_{nrr}) were sensitive to $C_r(0)$ (see figure B6 in appendix B). For T_d, the parameter estimates were sensitive for most of the range of this input but particularly for values of $C_r(0)$ between 75gC/m² and 175gC/m². For values of $C_r(0)$ in the ranges 0–75gC/m² and 175–200gC/m², the parameter estimates were approximately constant and were also closest to the truth. For F_{nrr}, its estimate was insensitive to changes in $C_r(0)$ (i.e. the band of blue lines is flat) of between 50gC/m² and 125gC/m², but decreased and became further from the truth when $C_r(0)$ was less than 50gC/m² and greater than 125gC/m². The parameter estimates were also closest to the truth when $C_r(0)$ was between 50gC/m² and 125gC/m².

For the eight parameters which were insensitive to changes in $C_r(0)$, the sensitivity plots are not shown but were very similar to figures B5(e)-(f) (appendix B).

LAI OBSERVATIONAL ERROR

Figure 3.7 shows six of ten parameter estimates (namely T_d, F_{nf}, T_{lab}, F_g, F_{nrr} and F_{ll}) which were sensitive to the LAI observational error. However the variation in the parameter estimates over the range of this input was far less than due to $C_{lab}(0)$. For

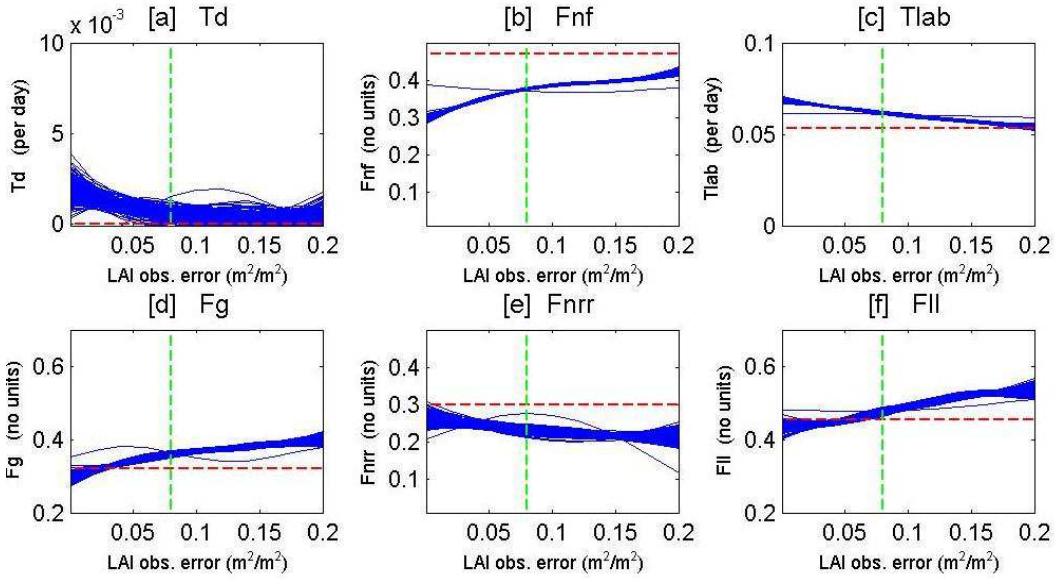


Figure 3.7 The expected value (blue line) of the T_d (a), F_{nf} (b), T_{lab} (c), F_g (d), F_{nrr} (e), and F_{ll} (f) parameters for different values of the LAI observational error. The red horizontal and green vertical dashed lines indicate the true values of the parameter and the true value of this input respectively.

two of these six parameters (T_d and F_{nrr}), the parameter estimates were mainly sensitive to changes to small values (i.e. less than 0.1 m^2/m^2) in the LAI observational error (figures 3.7a,e), while F_{ll} was mainly sensitive to changes in higher values (i.e. greater than 0.05 m^2/m^2). For the other three, the parameter estimates were sensitive for the entire range of the input.

The estimates for T_d, F_{nf}, and T_{lab} are closer to the true value of the parameter (red dashed line) when this input is greater than its true value of 0.08 m^2/m^2 (figures 3.7a-c), while the other three are closer to the truth when this input is less than 0.08 m^2/m^2 (figures 3.7d-f). In REFLEX, five of the Metropolis algorithms used an LAI observational error of between 0.04 and 0.11 gC/m^2, while the other one used values of 0.0005. Therefore this specification of different LAI observational errors may have contributed to the differences observed in the parameter estimation results from the REFLEX project.

3.3.4 Additional sensitivity analysis experiments

We first give the results of the two sets of follow-on experiments, labelled as experi-

ment 1 and experiment 2, as described in subsection 3.2.10. For the additional 20 runs of the Metropolis algorithm, two sets of 10 runs for each of experiments 1 and 2, the algorithm converged well before 5 million iterations and the Rhat plots were very similar to the ones in figure 3.3.

For experiment 1, all but one of the parameters are sensitive to $C_{lab}(0)$ is an almost identical way to the original sensitivity analysis results when the emulators were used, for example F_{nf} and T_{lab} (panels b and h of figure 3.8). For the remaining parameter, P_r, we can see from panel (e) of figure 3.8 that it is closer to the true value for a wider range of $C_{lab}(0)$ values than the original sensitivity plot (panel (d) of figure 3.8). For experiment 2, approximately half of the parameters had almost identical sensitivity plots to those from experiment 1, e.g. F_{nf} (panel c of figure 3.8). For all of but one of the remaining parameters, they were further from the truth compared to the experiment 1 plots (e.g. Pr, figure 3.8(f)). For the remaining parameter, T_{lab}, the blue line of sensitivity of this parameter to changes in $C_{lab}(0)$ make a horizontal shift to the left such that when the true $C_{lab}(0)$ is used this results in the true value of T_{lab} being estimated by the Metropolis algorithm. This is significant because in the original sensitivity analysis results, this was the only parameters whose estimate was not at or very close to the truth when the true $C_{lab}(0)$ was used. These results are discussed in the second half of subsection 3.4.2.

3.3.5 Treating the initial conditions as parameters

In this final run of the Metropolis algorithm, convergence was once again checked using the Gelman criteria as with all previous runs. As with the runs for experiments 1 and 2, the Rhat plot was very similar to those from figure 3.3. Figure 3.9 shows the plots of the estimates of the parameters, represented by the modal values from the posterior distribution, with the lower and upper bounds of the uncertainty intervals represented by the 5^{th} and 95^{th} percentiles. The parameter estimates and uncertainty broadly correspond to those from the 2 year run of the Metropolis algorithm from chapter 4, where differences between the two posterior distributions exist because of the different site used in each of chapters 3 and 4 and the fact that LAI data are used in chapter 3 but not 4. The fit of the modelled NEE and LAI to the data is very good. We can see this in figure 3.10a, where 500 parameter sets were randomly chosen from the posterior distribution and the corresponding trajectories of the NEE and LAI were

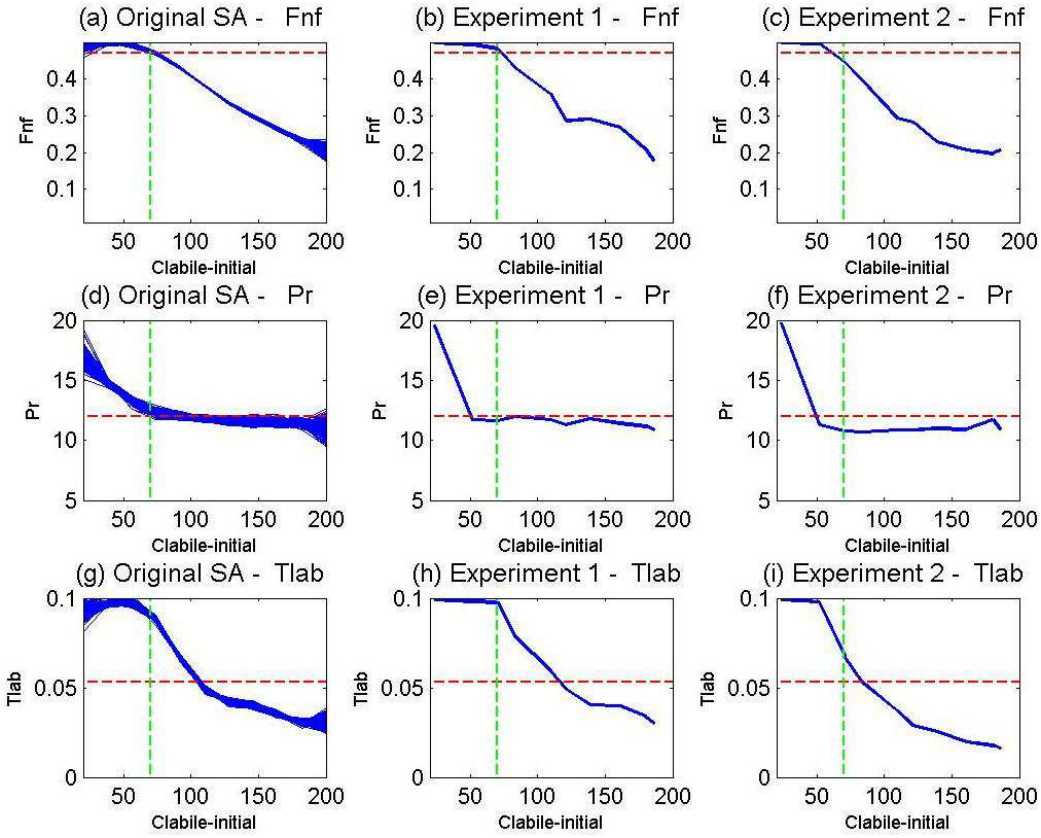


Figure 3.8 Graphical representation of the sensitivity analysis carried out in GEM-SA where $C_{lit}(0)$, $C_r(0)$ and the LAI observational error were kept fixed (Experiment 1) and where 90 extra LAI observations were also assimilated (Experiment 2).

plotted (shown in black), plotted alongside the trajectory of the true NEE and LAI (shown in red) and the NEE and LAI data used in the Metropolis algorithm (shown by the green crosses). For the NEE, we can see that the fit is very good compared to both the true NEE and the data. For the LAI, the fit is also good although the post-DA LAI is not quite reaching the summer maximum values of the true NEE. This is probably due to the fact that there were only a very small number of LAI data used (10), so the NEE data (of which were 231 data points) had a much stronger influence on the posterior distribution.

When the post-DA trajectories of the other C pools are plotted (figure 3.10b), we see reasonable correspondence with the true trajectory (i.e. the red line is mostly within the black area). This is encouraging given that for two of the three initial conditions treated as parameters, namely $C_r(0)$ and $C_{lab}(0)$, had posterior estimates far from the truth and whose posterior intervals did not include the truth. This is discussed in subsection 3.4.6.

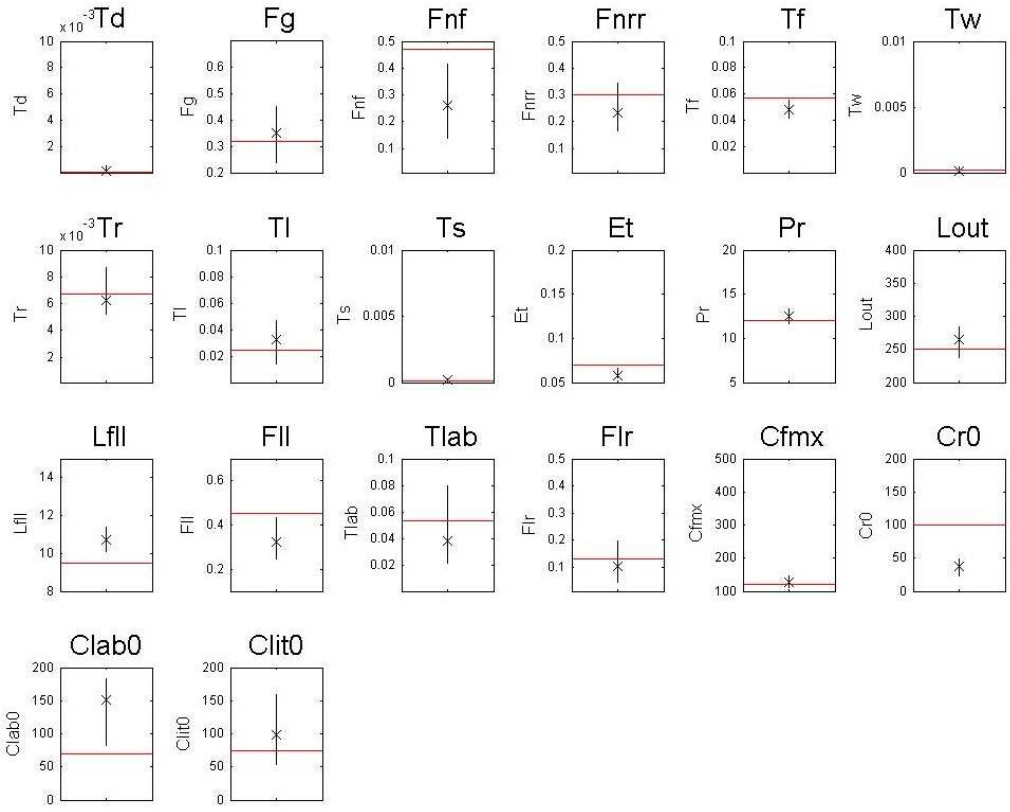


Figure 3.9 The parameter estimates given by the modal value of the posterior distribution along with the corresponding uncertainty bounds given by the 5th and 95th percentiles of the posterior distribution. The initial conditions of the root C, labile C and litter C pools are included as parameters. The red line corresponds to the truth.

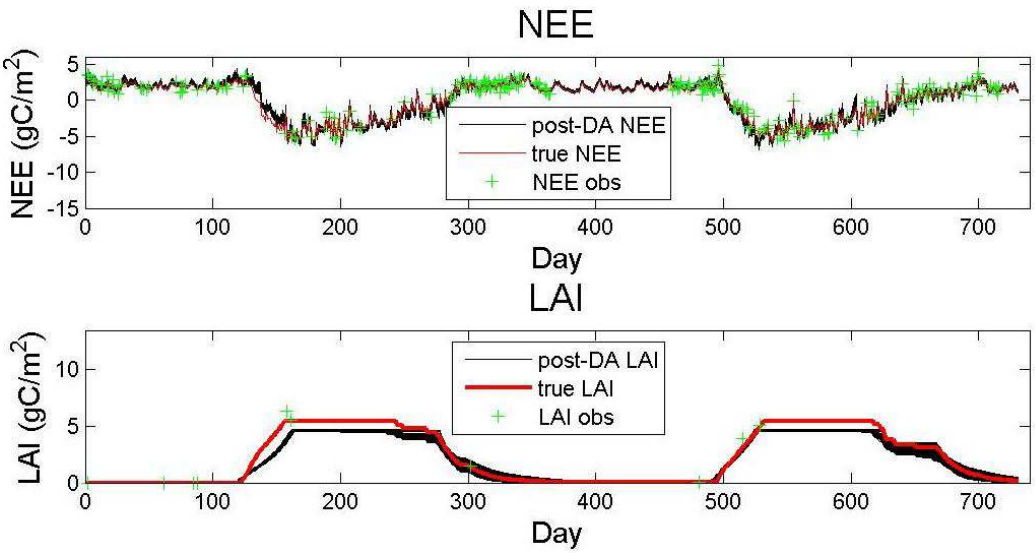


Figure 3.10a The black line are the NEE and LAI trajectories corresponding to 500 randomly selected parameter sets from the posterior distribution. The red line is the corresponding trajectory from using the true parameters and the green crosses are the data.

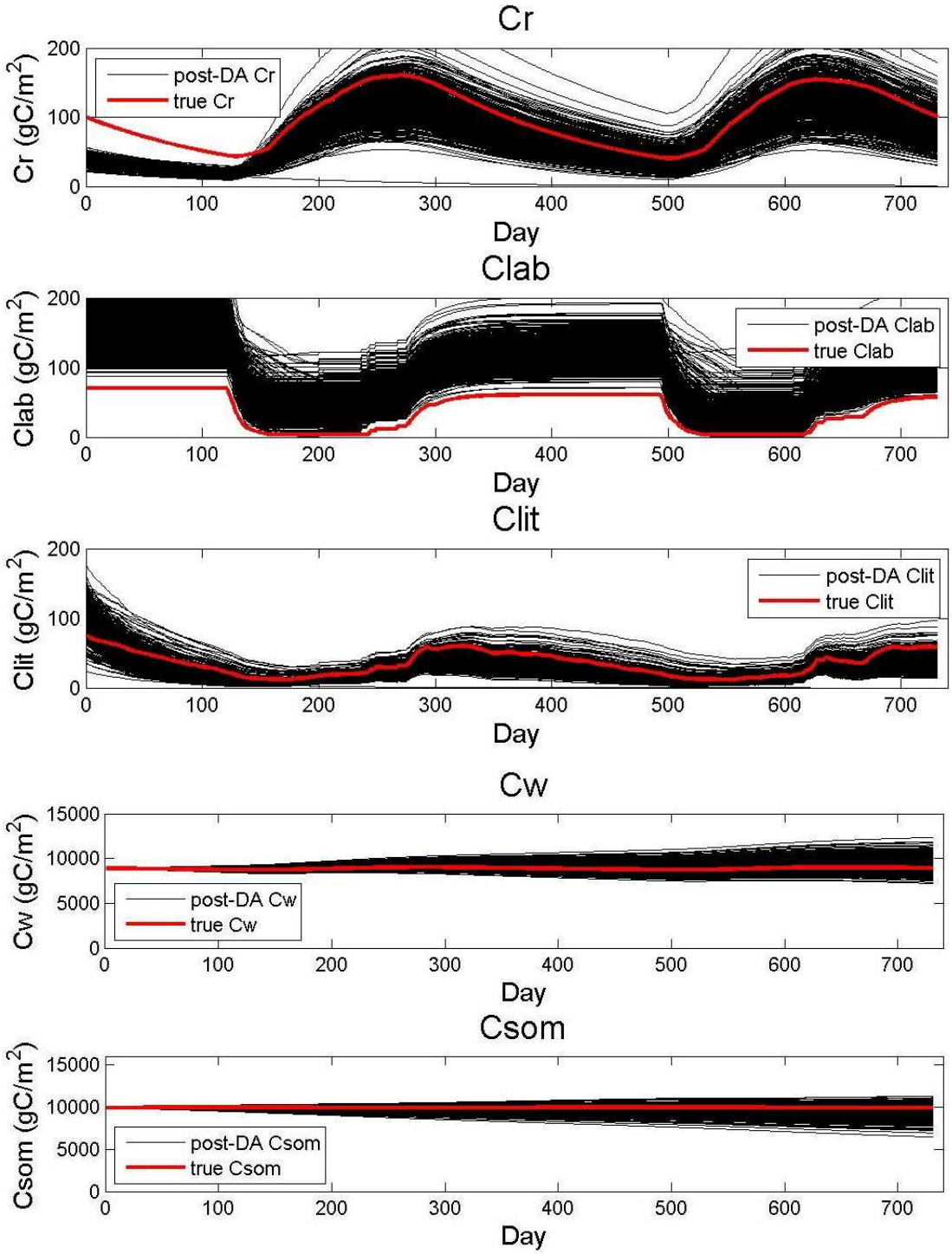


Figure 3.10b The trajectories (black lines) of the C pools (except for C_f) obtained from forward runs of the model using 500 randomly selected parameter sets from the posterior distribution. The red lines are the corresponding true trajectories.

3.4 Discussion

Section 3.3 has shown that $C_{lab}(0)$ causes the largest variation for more parameters than any of the other three inputs, and that the majority of the parameter estimates are close to their true values when the true value $C_{lab}(0)$ is used. However, some important questions arise from these results:

(i) Why is $C_{lab}(0)$ more important than $C_{lit}(0)$ and $C_r(0)$, and why were T_s and C_{fmax} insensitive $C_{lab}(0)$?

(ii) Is it a coincidence that the range of $C_{lab}(0)$ values for which the parameter estimates are close to the truth are either less than or greater than the true value of $C_{lab}(0)$?

(iii) What are the consequences when the right value for $C_{lab}(0)$ is not available (which is almost always the case)?

(iv) What other information can we use in order to accurately estimate the parameters if $C_{lab}(0)$ is poorly estimated?

(v) Why were more parameters sensitive to $C_{lit}(0)$ than to $C_r(0)$, and what inference can be made from the LAI observational error sensitivity plots?

(vi) Do these results help to explain the differences in the parameter estimates from the different DA schemes used in REFLEX?

(i)-(iii) is addressed in subsections 3.4.1-3.4.2, and (iv)-(vi) are assessed in 3.4.3-3.4.6. Finally, subsection 3.4.7 attempts to interpret the estimates of the parameters and uncertainties when Cr(0), $C_{lab}(0)$, $C_{lit}(0)$ are treated as parameters.

3.4.1 Sensitivity of parameters to $C_{lab}(0)$

The key reason for greater parameter sensitivity due to $C_{lab}(0)$ (figure 3.5) is that the C_{lab} pool is more closely linked to the C_f pool than to any of the other C pools, and C_f influences GPP. Since the Metropolis algorithm operates by optimising the fit between the modelled and observed NEE, and since GPP has a strong influence on NEE (where NEE = R_{tot} – GPP, see figure 3.11), factors that control the size of the GPP pool will directly influence the modelled NEE values. Modelled LAI values are also required to fit the observed LAI, since LAI is linearly related to C_f (subsection 3.2.2).

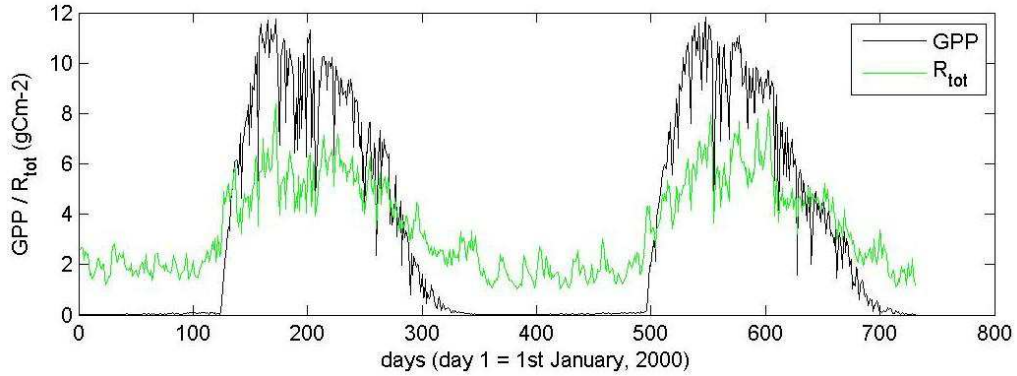


Figure 3.11 Trajectories of Gross Primary Production (GPP) and total respiration (R_{tot}) for 2000-01, using the true parameter set and true initial conditions of the DALEC-D model (table 3.1).

Two parameters, T_s and C_{fmax}, were completely insensitive to $C_{lab}(0)$. For T_s, the turnover rate of the C_{som} pool, this is not surprising because it was one the seven parameters from REFLEX whose estimates from the different DA schemes were very similar. It is also very small in magnitude, and so it is probably prevented from assuming large values otherwise too much C would get respired heterotrophically, which would cause a severe conflict between the modelled and observed NEE. For C_{fmax}, the maximum C_f value, this is probably insensitive to changes in $C_{lab}(0)$ because the LAI observations would be dictating the height of the peaks of the trajectory of the modelled C_f pool.

It is very noticeable that, of the eight parameters sensitive to $C_{lab}(0)$, seven were only sensitive for values either less than or greater than the true value of $C_{lab}(0)$. For example, in figure 3.5a, P_r is more or less insensitive to $C_{lab}(0)$ when this input is greater than its true value, but sensitive otherwise. In order to make sense of this, we plot the trajectories of total respiration, GPP, NEE and all the C pools (figures 3.12-3.14). For each of the nine plots, four trajectories are plotted when $C_{lab}(0)$ = 35 gCm^{-2}, 70 gCm^{-2} (the true value), 110 gCm^{-2}, and 170 gCm^{-2}, whilst keeping $C_r(0)$ and $C_{lit}(0)$ fixed at their true values, and the parameters fixed at their expected value using the emulator using the equation below (equation (3.7) from in subsection 3.2.10),

$$E[\hat{y}_i \mid x_1] = \int \int \int \hat{f}_i(x_1, x_2, x_3, x_4)\, p(x_2, x_3, x_4).dx_2 dx_3 dx_4,$$

but setting where x_1 (representing $C_{lab}(0)$) at 35, 70, 110 and 170 gCm^{-2} where $\hat{y}_i$ is the emulator's estimate of the *i*th parameters, and x_2, x_3, x_4 are the other inputs.

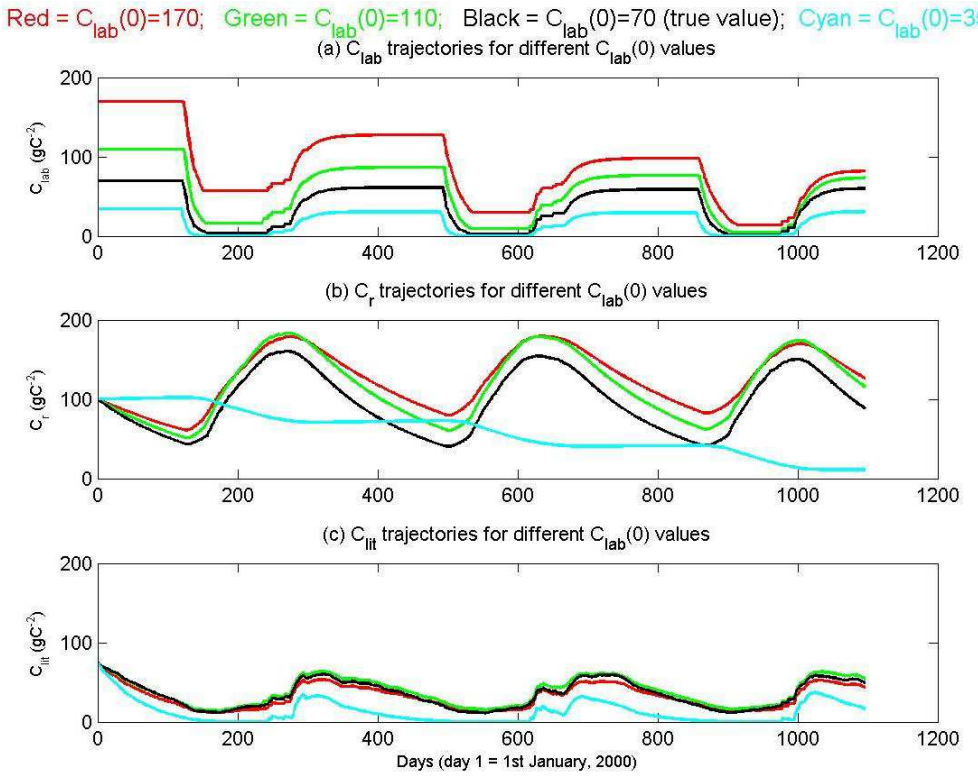


Figure 3.12 The trajectories of C_{lab}, C_r, and C_{lit} for different values of $C_{lab}(0)$, using the emulator's estimate of the parameters corresponding to the value of $C_{lab}(0)$.

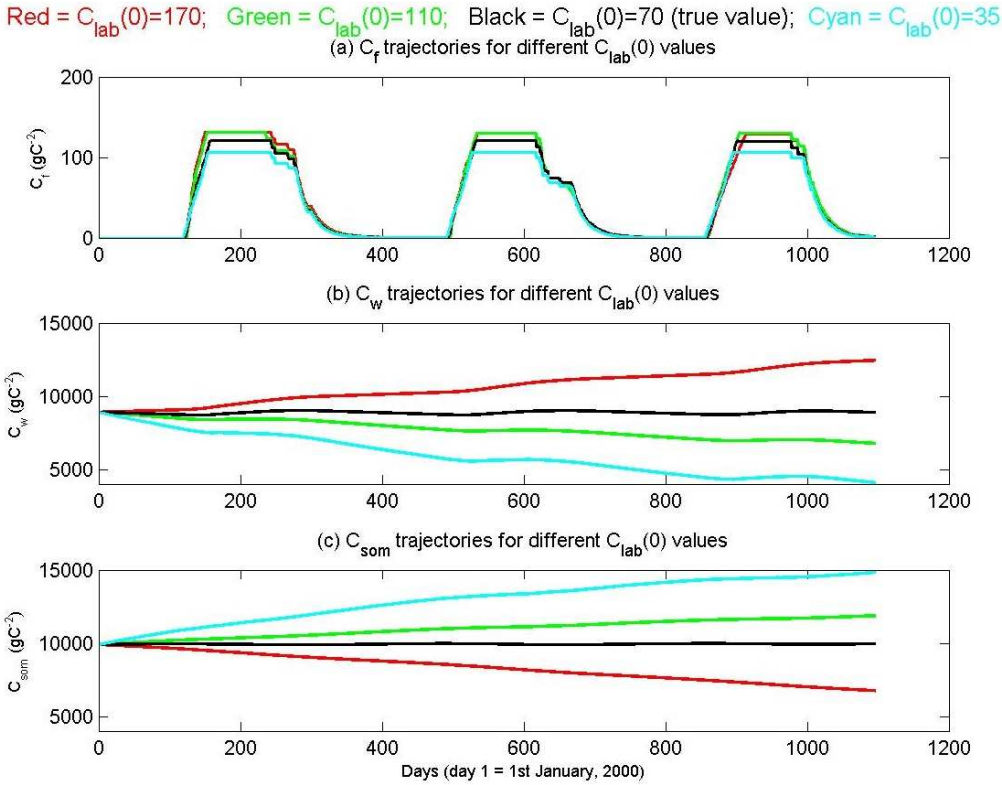


Figure 3.13 The trajectories of C_f, C_w, and C_{som} for different values of $C_{lab}(0)$, using the emulator's estimate of the parameters corresponding to the value of $C_{lab}(0)$.

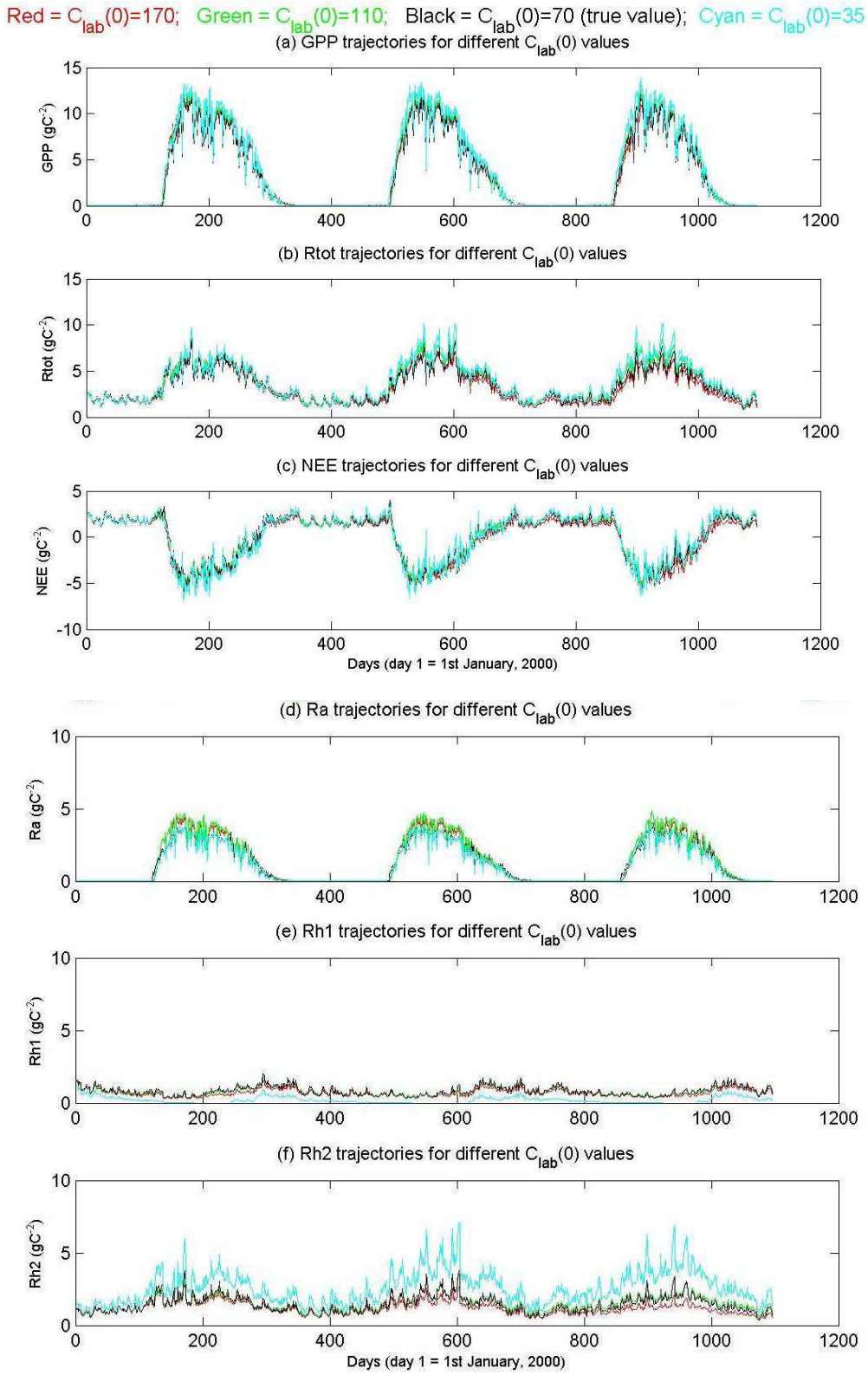


Figure 3.14 The trajectories of GPP, R_{tot}, NEE, R_a, R_{h1}, and R_{h2} for different values of $C_{lab}(0)$, using the emulator's estimate of the parameters corresponding to the value of $C_{lab}(0)$.

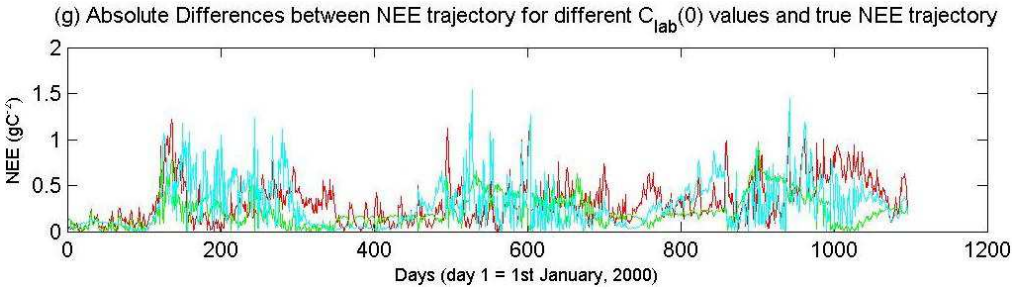


Figure 3.14 [cont.] The absolute difference between the NEE trajectories corresponding to $C_{lab}(0)$=35, 110 and 170 gCm^{-2} and the true NEE trajectory.

When $C_{lab}(0)$ = 110 or 170 gCm^{-2}, we can see that the C_{lit} trajectory (figure 3.12c) is closest to the true trajectory (i.e. when $C_{lab}(0)$ = 70 g C m^{-2}), whilst for the remaining C pools (figures 3.14a-b, 3.13a-c), the trajectories are closest only for $C_{lab}(0)$ = 110 gCm^{-2}. In particular, when $C_{lab}(0)$=35gCm^{-2}, the C_r trajectory (figure 3.12b) is not only furthest from the truth, but its trajectory is completely out of sync, in terms of the timings of the peaks and troughs, with the other trajectories. Despite the trajectories for GPP, R_{tot} and NEE (figure 3.14a-c), being close to the true trajectories whether $C_{lab}(0)$ is equal to 35, 110 or 170 gCm^{-2}, they are closer when $C_{lab}(0)$ = 110 gCm^{-2} – this can be seen for NEE if the absolute differences between each of the three trajectories and the true trajectory is plotted (figure 3.14g).

Therefore, although the parameters were not all consistently insensitive to changes in $C_{lab}(0)$ either less or greater than the true value of $C_{lab}(0)$, the trajectories of the C pools and key model states suggest that over-estimating $C_{lab}(0)$ by a small amount (i.e.110 gCm^{-2}) is better than under-estimating it by a similar amount (i.e. 35 gCm^{-2}) or over-estimating it by a larger amount (i.e. 170 gCm^{-2}).

A further interesting observation is that the labile pool in figures 3.10b and 3.12 appears to be decreasing year on year, both in the truth and the post-DA. There are two plausible explanations for this: (i) the pool stabilises at a pseudo steady state, or (ii) it reduces to zero. Unfortunately, it is not possible to answer this question directly by running the model forward because the driving data for the additional years is not easily available. Therefore we use Chuter (2013) to help determine the likely answer. Chuter (2013) showed very clearly that the foliar and labile C pools were tightly coupled. With this in mind and since the foliar C pool trajectories in figures 3.10b and 3.12 appear stable, it is reasonable to expect the labile C pool trajectory to also be stable and to settle on some pseudo steady state at some point in

the future.

3.4.2 Stability and transient issues

STABILITY

Although the previous subsection suggested that over-estimating the $C_{lab}(0)$ value by a small amount (i.e. 110 gCm^{-2}) is best, an important but overlooked issue is that any departure from the true set of parameters results in the results in a departure from stability. This is most clear in the trajectories of C_w and C_{som} in figure 3.13. When using the true parameters, we see that the system is in a pseudo equilibrium with the sizes of the C_w and C_{som} pools remaining virtually unchanged after a 3 year forward run of the model. Even when $C_{lab}(0)$ is increased from its true value by a small amount to 110gCm^{-2} (green lines of figure 3.13), we that after 3 years there has been roughly a 25% reduction in the size of the C_w pool. This means that either the woody part of the trees (which is predominantly the tree trunk and major branches) is reducing in size or there are fewer trees present. Both options are unlikely in the real world. Part of the reason this is happening is because the different values of $C_{lab}(0)$ correspond to different parameter sets. However this may also with the structure of the model. In particular, DALEC or DALEC-D does not incorporate tree allometry, that is relations between the height of the tree, width of the tree trunk, the amount of foliage, etc…. If it did, as the ACGCA model (Ogle & Pacala, 2009), then it is unlikely that we would have the close to same amount of foliage for the $C_{lab}(0)$=110gCm^{-2} setup (green line of figure 3.13a) as the true amount of foliage (black line) but then a significant difference in the amount of woody C.

As with the C_w trajectory, we also see that there is an increase in the C_{som} pool by around 20% under the $C_{lab}(0)$=110gCm^{-2} setup (green line of figure 3.13c) after 3 years, compared to the size of the pool at the start. Like the true C_w pool, the trajectory of the size of the true C_{som} pool remains relatively unchanged during the 3 year period. This change the C_{som} pool size by such a large amount in a relatively short amount of time is also unrealistic.

One thing that is noticeable is that under the $C_{lab}(0)$=110 gCm^{-2} setup (which is also true of the other setups), the amount of C lost by the C_w pool over the 3 year period is approximately the same as the amount gained by the C_{som} pool. Therefore, no net C is really lost or gained by the ecosystem it has just been moved from the

woody C store to the belowground soil C store. This is of course unrealistic, and so is really an issue of model structure. However this could also be resolved by assimilating data of one of these large C pools. Without this extra constraint, the modelled R_{h2} flux of respiration is able to mimic that of the true respiration (panel f of figure 3.14), as long as the sum of the C stored in the C_w and C_{som} stored is approximately the same as the true value by the end of the 3 year period. Improvements in the model structure can obviously be made so this unrealistic feature of the model does not happen, yet when estimating parameters (and uncertainties) in a DA scheme like this, assimilating data of one of the large C pools (as stated above) will likely enable the parameters controlling the amount of C going in and out of the C_w and C_{som} pools to be closer to their true values.

A final point to make about stability in relation to the comment from the previous subsection about the cyan line (corresponding to the $C_{lab}(0)=35gCm^{-2}$) being out of sync with the other C_r trajectories in figure 3.12b. An additional point to make is that not only is this trajectory out of sync, but it is also decreasing, so much so that after 3 years the size of root C pool is close to 0 gCm^{-2}. In reality this would eventually result in the trees dying as they would have no roots. This would undoubtably therefore result in a loss of woody carbon, as we see by the cyan line in figure 3.13b though one would not expect the transfer of woody C to shift to the C_{som} pool quite so dramatically as we do; though it worth realising the C_{som} pool also includes woody debris so as a number of the trees may not be alive it seems appropriate that the woody remains would thus be part of the C_{som} pool. Despite this, we can see that the resulting GPP and respiration trajectories from this $C_{lab}(0)=35gCm^{-2}$ setup (cyan lines of panels a and b from figure 3.14) are similar to the true trajectories. The point of this is that we that the calibrated parameters using this $C_{lab}(0)$ value of $35gCm^{-2}$ are still able to result in a good fit of the modelled to observed NEE despite this evidence that the ecosystem is in severe decline under this setup. This highlights the point that although poorly chosen initial conditions can result in a good fit of the observations to the model, the poorly estimated parameters can result in many part of the model to behave unrealistically.

TRANSIENT ISSUES

In figure 3.12(a), we can see that although the labile C pool is initialised from different

values, after 3 years using the corresponding sets of parameters, the C_{lab} trajectories are very similar to each other, particularly the black, green and red ones. Therefore, it could be suggested the different $C_{lab}(0)$ values make no difference really. However, there are important counter-arguments to consider. First of all the parameters in the optimisation were estimated using the NEE and LAI data from the two years whereas the plots in figures 3.12 – 3.14 cover an extra year (from days 732 to 1096 in the plots). At day 731, we can see that this is still a reasonable difference between the trajectories of C_{lab} under the different $C_{lab}(0)$ setups. Therefore if a short amount of data is being assimilated then knowing the initial conditions of the model are important. More importantly however, by under-estimating $C_{lab}(0)$ by a small amount ($35gCm^{-2}$) or over-estimating it by a large amount (by 100 gCm^{-2}) can result in severe deviation of the size of the C_w and C_{som} pools from the true sizes (red and cyan lines of panels b and c of figure 3.13). This is turn will result in posterior estimates of the parameters that could be far from the truths and whose uncertainty intervals are not containing the truth.

If we were using a longer time-series of data in the optimisation and using more types of data, then this transient issue would not be a problem. This is because the parameters would more likely be estimated correctly, with the increased amount of data able to constrain more of the model processes thus pin down the parameters that control these processes more precisely. Having a longer time-series of data would then mean that there would be more time to move away from the differing initial conditions.

WHAT DO THESE STABILITY AND TRANSIENT ISSUES MEAN FOR THIS ANALYSIS?

Under the setup we have, i.e. a short time series (2 years) of NEE and limited LAI data, despite slightly over-estimating $C_{lab}(0)$ by a small amount (i.e. by 40 gCm^{-2}) resulting in stability issue with the sizes of the C_w and C_{som} pools decreasing and increasing (respectively), it is least unstable in this respect compared to the other two $C_{lab}(0)$ setups (i.e. setting $C_{lab}(0)$ to $35gCm^{-2}$ or $170gCm^{-2}$). Therefore if $C_{lab}(0)$ is wrongly estimated, it is surely better to wrongly estimate by such an amount that there is the least deviation from the truth. In this case, over-estimating $C_{lab}(0)$ by $40gCm^{-2}$ is the best option. Furthermore, as stated in the previous discussions, if $C_{lab}(0)$ is

under-estimated by 35gCm^{-2} then in this situation it results in the severe instability in the plant ecosystem.

3.4.3 Dealing with lack of knowledge of $C_{lab}(0)$

The key observation about the results was that for nine of the ten parameters, the estimates were close to the truth when $C_{lab}(0)$ took its true value. However, accurately estimating $C_{lab}(0)$ can be very difficult to estimate accurately as it involves sampling tree tissue and determining the soluble starch and sugars in the laboratory (Gough, 2009), a process which is complex and time consuming. As a result, measurements of labile carbon are rarely available. In this subsection we assess whether incorporating other data and knowledge into the Metropolis algorithm may help to ensure that the parameter estimates, particularly F_{ll}, are close to the truth even if the $C_{lab}(0)$ values are not. This could include:

(1) Knowing the true values of $C_r(0)$ and $C_{lit}(0)$
(2) Increasing the number of LAI observations.
(3) Assimilating other types of observations, such as C_r and C_{lit}.
(4) Assimilating a longer time series of NEE and/or LAI.

In this chapter we only considered (1) and (2), while the other two are not considered for reasons that follow. Both $C_r(0)$ and $C_{lit}(0)$ can be measured more easily by direct means, but the measuring process can be time consuming, thus measurements are typically only made one or two times a year (Williams et al, 2005, Richardson et al., 2010). This means that we should be able to retrieve good estimates of $C_r(0)$ and $C_{lit}(0)$ if we began our assimilation when the measurements were made. However due to the lack of frequent measurements, assimilating C_r and C_{lit} data is unlikely to make a difference to the parameter estimates particularly over a 2 year period. Assimilation of longer time series of observations is discussed in chapter 4.

Increasing the number of LAI observations, from 10 to 100, is possible if synthetic observations from space are included as well as those based at the ground level. On the MODIS sensor on NASA's TERRA satellite, for example, LAI observations are made every 8 days over the entire globe which corresponds to 91 measurements over a 2 year period. Including more LAI observations into the DA scheme would result in better representation of the model processes surrounding the C_f pool (since LAI and C_f are linearly related) and consequently the processes

surrounding the C_{lab} pool due to the strong connection it has with C_f (figure 3E from appendix). To test whether knowing the true values of $C_r(0)$ and $C_{lit}(0)$ and increasing the number of LAI observations compensates for a poor estimate of $C_{lab}(0)$, the sensitivity analysis was repeated with these two changes included. A description of the how this was carried and what the results were, is described next.

ADDITIONAL COMPUTER EXPERIMENTS

Two computer experiments were carried out to test whether knowing $C_r(0)$ and $C_{lit}(0)$ and assimilating more LAI observations would compensate for not having an accurate estimate of $C_{lab}(0)$. For experiment 1, ten extra runs of the Metropolis algorithm were carried out, varying the $C_{lab}(0)$ value using a stratified random sample (which is the same as a Latin Hypercube design in one dimension), while fixing the other two initial conditions and the LAI observational error at their true values. Experiment 2 was exactly the same as experiment 1 except that the number of LAI observations was increased from 10 to 100, with the extra 90 observations randomly placed in those days which did not have one present. The value 90 was based on the number of LAI observations available if measured from space and from the ground.

For all of the ten parameters except P_r, the sensitivity plots from experiment 1 were almost the same when the emulator was used; for example see figure 3.8(a)-(b),(g)-(h) for F_{nf} and T_{lab} (subsection 3.3.4). For P_r, the sensitivity plot was almost the same except that the lower bound of the range of $C_{lab}(0)$ values for which the parameter estimate was close to the true value decreased from 70gC/m² (figure 3.8(d)) to around 50gC/m² (figure 3.8(e)). For seven of the ten parameters, their estimates remained close to the truth when $C_{lab}(0)$ was close to its true value. This means that knowing the true values of $C_r(0)$ and $C_{lit}(0)$ does not compensate for poorly estimating $C_{lab}(0)$.

For experiment 2, five of the ten parameters (F_g, F_{nf}, T_s, L_{out} and C_{fmax}) had sensitivity plots approximately the same as from experiment 1 (e.g. F_{nf}, figure 3.8(c)). For three of the parameters (T_d, F_{nrr} and P_r), the parameter estimates were further from the truth than in experiment 1 (e.g. P_r, figure 3.8(f)). However the most interesting finding was the T_{lab} sensitivity plot. In the original results, this parameter was the only one not to be close to the truth when the true $C_{lab}(0)$ value was used. In experiment 2 we found that T_{lab} was now at its true value when $C_{lab}(0)$ was at its true

value (figure 3.8(g)-(i)). This is probably because the extra LAI observations have helped the modelled C_f trajectory to be a better fit to the true C_f trajectory. Due to the strong connection between the C_f and C_{lab} pools (figure B7, appendix B), the extra LAI observations would have also encouraged a better fit of the modelled C_{lab} trajectory to the true C_{lab} trajectory when the true $C_{lab}(0)$ value was used. This results in the right amount of C leaving (and entering) this pool, implying a more accurate estimate of T_{lab}.

In summary, knowing $C_{lit}(0)$ and $C_r(0)$ made very little difference to the sensitivity of the parameter estimates to $C_{lab}(0)$. For most parameters, knowing these initial conditions and assimilating 90 extra LAI observations) did also not change the results. Therefore even if the initial conditions of most of the pools are known and even if two types of observations are assimilated, $C_{lab}(0)$ still needs to be right to correctly estimate the parameters. Since estimating $C_{lab}(0)$ correctly is unlikely, as explained in subsection 3.4.1, it is important to assess the consequences of when our estimate of $C_{lab}(0)$ is far from its true value. We explore this next.

3.4.4 The consequences of poorly estimating $C_{lab}(0)$

For some parameters it does not matter if they are poorly estimated, as a result of $C_{lab}(0)$ not being correct, since they may not have much influence on the model estimate of NEE. Therefore it is more important to estimate well the parameters which have the largest influence. In chapter four, we determined the sensitivity of the fit of the modelled to observed NEE over a period of 1 year for site in Harvard forest. Of the nine parameters which control the fast processes of DALEC-D, we found that the fraction of GPP respired autotrophically (F_g), the fraction of NPP allocated to foliage (F_{nf}), the turnover of foliar carbon (T_f), the fraction of foliar carbon lost to litter (F_{ll}) and the fraction of labile carbon respired (F_{lr}) had the largest influence on the NEE fit. Three of these, namely F_g F_{nf} and F_{ll}, were among the ten parameters which passed the diagnostic checks. Therefore we assessed whether the sensitivity plots associated with these parameters had any common features, in particular if there was a common range of $C_{lab}(0)$ values where the parameter estimates are closest to the truth. The $C_{lab}(0)$ sensitivity plots are displayed again for these three parameters in figure 3.15(a)-(c). We can see that for F_g and F_{nf} values of $C_{lab}(0)$ between 0 and 70 gCm^{-2} result in the parameter estimates being closest to the truth (i.e. the horizontal

red dashed line), but for F_{ll} this is when $C_{lab}(0)$ is greater 70 gCm^{-2}. Therefore, whatever value of $C_{lab}(0)$ is used we will get at least one of these three important parameters wrong. As an addendum, it is worth including here that by saying we get a parameter wrong, we mean that the posterior modal value will be far from the truth value in the parameter space, but the uncertainty interval may still include the truth.

As the estimated parameters and the associated uncertainty distributions are normally used in the model (e.g. in chapter four, 200 randomly parameter sets from the posterior distribution are used to estimate a distribution of NEE forecasts), it is perhaps more important to assess the impact of wrongly estimating $C_{lab}(0)$ on the time series of the C pools and other important quantities, such as NEE. In figures 3.10 – 3.12, we can see that higher values of $C_{lab}(0)$, i.e. values of 110gCm^{-2} and 170gCm^{-2}, result in the C_{lab} and C_r trajectories being further from the true trajectories. However, these larger values of $C_{lab}(0)$ result in the trajectories of the C_f pool (the most important of the small C pools due to its influence on GPP) and, C_w and C_{som}, the largest C pools where the long term C is stored. Larger values of $C_{lab}(0)$ also result in the GPP, R_{tot} and NEE trajectories to be closer to the true trajectories. For C_f, GPP, R_{tot} and NEE, this improvement is predominantly around Spring of both of the years. However, setting $C_{lab}(0)$ to 35 gCm^{-2} results in C_w being around 6600 gCm^{-2} at the end of the 2 year period, which is much further away from the true value of 8900 gCm^{-2} than 8600 gCm^{-2}, which is the end value of C_w when $C_{lab}(0)$ = 170 gCm^{-2}. We can see from figure 3.11(c) that the value of C_{som} on the final day of the 2 years period is much closer to the true value if $C_{lab}(0)$ = 170 gCm^{-2} than if $C_{lab}(0)$ = 35 gCm^{-2}.

3.4.5 Sensitivity of parameters to $C_{lit}(0)$, $C_r(0)$ and LAI observational error

The results showed that more parameters were sensitive to $C_{lit}(0)$ (five) than $C_r(0)$ (two) (figure 3.7 and figures B5 and B6 from Appendix B) However, for three of the five $C_{lit}(0)$ parameters (namely P_r, T_s, and L_{fll}) the sensitivity was very weak (i.e. in figure 3.7(a),(b) and figure B5(b), the blue lines are almost flat). One of the remaining two, F_{nrr}, is strongly sensitive to $C_r(0)$ as well as to $C_{lit}(0)$. This is of no surprise given that F_{nrr} controls the flow of C leaving the C_r and entering the C_{lit} pools.

Except for the multiple complex MCMC scheme, all the Metropolis based alg-

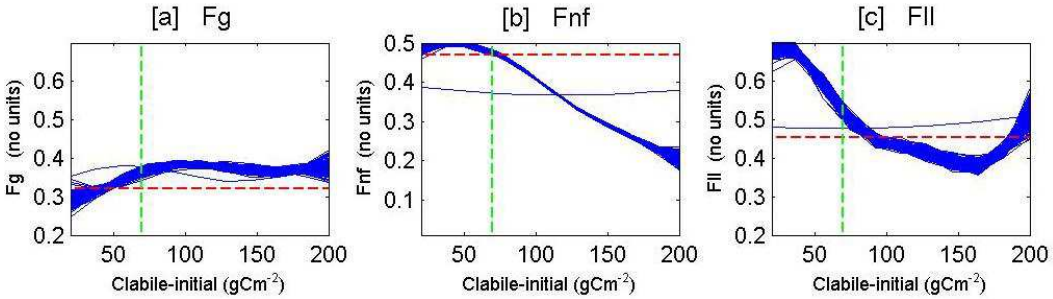


Figure 3.15 Panels (a)-(c) show the sensitivity of three of the most important parameters in DALEC-D to $C_{lab}(0)$.

orithms in REFLEX used an LAI observational error of 0.05 or 0.11 with this other scheme using a value of 0.0005 (see Fox et al., 2009). The 0.05/0.11 values were mostly used because the REFLEX participants were told that the variance of LAI errors should be approximated as 10% to 15% of the mean of the LAI observations. Four of the Metropolis based DA schemes used 10% while one used 15%, and these correspond to a variance of 0.05 and 0.11 respectively. The multiple complex MCMC scheme, on the other hand, used a variance of 1% of the mean of the LAI data which equalled 0.0005. The reasoning for such a low observational error for this scheme was to give the 10 LAI observations much greater weight in the cost function since there were significantly more (231) NEE observations. While there are some literature to support this idea (e.g. Vrugt et al., 2003), most Metropolis based DA schemes do not (e.g. Knorr & Kattge, 2005; Hill & Williams, 2007; Ziehn et al., 2011).

The results of this chapter do not appear to favour either option, with half of the six sensitive parameters (figures 3.7a-c) closest to the truth for larger LAI observational errors, and the other half (figures 3.7d-f) closest to the truth for smaller values of the observational error. However, when the trajectories of the C pools and key model states are plotted (figures B8-B10 from appendix B), we find that they are consistently close to the true trajectory when we use the parameters estimated from the Metropolis algorithm using an LAI observational error of 0.05gCm^{-2}.

3.4.6 Implications for REFLEX

Of the ten parameters whose corresponding emulators passed the diagnostic checks, nine were among the twelve parameters whose estimates varied among the different DA schemes in REFLEX. The results of this chapter suggest that a possible reason

could be due to poor estimation of the initial values of the labile carbon pool. However, the $C_{lab}(0)$ values are not available for all but one of the Metropolis based schemes, because $C_r(0)$, $C_{lit}(0)$ and $C_{lab}(0)$ were treated as parameters and were not outputted as part of the individual results. Given that there were differences in the estimates of 12 of the 17 parameters, it is unlikely that the estimates of $C_{lab}(0)$ were the same.

An interesting observation about the REFLEX results was that one of the Metropolis based schemes (the Multiple Complex MCMC method) produced noticeably different estimates to the remaining schemes for seven of the parameters, four of which (F_{nf}, L_{out}, F_{ll}, C_{fmax}) were a subset of the ten parameters, whose emulators passed the diagnostic checks. This method also had a significantly different prescribed variance for the observational error distribution for the LAI part of the cost function ($0.0005 gCm^{-2}$) compared to the other Metropolis based schemes ($0.04 – 0.11 gCm^{-2}$). Two of the six parameters whose estimates were sensitive to the LAI observational error (namely F_{nf} and F_{ll}) were among the above four parameters, however the other two were insensitive to changes in this input. Therefore, the results are inconclusive about whether the lower prescribed LAI observational error is causing the observed differences between the parameter estimates of the Multiple Complex MCMC method and the other Metropolis based schemes.

Finally, it is interesting to note that $C_{f,max}$ was the only parameter not sensitive to $C_r(0)$, $C_{lit}(0)$, $C_{lab}(0)$ or the LAI observational error, yet two of Metropolis based DA schemes from REFLEX produced very different estimates of this parameter compared to the other schemes. Therefore there must be other reasons causing the differences.

3.4.7 Treating the initial conditions as parameters

The results of the optimisation where the initial conditions of the C_r, C_{lab} and C_{lit} pools are treated as parameter are given in subsection 3.3.5. As expected, the NEE and LAI trajectories from propagating the model forward using 500 randomly chosen parameter sets from the posterior distribution resulted in a good fit to the data and the true NEE and LAI. For each of the other C pools, the trajectories tended to be close to the true trajectory. This is interesting because two of the three initial conditions which were treated as parameters, had posterior estimates that were far from the truth and

whose posterior intervals did not contain the truth. This suggests that it does not matter if these three initial conditions were wrongly estimated; in particular the $C_{lab}(0)$ parameter has a posterior mean value of around $150gCm^{-2}$, which is significantly higher than the true value, yet this has very little negative impact on other aspects of the model, such as the trajectories of the slow C pools which deviated significantly from the true trajectory when $C_{lab}(0)$ was treated as fixed and was estimated to be around $170gC^{-2}$ (figure 3.13).

3.4.8 Implications of emulation of a DA algorithm

A unique aspect of the work in this chapter is the emulation of a DA scheme, something which has not been done before. This is very noteworthy because it means that new insights can be gained into the inner workings of a DA scheme. In this chapter it was used to see how sensitive parameter estimates are to changes in the initial conditions of the model. It could be used to assess the sensitivity of parameter estimates or the uncertainty on those estimates to changes in other factors such as features of meteorological data which are generally used to drive for terrestrial ecosystem models. DA emulation could also be used for uncertainty analysis which is the quantification of uncertainty in an input (e.g. an initial condition or a prior estimate of a parameter) propagated forward to an output (e.g. a posterior parameter estimate). Both these sensitivity analysis and uncertainty analysis tools could help the implementation of DA algorithms, for example to help determine suitable uncertainty bounds on inputs or priors. Moreover, recent progress on multivariate emulators, such as Fricker et al. (2010) and Hankin (2012), will allow emulator outputs to be represented by multiple values rather scalars. With the exception of certain DA algorithms, such as variational schemes (e.g. 4D-VAR), most DA algorithms are stochastic in nature although behave approximately deterministic (in other words, repeating two run of the Metropolis algorithm will result in approximately the same posterior estimates of the parameters and uncertainties). Nonetheless, this extra source of uncertainty in needs to be included in the emulation process. Furthermore given that that a run of a DA scheme can be computationally expensive, with the exception of some schemes such as the Ensemble Kalman filter, DA emulation may in some circumstances be limited in application depending on the size of the model.

3.5 Conclusions

The objectives of this chapter were to determine the sensitivity of the DALEC-D model parameters to: (1) the initial conditions of the small C pools and, (2) the LAI weighting term of the cost function, when using DA to estimate the 17 parameters of the DALEC-D model. A major novelty of this chapter was the use of emulators to carry out the sensitivity analysis.

It was found that more of the parameters were sensitive to $C_{lab}(0)$ than to any of the other three inputs, and the parameter sensitivity was much stronger for $C_{lab}(0)$ than for the other inputs. Furthermore, the majority of the parameter estimates were close to their true values when the true $C_{lab}(0)$ value was used. However, obtaining accurate estimates of $C_{lab}(0)$ is very time consuming and so are rarely available. We tested whether this lack of information could be compensated by know the initial conditions of the litter and fine roots pools and assimilating more LAI observations. For most parameters incorporating this extra knowledge into the assimilation scheme did not help with the graphical representation of the sensitivity analysis being very similar to the original results.

As it was very difficult to accurately estimate $C_{lab}(0)$, we assessed the impact of wrongly estimating this initial condition on the trajectories of the C pools and key model states, using the parameter estimates obtained from the wrongly specified $C_{lab}(0)$ values. We found that over-estimating $C_{lab}(0)$ by a small amount (i.e. $C_{lab}(0)$ = 110 gCm^{-2}) compared to the true value (i.e. $C_{lab}(0)$ = 70 gCm^{-2}) resulted in the trajectories being closer to the true trajectories, than if $C_{lab}(0)$ had been under-estimated by a similar amount (i.e. $C_{lab}(0)$ = 35 gCm^{-2}) or over-estimated by a considerable amount (i.e. $C_{lab}(0)$ = 170 gCm^{-2}). Thus, even if the parameter estimates require the $C_{lab}(0)$ value to be accurately estimated, there is scope for $C_{lab}(0)$ being over-estimated by a reasonable amount whilst still being able to estimate the C pools and important model states such as NEE to a reasonable level of accuracy.

In the wider context of this thesis, this chapter has shown that the initial condition of the labile carbon pool is a factor which causes variations in the model parameters estimated from the Metropolis algorithm. We have also shown that parameter estimates are not consistently robust against a poor estimate of $C_{lab}(0)$, even if we assimilate more data and know $C_r(0)$ and $C_{lit}(0)$, i.e. we need to get $C_{lab}(0)$ right in order to get the parameters correct. Assimilating other measurements, particularly

ones which are orthogonal to existing ones (Richardson et al., 2010), may compensate for not knowing $C_{lab}(0)$, and this should be investigated using synthetic data. It was also shown that treating the initial conditions as parameters is preferable. Even though two of the three initial conditions had posterior estimates far from the truth, most of the other parameters were estimated well and the trajectories of C pools were in general reasonably consistent with the truth. In chapter 4, we explore other factors which cause parameter estimates to vary, which include the length of the dataset being assimilated, the size of the observational error and the level of size and frequency of gaps in the dataset; however we treat the initial conditions as parameters.

In this chapter we have investigated whether parameter estimates are sensitive to changes to various factors. However it would also be of great interest to examine the sensitivity of the error bars (or more specifically, the widths of the marginal posterior p.d.f.s) of the parameter estimates as this also varied in REFLEX for the different DA schemes amongst the parameters. Another useful extension to this work would be to use emulators which can deal with multi-dimensional outputs (i.e. use one emulator where the output is the 17 dimensional vector of parameter estimates, rather than using 17 scalar output emulators) once the theory behind these have been sufficiently developed – see Rougier (2008), Fricker et al. (2010), Conti & O'Hagan (2010), Hankin (2012).

The main limitations of this chapter are the use of synthetic observations and the use of emulators. Although the observational error and the size and frequency of gaps in the synthetic dataset is an accurate representation of reality, the model processes used to generate the 'true data' may not accurately represent the real world processes well enough. Furthermore, synthetic data does include a representation of systematic error. This is usually ignored because such errors can usually be identified resulting the affected parts of the dataset being removed. However, it is possible that it is not identified. A simple remedy to this would be to assign a probability for each time-step of systematic error not being picked up, and then adding on either a predetermined or random amount to the affected part of the dataset.

Although highlighted as a major novelty to this chapter, the use of emulators is also a further limitation, because the emulator is usually built to model simulators. However, in this chapter it is being used to represent the Metropolis algorithm which

is not deterministic (though is approximately so). This may have resulted in slight inaccuracies in the results.

Chapter 4

Sensitivity of estimates of model parameters and NEE predictions using Data Assimilation to changes in the characteristics of an NEE dataset.

Abstract

It is becoming increasingly common for Data Assimilation (DA) to be used to estimate parameters and states of terrestrial ecosystem models, using land-based measurements. Since the characteristics of a dataset vary for different studies, this chapter investigates how the parameter estimates and corresponding forecasts by assimilating datasets of different lengths, data density and observational error. Two DA methods were compared, namely the Ensemble Kalman Filter (EnKF) and the Metropolis algorithm. The model used was a simple carbon box model called DALEC-D (Data Assimilation Linked Ecosystem model for Deciduous vegetation), which models the flow of carbon from photosynthesis through to respiration; and the observations assimilated were the net vertical flux of CO_2, referred to as NEE or Net Ecosystem Exchange, which is measured using eddy-covariance methods. The

assimilation schemes were used to estimate the DALEC-D model parameters and Carbon (C) stocks on the final day of the dataset. These estimates were then used to forecast cumulative NEE 10 years into the future.

The results suggested that at least 5 years of NEE observations are required to obtain good estimates of the parameters, final C stocks and cumulative NEE forecasts for both DA schemes, with the estimates improving with an increase in dataset length and the uncertainty intervals becoming narrower whilst still containing the truth. For the Metropolis algorithm, the runs involving datasets longer than 5 years were excluded from the final analysis for reason given below. The level of observational error and density of observations had far less impact on the parameter and C stock estimates for the EnKF, but the uncertainty in the estimates did reduce which is encouraging. There is little or no evidence that the slowly changing parameter and final C stocks are better constrained by assimilating the longer time-series datasets. For the Metropolis algorithm, the corresponding results were excluded from the final analysis for reasons described next.

These unexpectedly good forecasts for the Metropolis algorithm, given the poor parameter estimates, is most likely because that DALEC-D is susceptible to a phenomenon called “equifinality”, which in this context is when a parameter set which is not similar to the true parameter set can produce an NEE trajectory close to the true NEE trajectory. Equifinality occurs because the modelled NEE is likely to be insensitive to changes in a number of “redundant” parameters, with only a certain number having significant influence. To test this, an emulator called GEM-SA (Gaussian Emulation Machine for Sensitivity Analysis) was used to see how sensitive the root mean squared error (RMSE) of the trajectory of modelled NEE values against the true NEE values over 1 year NEE over one year is to changes in the nine parameters that control the rapidly changing C pools. It was found that the cumulative NEE was very insensitive to changes in 3 of these parameters (the turnover rates of fine root carbon, litter carbon and labile carbon), and very strongly sensitive to two parameters (the fraction of GPP respired autotrophically and the turnover rate of foliar carbon). The Metropolis algorithm is particularly affected by equifinality because the algorithm operates by searching for parameter sets that give a good fit to the observations. As a result the poor parameter estimates resulted in NEE trajectories, during the DA period, that were still close to the truth. The effects of equifinality can

be limited by assimilating multiple types of data (Richardson, 2010), though altering the structure of the model could also be beneficial.

However the poor parameter estimates from the Metropolis algorithm did not occur for all of the runs. For the ones which did see poor results, the residuals in the cost function used in the Metropolis algorithm were autocorrelated, whereas for the other runs where there was good agreement with the EnKF the residuals exhibited no autocorrelation. A critical finding described in the discussion found that for the poor runs, the true parameter set had a significantly lower cost function value than those corresponding to the posterior parameter space. This suggested that the posterior parameter distribution was not at the global minimum and for this reason among others these poor runs were excluded from the final analysis. The runs from the Metropolis algorithm which remained were mainly ones where the dataset length varied. Despite the exclusion of a number of the Metropolis algorithm runs, this autocorrelation of the residuals was an interesting finding and serves as a warning to other studies where this DA scheme is used with synthetic data.

4.1 Introduction

4.1.1 Background and Motivation

A key uncertainty in the global carbon budget is due to poor estimates of the amount of carbon naturally taken up by terrestrial ecosystems from the atmosphere each year (Denman et al., 2007). Until recently, this was indirectly estimated as the residual when all the other elements of the carbon budget had been accounted for (Denman et al., 2007). Efforts are now being made to measure this atmosphere-land flux of carbon, which can be either land based or satellite based observations. Alternatively, forward runs of land-surface ecosystem models can be conducted and compared with observations. Data assimilation is a way of absorbing observations into a model, rather than making the comparison afterwards. A key advantage is that the uncertainty of observations and the model can be incorporated into the procedure. DA can be used to estimate the states of a model (e.g. how much carbon is stored in foliage), over multiple time-steps (chapter 2). It can also be used to estimate parameters of the model. Two important questions which need to be asked are: (i)

What are the strengths and weaknesses of DA? (ii) Under what conditions does DA perform well and not so well?

The REFLEX study (Fox, 2009) showed that the choice of DA algorithm is important. The study found that 8 different DA schemes, using the same model and the same datasets, resulted in significantly different parameter estimates. One possible explanation is that the initial conditions of the small carbon (C) pools (i.e. C stored in foliage, fine roots, litter and labile C) were estimated separately for each scheme. Chapter 3 showed that increased knowledge of these initial conditions improved the parameter estimates. In this chapter we investigate other factors which might affect a DA algorithm's ability to estimate parameters.

An important study by Tang and Zhuang (2008), which used a Bayesian inference DA scheme, highlighted that assessing a DA scheme's performance based solely on its estimates of the parameters may be misleading. They found that equally good fits to the observations can be achieved using different sets of parameters, a phenomenon they termed "parameter equifinality"; this occurs because the observations fail to constrain many parameters. They also found that when site level parameterisation is extrapolated to estimate the carbon dynamics on a regional scale, parameter equifinality can cause large uncertainty in these regional estimates. That study nonetheless seems flawed since parameters estimated at the site-level are based on site-level observations and meteorological data, and it seems inappropriate to use these parameters to estimate the carbon dynamics over a much larger spatial scale, many parts of which will have different meteorological data and vegetation types. However in fairness to the authors, they do state in their conclusion that a more sensible approach to estimating regional carbon dynamics would be to aggregate the estimates obtained from parameterizations from all sites within the region. This, though, is more computationally expensive. A further criticism of Tang and Zhuang (2008) is that they do not show the posterior distributions of the site-level parameters. If they had done, an interesting discussion point would have been whether if greater equifinality results in greater uncertainty of the regional carbon dynamics.

Other work has also considered factors that might affect the performance of DA to obtaining unbiased estimates of parameters. Carvalhais et al. (2008) found that when steady state is assumed, the parameter estimates may be biased, and estimates improved when this assumption is relaxed. Specifying different cost functions (to

describe the model-data mismatch) whilst using the same model and the same dataset can also result in different parameter estimates, as was found by Trudinger et al. (2007). More specifically, if the cost function is specified by the equation below (where $x_1(t_i)$ and $x_2(t_i)$ are the observations at time t_i with $z_1(t_i)$ and $z_2(t_i)$ being the corresponding modelled values), they found that if the weights, w_1 and w_2, are not constant (i.e. functions of time) and vary with noisy observations, this leads to biased parameter estimates.

$$\Theta = \sum_i \left(\frac{(x_1(t_i) - z_1(t_i))^2}{w_1} + \frac{(x_2(t_i) - z_2(t_i))^2}{w_2} \right) \quad (4.1)$$

Another interesting finding in the study by Trudinger et al. (2007) was that all 9 DA schemes used in that study gave biased estimates when the observational noise was treated as temporally correlated or non-Gaussian. However, only one non-Gaussian noise type was used (the lognormal distribution). It would have been interesting if other non-Gaussian noise types had been included, e.g. the double exponential distribution as used in Hollinger & Richardson, 2005; Richardson et al., 2006.

However, there are other possibly influential factors which have not yet been considered. The first is the length of the dataset. Different DA studies, using terrestrial ecosystem models, assimilated different numbers of years of observations; for example, the REFLEX study used 2 years (Fox et al., 2009), Quaife (2007) used 3 years, Mo et al. (2008) used 8 years, and Braswell (2005) used 10 years. Since more information is assimilated, longer datasets might be expected to lead to more accurate parameter estimates. Thus, one of the questions we set out to answer in this chapter is whether longer datasets help to constrain more of the parameters. Other aspects of the observations might also be important, such as the observational uncertainty and the size and frequency of data gaps, both of which are features of real datasets. In all of the studies mentioned above, these were kept fixed. Therefore in this chapter, we also investigated whether smaller observational uncertainty and greater density of observations improved the parameter estimates.

4.1.2 Hypotheses

This chapter tests five hypotheses:

(H1) Increasing the length of the NEE dataset improves parameter and carbon stock estimates;

(H2) Slowly varying C pools and their controlling parameters are only constrained by longer datasets;

(H3) Increases in the density of observations (or decreases in the frequency and size of gaps) will result in more accurate estimates of the parameters and carbon stocks;

(H4) Increases in observational error will result in less accurate estimates of the parameters and carbon stocks;

(H5) Datasets of greater length, with greater data density and with smaller observational uncertainty will result in more accurate forecasts.

H1, H3 and H4 arise from the idea that the greater the amount or quality of information contained in the dataset, the better the estimates of the parameters. H2 arises from the fact that some of the parameters in the model correspond to fast processes (e.g. the fraction of Net Primary Production allocated to foliage), which have an approximately periodic annual behaviour, while others relate to slow processes, e.g. the turnover rate of C from woody biomass (e.g. tree trunks). Therefore, while short datasets may be long enough to constrain the parameters related to the fast-changing pools, much longer ones may be required for the parameters of the slow-changing pools.

4.1.3 Other considerations

In testing these hypotheses, it is clear that the DA scheme employed is important, as different algorithms may give different results. In studies which have compared different DA schemes (Fox et al., 2009, Trudinger et al., 2007), a large variety of different methods were used. Instead, we here consider two broad groups of DA schemes. The first includes recursive filters and smoothers where the outputs at one time-step are used as the inputs of the next (e.g. the Ensemble Kalman Filter or EnKF); the second includes batch methods which unlike recursive filters and smoothers involve repeated forward runs of the model, e.g. the Metropolis algorithm and the Genetic algorithm. There is also a third group, called variational methods (e.g. 4D-VAR), but these are not considered here. We picked an algorithm from each of the first two groups, namely the EnKF and the Metropolis Algorithm.

To carry out the experiments, we used the DALEC-D ecosystem model (Fox et al., 2009, Williams et al. 2005). The datasets were based on ground-based measurements, for which there was good understanding of how to quantify the uncertainty of the observations (Richardson et al., 2007); this is important since one of the factors being varied in the experiment is observational uncertainty. Several types of observations could be used, but for this experiment NEE (Net Ecosystem Exchange) was chosen, as it is frequently used in studies and is measured continually and frequently (half-hourly) in a network of measuring stations around the globe.

Several complicating factors were not considered including: error covariances, non-normal error distributions, the form of the cost functions and systematic and selectively systematic biases in observations and models.

4.1.4 Layout of chapter

The methodology section (4.2) describes the specifics of the model (DALEC-D) and two DA schemes used (the EnKF and Metropolis algorithm) to this chapter. Fifteen synthetic dataset types were used, which varied in length, magnitude of observational error and size and frequency of gaps. There were two main stages to the study: (i) First, we assimilated the 15 datasets to obtain 15 sets of parameter distributions (where the mode was used as the estimate), which were compared with the true parameter set; (ii) The second stage involved using each set of parameters to forecast cumulative NEE 10 years into the future; this was compared with the true cumulative NEE for the forecast period. Also explained in (4.2) is how the synthetic NEE and meteorology data were created. In the final part of section 4.2, we outline the expectations of the results of this chapter, particularly in regard to expectations of similarities and disparities between the parameter estimates and uncertainties from the two DA schemes.

The results are displayed and described in section (4.3). In section (4.4) the results are discussed and in particular whether they support or invalidate the hypotheses. The conclusions are given in section 5.5.

4.2 Methodology

4.2.1 Outline of experiment

We used the DALEC-D ecosystem model (Fox et al., 2009; Williams et al., 2005), which includes 7 carbon pools and 17 parameters. Two data assimilation algorithms were used, namely the Markov Chain Monte Carlo (MCMC) Metropolis algorithm (Knorr and Kattge, 2005), and the Ensemble Kalman Filter or EnKF (Evensen, 2003). Only observations of Net Ecosystem Exchange (NEE) were used in DA. The observations were synthetic, which confers a number of advantages over real data, the main ones being: (1) the complications of poor model process representation (Abramowitz & Pitman, 2007) and systematic biases in eddy covariance measurements (Goulden et al., 1996; Kutsch et al., 2008; Moncrieff et al., 1996) are avoided; (2) they can provide as long a time series as needed; (3) any level of observational noise and data gaps can be specified. Section 4.2.6 provides a discussion of the use of synthetic data, including the limitations of doing so.

The synthetic NEE time series data were created by running the DALEC-D model forwards in time, with appropriately chosen initial conditions and parameters. These are referred to as the true initial conditions and parameters. To this time-series, observational error and gaps were added, based on the characteristics of the eddy-covariance NEE observations from the Harvard forest site (Munger and Wofsy, 1999). The forward run covered 70 years in total, with initial conditions for the C pools being appropriate for the site having been recently reforested. The first 10 years of the forward run were used as an initial growth period, the next 50 years for data assimilation, and the final 10 years to test the predictions. The 70 years of meteorology data (used to estimate Gross Primary Production by the model) used for the forward run were obtained by repeatedly sampling whole years from a total of 7 years of real meteorology data from the Harvard field site (Boose, 2001) (see subsection 4.2.5). Seven different lengths of the NEE dataset were assimilated (1, 2, 5, 10, 20, 30 and 50 years), along with five levels of observational error and five levels of data density. The central levels of observational noise and data density were based on real data, as explained in subsections 4.2.5 – 4.2.8. Each data assimilation run ended at the same point in time (rather than starting from the same point). This allowed easier comparison of the NEE time-series during the 10 year prediction

period, all of which started from the same point in time but with different sets of parameters arising from each of the different runs.

The experiment contained two stages. In stage 1, we estimated the model parameters and corresponding 90% posterior intervals (defined by the 5^{th} and 95^{th} percentiles of the posterior distribution) corresponding to each dataset using the two DA algorithms. In stage 2 we used the sets of estimated parameters to predict cumulative NEE 10 years into the future. We were then able to observe how the parameter estimates, posterior intervals and predicted NEE varied with dataset length, data drop-out level and observational error.

4.2.2 The DALEC-D Model

The DALEC-D model (Data Assimilation Linked Ecosystem Model for Deciduous vegetation) (Williams et al., 2005) is a simple box model that tracks the journey of carbon from photosynthesis through to C pools and respiration fluxes. It is described in full detail in subsection 2.4.4 of chapter 2.

Table 4.1 lists all the parameters and the C pools, with their corresponding symbol, together with their lower and upper bounds for this study. All of the parameters were estimated using the two DA schemes. The true values were used to create the synthetic observations (see subsection 4.2.7 for details); they were also used to compare the accuracy of the parameter estimates from the two DA schemes. The reasoning behind the choice of the lower and upper bounds given for the parameters is the same as that of Fox et al. (2009), who also used the DALEC-D model.

For this chapter, the initial conditions of the C pools were treated as parameters given that chapter 3 found that treating them as fixed could result in biases in the posterior parameter estimates for the Metropolis algorithm.

4.2.3 The Metropolis algorithm

The Metropolis algorithm is a full Markov Chain Monte Carlo (MCMC) algorithm. It uses a numerical least squares approach to find the distribution of model parameters that provides optimal fits between model's output and the observations, and can avoid local minima. Unlike traditional least squares, which finds a single set of parameters (the optimal set), the Metropolis algorithm finds a distribution of parameter sets which

Symbol	Description	Units	True value	Range (low/high)	$\underline{P}_{i,0}^{(m)}$
T_d	Litter decomposition rate parameter	day^{-1}	$5x10^{-4}$	$1x10^{-6}$/0.01	0.000739
F_g	Fraction of GPP respired autotrophically	-	0.45	0.2/0.7	0.632
F_{nf}	Fraction of GPP allocated to foliage	-	0.4	0.01/0.5	0.328
F_{nrr}	Fraction of GPP allocated to roots	-	0.4	0.01/0.5	0.293
T_r	Turn over rate of foliage	day^{-1}	0.06	$1x10^{-4}$/0.1	0.0474
T_w	Turn over rate of wood	day^{-1}	$7x10^{-5}$	$1x10^{-6}$/0.01	1.00E-06
T_r	Turn over rate of roots	day^{-1}	0.008	$1x10^{-4}$/0.01	0.00668
T_l	Turn over rate of litter	day^{-1}	0.03	$1x10^{-5}$/0.1	0.0325
T_s	Turnover rate of C_{SOM} pool.	day^{-1}	$3x10^{-5}$	$1x10^{-6}$/0.01	3.67E-05
E_t	Exponential temperature dependent rate parameter	-	0.073	0.05/0.2	0.148
P_r	Nitrogen use efficiency parameter	-	14	5/20	10.53
L_{out}	Growing degree days for leaf out	°C day	240	200/400	345.8
L_{fall}	Minimum temperature for leaf fall	°C	9	8/15	14.09
F_{ll}	Fraction of carbon loss in transfer to litter	-	0.48	0.1/0.7	0.7
T_{lab}	Turn over rate of labile carbon	day^{-1}	0.09	$1x10^{-4}$/0.1	0.0269
F_{lr}	Fraction of labile carbon respired	-	0.15	0.01/0.5	0.172
C_{fmax}	Maximum C_f	gCm^{-2}	300	100/500	350.462
C_f	Foliar carbon pool	gCm^{-2}	0*	0/100†	0.337
C_w	Wood carbon pool	gCm^{-2}	5*	0/30000	1550.2
C_r	Fine root carbon pool	gCm^{-2}	5*	0/300	138.37
C_{lab}	Labile carbon pool	gCm^{-2}	100*	0/200	92.201
C_{lit}	Fresh foliar and fine root litter carbon pool	gCm^{-2}	5*	0/200	71.292
C_{SOM}	Soil organic matter & coarse woody debris carbon pool	gCm^{-2}	9900*	0/40000	9733.56

*Table 4.1 The description of the parameters and C pools of the DALEC-D model. The * refers to the initial values for the 70 year forward run which was used to create the observations (see subsection 4.2.7). The † refers to the initial values of the C pools also at the start of one of the runs where the dataset being assimilated into the model was 50 years in length (the initial values were different depending on the length of the dataset for reasons explained in subsection 4.2.7). These initial values of the C pools were treated as parameters for the Metropolis algorithm. The range column shows the lower and upper bounds used for each parameter. The final column labelled $\underline{P}_{i,0}^{(m)}$ refers to the best guess of the ith parameter for the EnKF; while for the Metropolis algorithm it is where the algorithm was initialised from.*

are close to the set that gives the global minimum of the weighted sum of squared difference between the observed and modeled NEE. It is described in full detail in subsection 2.2.8 of chapter 2. As only NEE data is used in this chapter, the likelihood function is of the form:

$$L(m^{(k)}) = \exp(-J^{(k)})$$

where $m^{k}_{NEE,i}$ is the modelled NEE at the *k*th iterative step for the *i*th day when an observation is present, and $J^{(k)}$ is the cost function. The cost function is a weighted sum of squared differences between the observations and the corresponding model predictions and is given as:

$$J^{(k)} = \frac{1}{2}\sum_{i=1}^{L}\left[\frac{\left(m^{k}_{NEE,i} - obs_{NEE,i}\right)^{2}}{w_{NEE}}\right]$$

where w_{NEE} is the weight of the cost function, which is represented by the variance of the Normally distributed observational errors (denoted s^{2}_{NEE}). As described in chapter 2, the convergence of the Metropolis algorithm is determined formally using the Gelman criteria. A plot of the trajectories of the Gelman criteria statistic, namely Rhat, is given for each parameter at the start of the results subsection.

4.2.4 The Ensemble Kalman Filter (EnKF)

The Ensemble Kalman Filter (EnKF) is a sequential data assimilation algorithm, which, like other DA algorithms, is based on Bayes' theorem (Bayes, 1763). For terrestrial ecosystem models, the EnKF has been successfully used for state and parameter estimation (Williams et al., 2005, Quaife et al., 2007). The formulation of the EnKF used in this study is described in detail by Evensen (2003, 2009). A complete a derivation and description of how it is used for state and parameters is given in subsection 2.2.4; in particular the 'posterior' parameter distributions used here are those take from the final day of time-series.

For this study, we used 500 ensemble members for the EnKF. As stated in step 1 of subsection 2.2.4, before using the EnKF the error covariance matrices of the model and observational errors ($\mathbf{Q}_k$ and **R**), and the initial conditions ($\hat{\mathbf{x}}_0$ and $\mathbf{P_0}$) need to be estimated:

(i) **R**, the observational error covariance matrix, had dimension 1×1, since there was only one type of observation being used, namely NEE. Since observational error was one of the factors being varied, **R** was given different values according to the experimental run. Details about this are provided in subsection 4.2.7.

(ii) $\mathbf{Q}_k$, the error covariance matrix for the model, was used to represent the model uncertainty when propagating the augmented state vector of each ensemble forward from the previous time-step. In previous studies which have used the EnKF with DALEC (Williams, 2005, Fox, 2009), the off-diagonal elements of $\mathbf{Q}_k$ were set to zero, since wrongly prescribing them could lead to biases. As with these two previous studies, the standard deviations corresponding to each element of the mean augmented state of the mean augmented state vector (i.e. the square roots of the diagonal elements of $\mathbf{Q}_k$), were expressed as a fraction of the corresponding element. In other words, if $\overline{\mathbf{x}^f_{k\ (m)}}$ represents the *m*th element of the mean augmented state vector, and $(s_k)_{(m)}$ is the corresponding standard deviation, then $(s_k)_{(m)} = a \times \overline{\mathrm{x}^f_{k\ (m)}}$, where *a* is a constant between 0 and 1. For this study *a* was set to $\sqrt{10^{-6}}$ (=0.1%) and $\sqrt{10^{-5}}$ (≈0.32%) for the model state elements and the parameter elements of the augmented state vector, respectively. In Williams (2005) and Fox (2009), *a* was set to 20% for most of the elements of the state vector (in both cases, state estimation was performed rather than parameter estimation, so the parameters were not part of the state vector). However, tests carried out in this study showed that 20% caused each ensemble to behave almost entirely like a stochastic process with the model having very little influence. This is because *a* represents the model uncertainty on a day to day basis. The tests showed that after propagating the ensembles forward by 1 year, using perfect initial conditions and a daily model error with *a* = 20%, the standard deviation of the ensembles was around 450% of the size of the mean ensemble. In reality, for state estimation, whether *a* is set to 0.1% or 20% does not matter too much when observations are frequently assimilated, as in both Williams (2005) and Fox (2009). This is because the observations stop the trajectories of the ensembles spreading too much. However, large gaps in the datasets could lead to major problems. Finally, *a* was set to a lower value for the model formulation than for the parameters, since this is a synthetic study, so we treat the model formulation as close to perfect.

(iii) For the initial conditions, $\hat{\mathbf{x}}_0$ was set to $\mathbf{x}^t_0 + \kappa$, where $\mathbf{x}^t_0$ is the true initial value of the augmented state vector, and $\kappa \sim N_n(\mathbf{0}, \mathbf{P_0})$.

(iv) $\mathbf{P_0}$ was chosen to have off-diagonal elements which are zero, for the same reasons

as $\mathbf{Q_k}$. Like $\mathbf{Q_k}$, the standard deviations of the elements of $\hat{\mathbf{x}}_0$ are represented as a fraction *b*, of $\hat{\mathbf{x}}_0$, where $b = \sqrt{0.5} \approx 0.71$, so that the relative variance is 0.5. Chapter three showed that some of the variation in the parameter estimates found in the REFLEX study (Fox et al., 2009) could be attributed to different initial conditions used for the small C pools. Although this finding is based only on the Metropolis algorithm, assigning the level of uncertainty on the initial conditions of the C pools also causes difficulty for EnKF. To take this into account, the dataset length runs were repeated using a relative variance on the priors of the initial C pools of 0, 0.25, and 1.0, while keeping the relative variance on the prior values of the parameters (Metropolis algorithm) and the initial conditions of the parameters (EnKF) at 0.5 throughout. The estimates of the parameters and final C stocks were calculated by computing the mean, 5^{th} and 95^{th} percentiles corresponding to the ensemble values of the final time-step.

4.2.5 Generation of the meteorology data

To carry out the experiment, a 70-year time-series of minimum and maximum daily temperatures, total daily irradiance, and atmospheric CO_2 concentration was needed. Due to the length of the time-series, it was desirable to avoid real meteorological datasets which could contain characteristics from local climate change and therefore make it difficult to interpret the results. Hence, a synthetic dataset was generated, in which the variance of the meteorological data does not change across the entire 70 years of the time-series (in Statistics, this is referred to as a homoscedastic time-series). To do this, each year of the 70 year dataset was formed by randomly selecting a complete year from 2002 to 2008 of daily observations from the Fisher Meteorological Station (HF001), Harvard Forest (Boose, 2001) (Latitude 42.5° North, -72.2° West, elevation 340 meters). This approach maintains the intra-annual variability of seasonality of the meteorological data, whilst long term climate changes are avoided.

4.2.6 Synthetic NEE generation

Synthetic studies have several advantages over those which use real data:

- The processes which generated the synthetic NEE data and those in the ecosystem model are consistent, avoiding model and/or measurement biases.
- A synthetic dataset allows complete control over the factors being varied.
- We know the true value of the parameters and the true value of the NEE observations. This allows us to quantify how well each of the DA algorithms performs in: (1) estimating the parameters/final stocks and, (2) predicting future NEE. It also tells us which datasets result in the most accurate estimates of the parameters and predicted NEE.

Other studies have also used synthetic observations. In the OpTIC study (Trudinger et al., 2007), pseudo-data from a simplified biogeochemical test model were assimilated using 10 different DA algorithms. One interesting finding was that all DA schemes gave biased estimates when the observational noise was treated as temporally correlated or non-Gaussian. A weakness of that study is that only one non-Gaussian noise type was used (the lognormal distribution). It would have been interesting if other non-Gaussian noise types had been included. While, in reality, the distribution of errors is unlikely to be Gaussian (as the authors state), there is evidence that it could be close to the correct distribution (Hollinger & Richardson, 2005; Richardson et al., 2006). In the REFLEX study (Fox, et al., 2009), 3 years of synthetic NEE observations were generated using the DALEC and DALEC-D models (subsection 4.2.2) with temporally uncorrelated Gaussian noise added. For some parameters, the estimates from the 8 different DA algorithms used were consistent with each other and close to the true value, whilst for others there was large variation between schemes. It is thought that this was due to different initial conditions for some of the model states being used for the different schemes. Chapter 3 found some evidence of this.

It is nonetheless important to realise that there are limitations with using synthetic data. The model processes which are being used to generate the data may not relate well to real world processes. Significant work may be needed to extend the work to real datasets which might exhibit complex processes and types of observational errors not accounted for in synthetic data. For example, real datasets may contain systematic errors (see subsection 4.2.7), which were not considered in this study or REFLEX or OpTIC. Also, real observations can contain extreme or near extreme values which are not always removed.

4.2.7 Using site data in the synthetic NEE observations

The synthetic observations were created by adding synthetic gaps and observational error to the true NEE. These were constructed to be consistent with real hourly NEE measurements, taken from the eddy covariance site (site reference HF004) at Harvard Forest (Munger & Wofsy, 1999). This site is located approximately 1 mile from the Fisher meteorological weather station. The vegetation of the site is mixed, with dominant species including red oak (Quercus rubra), red maple (Acer rubrum), black birch (Betula lenta), white pine (Pinus strobus), and hemlock (Tsuga canadensis).

SPECIFYING THE DATASET LENGTH

Synthetic NEE datasets of length 1, 2, 5, 10, 20, 30 and 50 years are used. There were a number of reasons for choosing these particular lengths. The model contains fast processes (e.g. the amount of carbon stored in foliage changes seasonally), and so it is possible that the parameters directly linked to these fast processes may only need a few years of data to be estimated well. This is the reason for including the shorter lengths (1, 2 and 5 years), and is consistent with other DA studies which used a terrestrial ecosystem model, e.g. the REFLEX study used 2 years (Fox et al., 2009), while Quaife et al. (2007) used 3 years. However, other DA studies involving terrestrial ecosystem models used around a decade of data, e.g. Mo et al. (2008) used 8 years, and Braswell et al. (2008) used 10 years. Therefore, we include 10 years as a further length. The motivation for assimilating datasets longer than 10 years was from Braswell et al. (2005), who found that using parameter estimates obtained from a stochastic Bayesian technique gave a better fit to the variability patterns of NEE at the daily and seasonal timescales, and a worse fit at the annual and decadal timescales. The authors believed this was because the longer timescales were “governed by more complex processes”, and one of the ways they suggested overcoming this was to use longer datasets of at least 10 years. A further reason to use the longer datasets is that the model contains slow processes (e.g. the amount of carbon stored in woody biomass or soil organic matter) which operate over much longer time-scales, and are unlikely to be constrained by short datasets. By assimilating datasets of lengths 20, 30 and 50 years, we tested whether it might be possible to constrain the parameters linked to the slow processes. Datasets longer than 50 years are not considered because of the length of time needed to run the Metropolis algorithm. Datasets shorter than 1 year

were not considered since assimilating such datasets might bias our results depending on the period in the year we choose (e.g. summer versus winter).

The different datasets were designed in the following way: first, a 70 year synthetic dataset was generated using NEE from DALEC-D. At the start of the 70 year period, we assumed a young forest with a small woody biomass; the size of the soil carbon pool was based on Harvard forest inventory data for a time when the site had a young forest. For the first 10 years, the forest was allowed to grow, and none of this part of the 70 year period was used for assimilation. For years 11 to 60, different amounts of gaps and observational noise were added to create the synthetic datasets which were assimilated into the model to estimate the parameters and final stocks (stage 1), with years 61 to 70 used as the prediction period (stage 2). All of the datasets ended at the same point in time to make it easier to compare the forecasts made from runs corresponding to different dataset lengths. For example, the dataset of length 1 year will be in year 60, the dataset of length 2 years will cover years 59 and 60; the 5-year dataset will cover years 56 to 60; etc… (see figure 4.1).

NEE DATA DROPOUT

Daily NEE observations were created from the hourly eddy-covariance NEE observations. A day was regarded as a gap if more than two hourly measurements were missing. Thus, to determine the daily gap characteristics of the NEE data from the Harvard site, the number and the length of gaps between observations were counted. For example, if on days 5 and 10 observations are present, but none in between, then the gap between these two observations has a length of 4 days. The observed gap length is shown figure 4.2(a). To vary the frequency and size of gaps, the probability of there being data at a particular time-step was varied. The probability was determined by the proportion of data present. This proportion, labelled R_g, is defined as:

$$R_g = \frac{\{no.\ of\ \ days\ with\ observations\}}{\{total\ \ no.\ of\ \ days\}} \tag{4.2}$$

When there are data at every time-step in the dataset, $R_g = 1$, and when there are no data at all, $R_g = 0$. For the Harvard site, $R_g = 0.49$, in other words 49% of the days had observations. To generate synthetic gaps, R_g was first fixed, then the following steps were used starting with day 1:

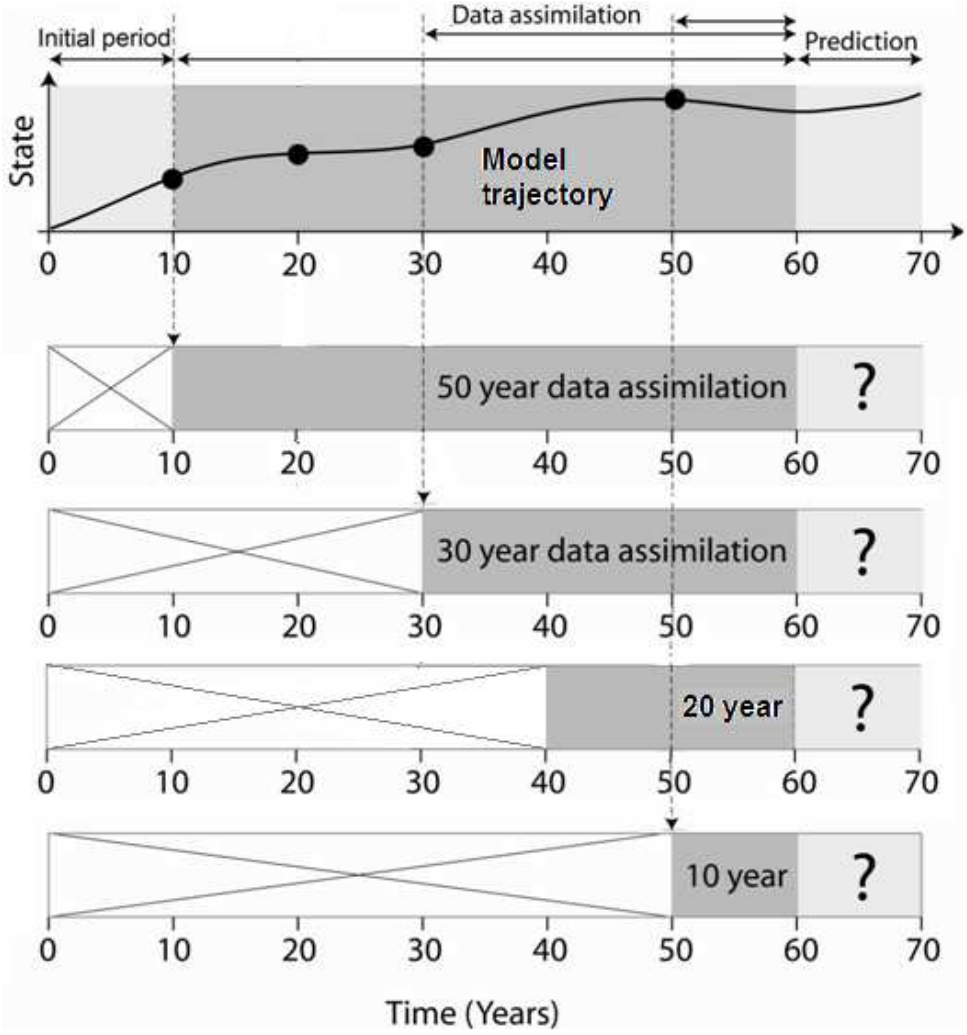


Figure 4.1: Schematic of the synthetic data usage. For clarity the 1, 2 and 5 year periods have been omitted from the diagram.

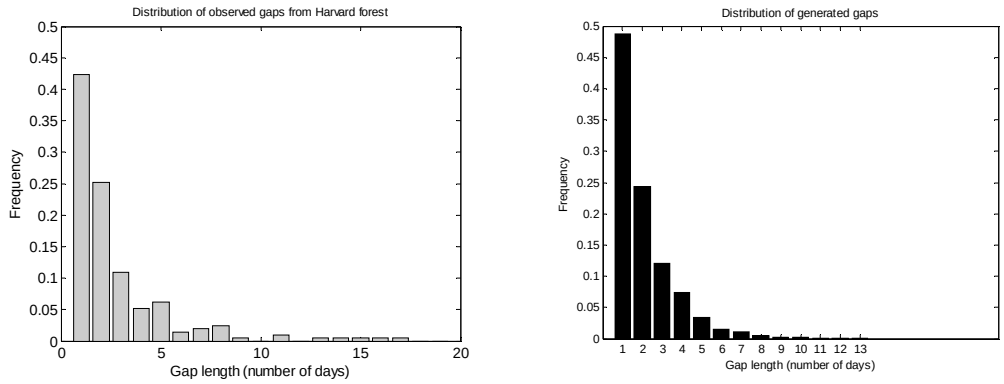
(1) Generate a random number between 0 and 1 from the uniform distribution.

(2) If this number is less than R_g then an observation is present for the current day.

(3) If the random number is greater than R_g, then no observation is present for the current day.

(4) Go to the next day and repeat (1) to (3) until the end of the dataset is reached.

The probability distribution that describes this distribution of gaps is the Geometric distribution. We can check the accuracy of the method by generating gaps using the R_g value from the Harvard forest (0.49). One can see that the distribution of the generated gaps (figure 4.2b) is close to that of the observed gaps (figure 4.2a).



(a) (b)

Figure 4.2 (a) The distribution of observed gap lengths from Harvard forest; (b) The distribution of generated gaps.

Five levels of data dropout were formed by assigning the values of 0.2, 0.4, 0.6, 0.8, and 1.0 to R_g. These correspond to 20%, 40%, 60%, 80%, and 100% of the days having data present.

OBSERVATIONAL ERROR

Observational error in NEE is defined as the difference between the measured and true values of NEE. This can be split into random and systematic elements Goulden et al.,1996; Moncrieff et al., 1996). Systematic errors result from, for example, the measuring equipment, if incorrectly calibrated. The eddy-covariance method has various assumptions, the main ones being: (1) the terrain must be flat; (2) the environmental conditions are steady (Baldocchi, 2003). If these assumptions are not met, this may cause systematic errors in the measurements. Another type of systematic error can arise from low winds at night-time causing under-estimation of NEE (for details of all types of systematic error, see Baldochi (2003)). Modelling systematic errors is difficult due to the variety of different factors that can cause them, and the likelihood of them varying from site to site. Therefore, for simplicity, systematic errors were not considered here. A random error is defined as:

> *An error of measurement as a consequence of recording the value* $y + \varepsilon$ *instead of the true value y, with* ε *being an observation on a random variable. The random error is often assumed to have a normal distribution with mean 0 and constant variance (though this assumption should always be verified)*
>
> (Oxford dictionary of Statistics, 2004).

The distribution of the noise was determined by looking at the variability of hourly NEE measurements from the eddy covariance site, which for this study was at Harvard Forest. To do this, we compared hourly NEE measurements taken at exactly the same hour in the day and with similar conditions. A pair of measurements, x_1 and x_2, with similar conditions are defined as those where the difference in photosynthetic flux densities (i.e. the rate of photosynthesis) is less than 75 μmol $s^{-1}m^{-2}$, the difference in air temperatures is less than 3°C, and the difference in wind speeds is less than $1ms^{-1}$ (Hollinger & Richardson, 2005; Richardson et al., 2006). Let $x = x^* + \varepsilon$ be a general measurement, where the random error ε is an i.i.d. (identically and independently distributed) measurement, distributed with mean zero and variance $\sigma^2(\varepsilon)$ and x^* is the true value. These hourly noise term ε, was found to approximately follow a Laplace distribution (see subsection 2.5.3 for details), which has probability density function:

$$f_\varepsilon(x) = \frac{1}{2b}\exp\left(\frac{-|x-\mu|}{b}\right), \tag{4.3}$$

where μ and b are referred to as the location and scale parameters, with the hat denoting that they need estimating. This distribution has a mean and variance given by μ and $2b^2$. Using the hourly NEE data from the Harvard site, the Laplace distribution with $\hat{\mu}$ =0.002 μmol/m²/s (1 s.f.), and $\hat{b}$=1.94 μmol/m²/s (2 d.p.) gave the best fit to the noise terms. However, DALEC-D works on a daily time-step, and to determine the distribution of random errors in daily NEE observations, the hourly random errors were first converted to daily ones. If the daily noise term is represented by $\phi \sim \varepsilon_1 + \varepsilon_2 + \varepsilon_3 + ... + \varepsilon_{24}$ where ε_i is the noise term for the ith hour of the day, then using the central limit theorem, ϕ is approximately normally distributed (Pólya, 1920).

As well as summing up 24 hourly noise terms, it is also necessary to convert from the units used for the eddy-covariance NEE measurements (μmol/m²/s) to those used in DALEC-D at the hourly time-scale (gC/m²/hr). First note that there are 12 grams of carbon for 1 mole, and 1 mole is equal to 1,000,000 micromoles (μmols). Thus, letting X_i be the NEE noise term with units μmol/m²/s and Y_i with units gC/m²/hr calculated on the ith hourly time-step ($1 \le i \le 24$), the conversion between the two is:

$$Y_i = \frac{12 \times 60 \times 60}{1{,}000{,}000} X_i$$

We multiply by 60 twice since the units of the units of X_i is per second. This can be simplified to:

$$Y_i = c\, X_i$$

where $c=0.043$ (2 s.f.). Now let Z be the DALEC-D NEE noise term at the daily time-step (i.e. gC/m²/day). Then,

$$Z = \sum_{i=1}^{24} Y_i \qquad \Rightarrow \qquad Z = \sum_{i=1}^{24} cX_i$$

We assume that X_i and X_j are i.i.d for all $1 \le i, j \le 24$, thus $E(X_i)=E(X_j)=E(X)$, $Var(X_i)=Var(X_j)=Var(X)$, and $Cov(X_i, X_j)=0$. Therefore, the expectation and variance of Z is:

$$E(Z) = \sum_{i=1}^{24} cE(X_i)$$

$$\Rightarrow \qquad E(Z) = 24cE(X) \qquad (4.4a)$$

$$Var(Z) = \sum_{i=1}^{24} c^2 Var(X_i)$$

$$\Rightarrow \qquad Var(Z) = 24c^2 Var(X) \qquad (4.4b)$$

From the text after equation (4.3), $E(X)=0.002$, $Var(X)=2b^2=2\times 1.94^2=7.5272$, and with $c=0.043$, equations (4.4a) and (4.4b) become:

$$E(Z) = 24cE(X) = 24 \times 0.043 \times 0.002$$

$$\Rightarrow \qquad E(Z) = 0.002 \text{ (1 s.f.)} \qquad (4.5a)$$

$$Var(Z) = 24c^2 Var(X) = 24 \times 0.043^2 \times 7.5272$$

$$\Rightarrow \qquad Var(Z) = 0.3340270272\ldots$$

$$\Rightarrow \qquad \sigma(Z) = 0.58 \text{ (2 s.f.)} \qquad (4.5b)$$

Thus from equations (4.5a) and (4.5b), the resulting distribution of the NEE noise term using gC/m²/day units is Gaussian with mean 0.002gC/m²/day and standard

deviation 0.58gC/m²/day. The observational error was varied by varying the value of *b* over five levels, which were determined by computing 50%, 75%, 100%, 125% and 150% of b = 1.94, which results in: b = 0.97, 1.455, 1.94, 2.425 and 2.91 μmol/m²/s with μ=0.002 μmol/m²/s. At the daily time-step and using equation (4.4b), this results in errors corresponding to Gaussian distributions with standard deviations given by: σ = 0.29, 0.43, 0.58, 0.72 and 0.87 gC/m²/day (2 d.p.) with a fixed mean of μ=0 gC/m²/day.

4.2.8 Design and set-up of the experiment

Summarising the last three subsections, estimates of parameters and the final C stocks were obtained using the EnKF and Metropolis algorithm DA schemes, and by varying three different factors by the following levels:

Dataset length: 1, 2, 5, 10, 20, 30, 50 years.
Data dropout (R_g): 0.2, 0.4, 0.6, 0.8, and 1.0 (no units).
Observational error (σ): 0.29, 0.43, 0.58, 0.72 and 0.87 gC/m²/day.

The most thorough type of design is the full factorial design, in which every level of each factor is combined with each level of the other factors. This produces 7×5×5=175 runs for each DA scheme. An advantage of this design is that it allows one to explore interactions between the different levels of the 3 factors. For example, suppose we did runs corresponding to the values of the 3 factors: {10 years, 0.58 gC/m²/day, R_g } and {1 year, 0.58 gC/m²/day, R_g } where R_g varies by the 5 specified levels. Then the pattern of estimates for each parameter and final stocks would be different, and this may provide some insight into the effect of the dataset length. However, the computation time to carry out the full factorial analysis would be too long, particularly for the Metropolis algorithm. For example, carrying out the 50 year run takes roughly 1 day. An alternative, which requires only 15 runs, is to vary the levels of one factor, while keeping the other two factors fixed at their central values. The resulting 15 runs are displayed in table 4.2.

Both the EnKF and the Metropolis algorithm needed prior estimates of the parameters and the initial carbon pools, but each uses this information in different ways. For the EnKF, since the parameters are treated as states and allowed to evolve,

Run	Dataset length (years)	Data dropout (R_g)	Observational error, σ (gC/m²/day)
1	**1**	0.6	0.58
2	**2**	0.6	0.58
3	**5**	0.6	0.58
4	**10**	0.6	0.58
5	**20**	0.6	0.58
6	**30**	0.6	0.58
7	**50**	0.6	0.58
8	10	**0.2**	0.58
9	10	**0.4**	0.58
(4)	*10*	***0.6***	*0.58*
10	10	**0.8**	0.58
11	10	**1.0**	0.58
12	10	0.6	**0.29**
13	10	0.6	**0.43**
(4)	*10*	*0.6*	***0.58***
14	10	0.6	**0.72**
15	10	0.6	**0.87**

Table 4.2 The 15 different runs which were carried out using both DA schemes.

they needed to be initialised at the start of the time-series, along with the C pools; the prior estimates were used for this. For the Metropolis algorithm, the prior values of the parameters were used as a starting location from which the algorithm began its search of different parameter sets. In this DA scheme, the initial values of the C pools were treated as 6 additional parameters, which were initialized from the same values (i.e. $\underline{P}_{i,0}^{(m)}$ in model parameter space) as the means of the prior distributions used for the C pools in the EnKF. The prior distribution for model parameter P_i is assumed to be Gaussian for the EnKF with a mean denoted by $P_{i,0}^{(m)}$ (for simplicity we use the same notation from the initialization value of model parameter P_i for the Metropolis algorithm). The relative variance and standard deviation of this prior distribution is given by $V_i = 0.5 \times \left(P_{i,0}^{(m)}\right)^2$ and $SD_i = 0.71 \times \underline{P}_{i,0}^{(m)}$. For the Metropolis algorithm the prior distribution of the log-normalized parameters were also assumed Gaussian but each had a mean of 1 and a variance 0.5. The choice of a large relative variance reflects the fact it may not be possible, or if possible very infrequently, to make measurements directly related to the parameters. A variance of 50% means that we know the correct order of magnitude of the model parameter, which is approximately

correct in terms of the turnover rates and sizes of the C pools. $P_{i,0}^{(m)}$ was determined by generating a random number from the Gaussian distribution with mean 0 and variance V_i, and then added to $P_i^{(t)}$, the true value of the parameter. $P_i^{(t)}$ was determined arbitrarily by Mat Williams, the creator of DALEC-D.

4.2.9 Analysis of predictive ability

Estimating the parameters is important, but being able to accurately predict NEE is also critical. Therefore, for each of the 15 runs listed in table 4.2, we used the corresponding parameter estimates and final stock estimates to run the DALEC-D model 10 years into the future. To incorporate uncertainty into the forecasts, all 500 of the ensembles were used for the EnKF, and 500 sets of parameters and corresponding final stock estimates were selected at random from the posterior distribution for the Metropolis algorithm.

4.2.10 Expectations

In subsection 2.3.4, expectations of the EnKF and Metropolis algorithm for estimating parameters and associated uncertainties was given, with direct reference to chapter 4. The purpose of this subsection is to provide a summary of subsection 2.3.4 in order to outline the key points where we would expect differences and similarities in the results from the two algorithms.

The two DA algorithms are both based on Bayes' theorem, yet how they apply this theorem is different with the EnKF being sequential in nature (subsections 2.2.2 to 2.2.7) whereas the Metropolis algorithm is a global search algorithm (subsections 2.2.8 and 2.2.9). Despite this difference, they are still solving Bayes' theorem and so theoretically should lead to the same solution or posterior distribution. In this chapter, there are two minor differences in the setup of the two algorithms which potentially could result in differences in the resulting posterior distributions. These include: (i) a slightly different prior distribution on the parameters used and (ii) slightly different model errors specified. In subsection 2.3.4, it is explained in detail why these differences in setups are very likely to produce insignificant differences in the respective posterior distributions. In summary, for (i) this is because although the

model parameters in the case of the Metropolis algorithm are assumed to follow a log-Normal distribution, the log-normalised parameters as used in the actual algorithm follow a Normal distribution. Moreover the size of the variance used for the Normally distributed log-Normalised priors is the same in relative terms to the variance used in the prior distribution for the EnKF approach. For (ii), the Metropolis algorithm assumes no model error whereas the EnKF incorporates very small model error. In this chapter, 13 of the 15 runs assimilate datasets with at least 60% of daily data present; therefore the gaps between the data points rarely become large for the model error to grow to anything beyond insignificant. As a result of this and due to the differences between the two approaches being so minor it is not expected that this will yield anything beyond very slight differences in the respective posterior distributions.

To demonstrate that these slight differences in the setups between the Metropolis algorithm and the EnKF result in very similar parameter estimates and uncertainties, figure 2.8 in subsection 2.3.4 shows such estimates of the DALEC-D parameters from the REFLEX study (Fox et al., 2009) obtained from these two DA schemes. As described in subsection 2.3.4, we can see that for 13 of the 17 parameters, the estimates from the two schemes are very similar and for 3 of the remaining parameters the estimates are fairly close. The length of the uncertainty intervals are also in good agreement between the two schemes, with either good or very good agreement for 15 of the 17 parameters.

In REFLEX, the length of the assimilated dataset was two years, which is the same as run 2 used in this chapter except that no LAI data was used here. Therefore it is reasonable to assume that the parameter estimates and uncertainty intervals are likely to be the same or similar for the Metropolis algorithm and the EnKF when assimilating short datasets (1 to 5 years in length), which correspond to runs 1 to 3 (table 4.2). Virtually no terrestrial carbon DA study and no known DA inter-comparison study has assimilated a dataset 10 years of more in length or under varying observational uncertainty or varying data density. Therefore it is difficult to back up any expectation of differences or similarities of the two DA schemes in estimating the posterior distribution of the parameters. Instead we briefly outline factors that might result in differences in the solution between the two schemes as a result of assimilating a longer time-series of data or data with varying observational uncertainty or data density, with more detail given in subsection 2.3.4.

Equifinality is a known issue which can cause variations in parameter estimates produced by global search algorithms such as the Metropolis algorithm – see Mdlyn et al. (2005), Tang & Zhang (2008), Williams et al. (2009) and Luo et al. (2011). Equifinalty occurs when different parameter sets result in similar outputs, in this case similar fits of the modelled NEE to the observed NEE. In DA, one reason for its cause is the existence of parameters which may have little impact on the resulting NEE, i.e. the NEE is relatively insensitive to variations of these redundant parameters. In the above four studies, equifinality was identified to occur using datasets of lengths up to 5 years. However its impact when longer datasets are used or those with varying observational error and data density is unclear. It is reasonable to expect it to occur to the same level for longer datasets at least for synthetic datasets used in this chapter where the inter-annual variations in the NEE data is minimal. In other words, a parameter set which results in a good fit of the modelled to observed NEE using a short dataset, may be expected to give a similar relative fit if a longer dataset was assimilated. If we vary the density of data in the dataset, it is reasonable to expect equifinality to occur for the more sparse dataset because the cost function is based on fewer data points. For example, imagine that the dataset consisted of just one data point, then the trajectory of the modelled NEE for the other time points could vary immensely as long as the fit of the modelled NEE to the observed NEE and the location of the observed NEE was the same. As we increase the number of data points, the trajectory of the modelled NEE is constrained more to a particular shape. In a similar way, it is reasonable to expect equifinality to occur more when assimilating a noisy dataset compared to a non-noisy one.

Medley et al. (2005) also states that equifinality could occur as a result of model structure (see subsection 2.3.4 for detail), therefore the EnKF may also suffer from equifinality for this reason. However the impact of equifinality is likely to be less on the EnKF than for the Metropolis algorithm. This is because the EnKF is a sequential DA scheme, and does operate by trying to improve the fit of the modelled to observed NEE. Instead the EnKF operates by applying Bayes's theorem one time-step at a time, rather than all in one go as is the case of the Metropolis algorithm.

Another possible reason for the differences in the parameter estimates and uncertainties between the two DA schemes could be because the residuals between the modelled and observed NEE are autocorrelated. Although the Gaussian noise, added

to true NEE when creating the synthetic NEE data, are independent in time the same is not guaranteed in the residuals. As the independence of the residuals is an assumption of the likelihood, breaking this could result in biased parameter estimates and inaccurate uncertainty estimates. Therefore, this could potentially cause discrepancies in the estimates of the parameters and uncertainties from the two algorithms. However in chapter 3, it was found the residuals were not autocorrelated so this is not expected to be true in this chapter either.

A final potential reason is that the Metropolis algorithm fails to converge to the posterior distribution. To ensure that this is not the cause of any potential differences, the Gelman criteria is applied as was done in chapter 3.

4.3 Results

4.3.1 Overview of the results section

Since the results shown here are very comprehensive and detailed a quick overview of the layout of this section is first given. The results can be split up into four parts:

Part 1: In subsection 4.3.2, we assess the convergence of the Metropolis algorithm and check the Normality of the resulting posterior distributions.

Part 2: The parameter estimates and the corresponding 90% posterior intervals for both DA schemes are then presented in subsection 4.3.3. In the four subsections that follow (4.3.4 – 4.3.7), we assess the performance of the parameter retrievals and uncertainties in different ways which include: the proximity of the posterior parameter sets to the truth, how many posterior intervals contain the truth, whether there was a reduction in the length of uncertainty intervals from the posterior compared to the prior, and how well the parameters in the posterior distributions are correlated with other.

Part 3: In subsections 4.3.8 and 4.3.9, we assess the accuracy and robustness of the iNEE (cumulative NEE) forecasts.

Part 4: Finally, results relevant to the discussion section (section 4.4) are given. These include: (i) assessing the sensitivity of modelled NEE (using the DALEC model) to changes in the parameters; (ii) assessing the autocorrelation of the residuals in the

likelihood function by constructing lag plots; (iii) plotting the Root Mean Squared Error between the modelled and observed / true NEE for each of the 15 runs.

4.3.2 Assessing convergence of the Metropolis algorithm and assessing Normality of posterior PDFs

CHECKING CONVERGENCE OF THE METROPOLIS ALGORITHM

The convergence of the chains in the Metropolis algorithm was assessed by monitoring the Gelman criteria statistic Rhat for each of the iterations (Gelman, 1995). Figure 4.3 shows the trajectories of Rhat for the length of the chains for four of the 15 runs used in this chapter, namely runs 1, 6, 11 and 15. The corresponding plots for the remaining 11 runs are very similar to these plots. Some of the runs converged very quickly, for example run 1 converged within 100,000 iterations, whereas others took a little longer with all converging before 5 million iterations and in the vast majority of cases well before this point.

ASSESSING NORMALITY OF THE POSTERIOR PDFS

In this section, we describe the estimates of the parameters, final stocks and forecasts. For both DA schemes, the mean was used as the estimate of each parameter under the assumption that all the posterior distributions were Gaussian. To check this, quantile-quantile plots of the marginal distributions for each parameter were plotted for each run and for each algorithm. A quantile-quantile plot, or q-q plot, plots each of the N-quantiles from the marginal posterior distributions and the theoretical Normal distribution. In this case, we set N to be 100 meaning that we plot each of the percentiles from both distributions. If the posterior distribution is Normal then the percentiles from the posterior distribution should match those of the theoretical Normal distribution, and this is indicated by the points following the y=x line. The marginal distributions were also plotted alongside the q-q plots.

For the Metropolis algorithm, most parameters had marginal distributions that were approximately Normal for most or all of the runs, for example the marginal distribution for the T_l parameter of run 1 is given in figure 4.4a and the corresponding q-q plot (figure 4.5a) shows the points following the y=x line very well. Evidence of non-normality for around half of the runs was evident in 7 of the 23 parameters, name-

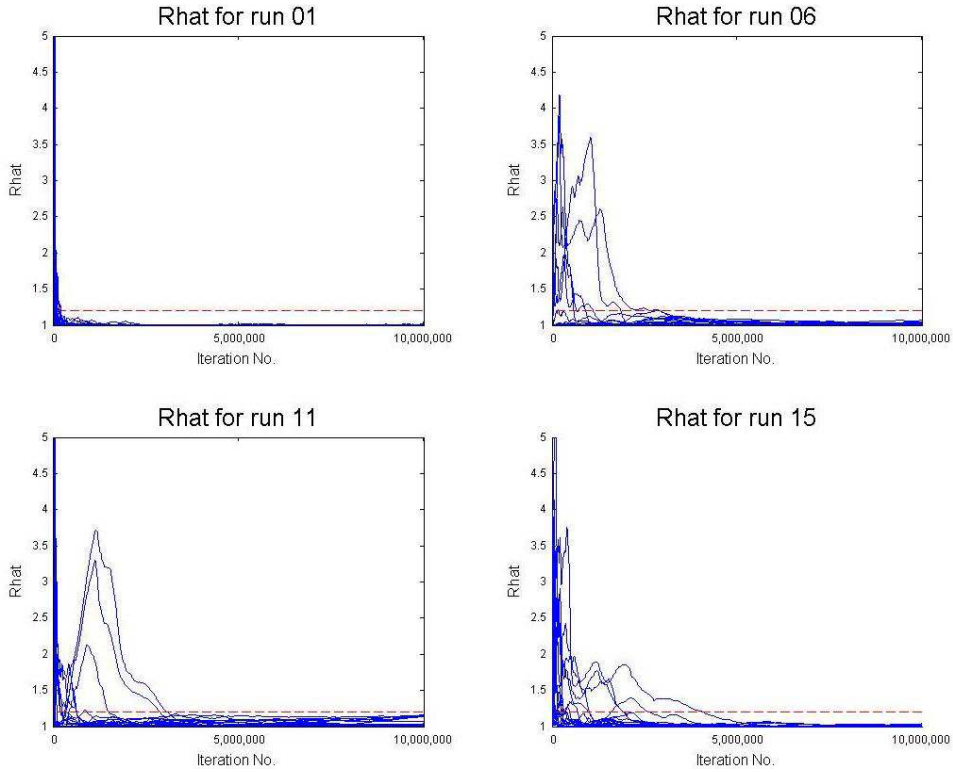


Figure 4.3 Plots of Rhat against the iteration number for 4 of the 15 runs from the Metropolis algorithm. The red dashed line is the y=1.2 line and when the Rhat trajectories all fall below this threshold,the chains have converged for the parameters.

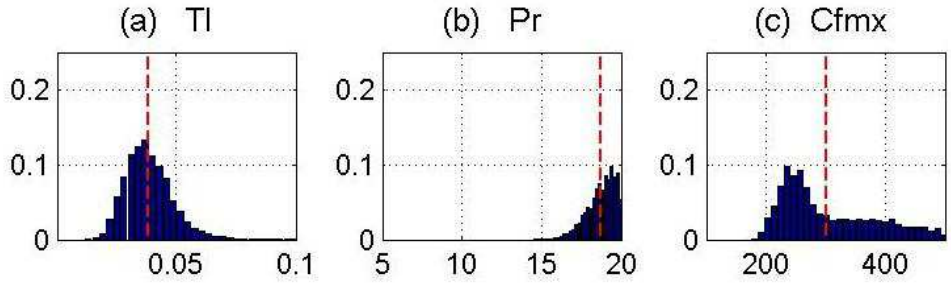


Figure 4.4 The posterior distributions corresponding to three parameters of the Metropolis algorithm: (a) T_l (Turnover of labile carbon) for run 1, (b) P_r (Nitrogen use efficiency parameter) for run 5; (c) C_{fmax} (Maximum C_f value) for run 3. The red dotted line is the mean of the distribution.

ly T_d, F_{nf}, T_r, P_r, T_{lab}, $C_{w,0}$ and $C_{som,0}$. For these parameters, the distributions had a tail on one side, but because the mean was close to one of the bounds, there was no tail on the other side. A good example of this is the P_r parameter in run 5 with the marginal distribution shown in figure 4.4b and the Q-Q plot showing serious deviation away from the y=x line for the 100^{th} percentile point. Evidence of non-normality for most of the runs was evident in one parameter and 2 final C stocks, namely C_{fmax}, C_w and

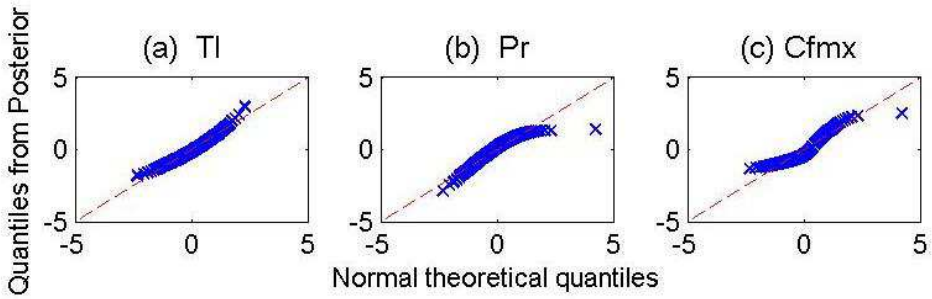


Figure 4.5 The Q-Q plots corresponding to the corresponding marginal posterior distributions given in figure 4.4.

C_{som}. Due to the non-normality of some of the parameters, the mode of the posterior distribution was used to represent each parameter estimate. For the parameters which had Normal posterior distributions, the mode is the mean so using the mode made no difference to the estimate.

For the EnKF, with the exception of T_{lab} and C_{fmax}, all the parameters and final C stocks had posterior distributions which were approximately Normal for all or virtually all of the 15 runs. For just under half of the runs for T_{lab}, the distributions were similar to figure 4.4b, and for most of the runs for C_{fmax} the distributions were similar to figure 4.4c (with corresponding Q-Q plot given in figure 4.5c). For virtually all of the forecast runs, the distributions were approximately Normal.

4.3.3 Parameter estimates and their 90th posterior intervals

The complete set of parameter and final C stock results is displayed in figures 4.6 – 4.9. In all of figures 4.6 – 4.8, the upper bound of the T_w parameter is 10% of the actual upper bound given in table 4.1. This was done because using the actual upper bound meant that it was difficult to distinguish between the parameter estimates and uncertainties from the different runs. However for completeness, the results for this parameter have been plotted again with the actual upper bound from table 4.1 used (figure 4.9).

The behaviour of the parameter estimates can be split into three broad categories, exemplified by the behaviour of the three parameters shown in figure 4.10:

(i) <u>The parameter estimate improved with dataset length</u>: As dataset length increased, (I) the parameter estimate got closer to the true value (red line in figure 4.10), <u>and/or</u> (II) the 5th and 95th percentile range became narrower, while still containing

the true value. L_{out} (growing degree day for leaf out) is an example of this, satisfying both (I) and (II) (figure 4.10a). Note that the blue line is the prior value.

(ii) The parameter estimate deteriorated with dataset length: As dataset length increased, (I) the parameter estimate got further away from the true value, and/or (II) the 90^{th} posterior interval became narrower, while not containing the true value, and/or (III) the 90^{th} posterior interval widened, while still containing the true value. F_{lr} (the fraction of labile carbon respired) is an example of this, and satisfies (I) and (II) (figure 4.10b). Note that the blue line is the prior value.

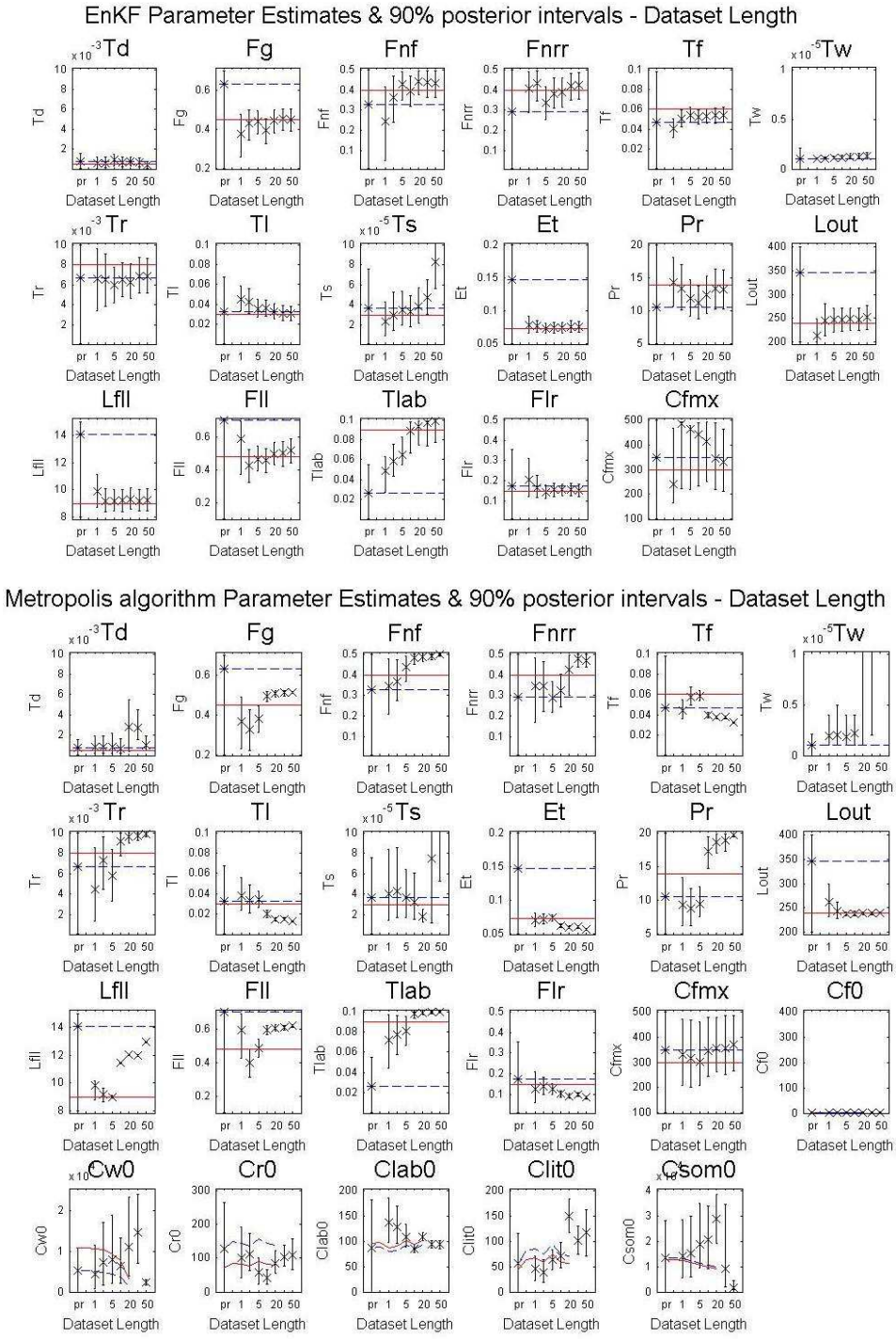
(iii) The parameter changed with dataset length, but with no consistent pattern: As dataset length increased, the parameter estimate neither consistently improved nor deteriorated. P_r (the Nitrogen use efficiency parameter for photosynthesis) is an example of this (figure 4.10c).

The parameter estimates can be grouped similarly as observational error decreases and the density of observations increases.

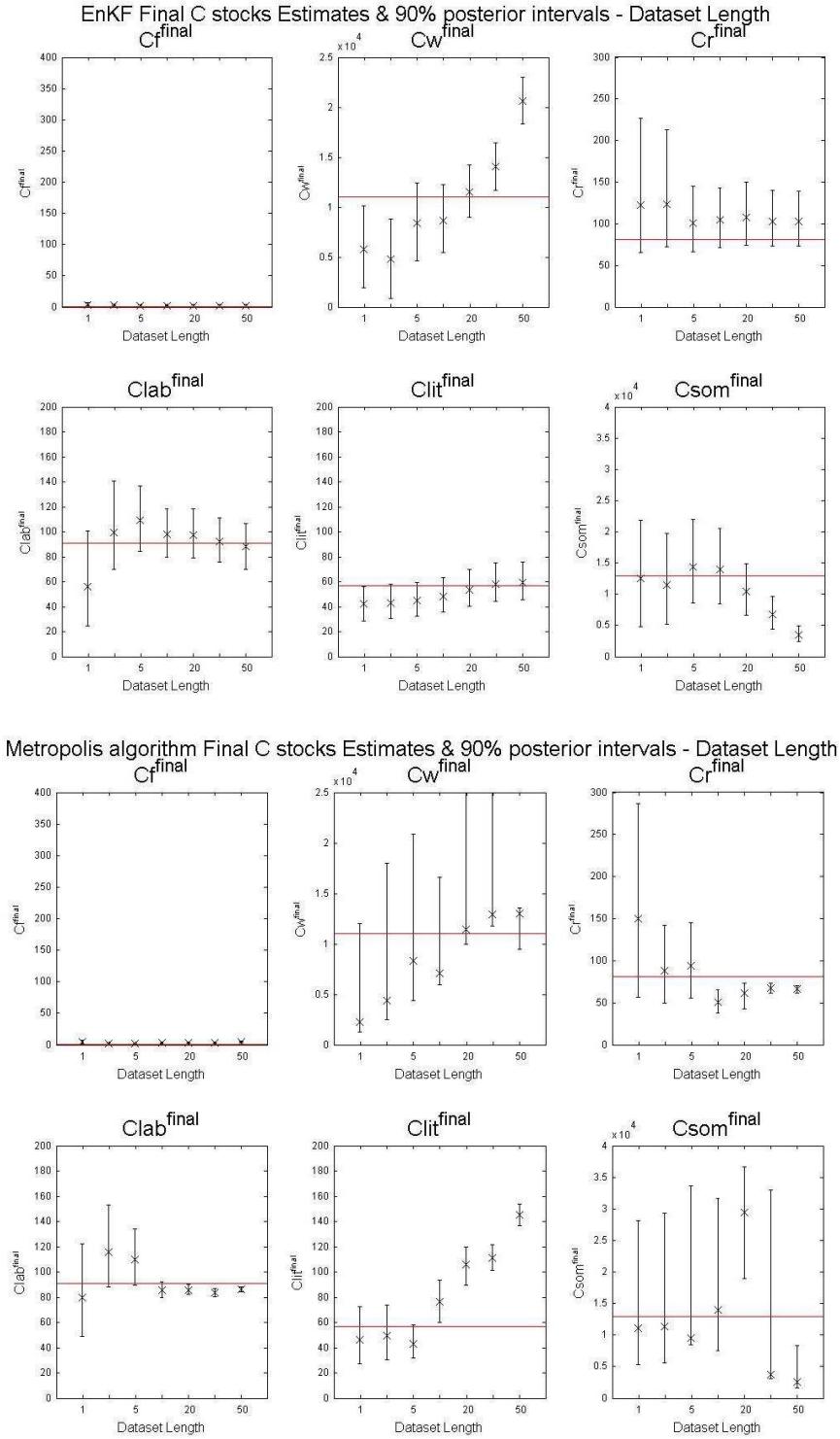
A convenient way to display these 3 types of performance for parameter estimates and final stock estimates is through the colour representation used in tables 4.3 and 4.4, with green denoting improvement, red representing deterioration and yellow for no consistent pattern. For the EnKF, most of the parameter and final stock estimates showed improvement, though this was predominantly true with the dataset length runs, i.e. increasing the density of observations and reducing the observational error gave only minimal improvement in the estimates. Moreover, most of the EnKF parameter estimates took 5 to 10 years of data to converge to a value close to the truth. Interestingly, the only parameter which shows deterioration for the EnKF is T_s.

In contrast, most of the estimated parameters from the Metropolis algorithm deteriorated. Those showing improvement related to 3 parameters and 2 initial C pools, namely F_{nrr} (fraction of NPP allocated to roots), L_{out} (growing degree day value causing leaf out), C_{fmax} (maximum C_f value), and $C_w^{(0)}$ and $C_{lab}^{(0)}$ (initial values of the woody and labile carbon pools).

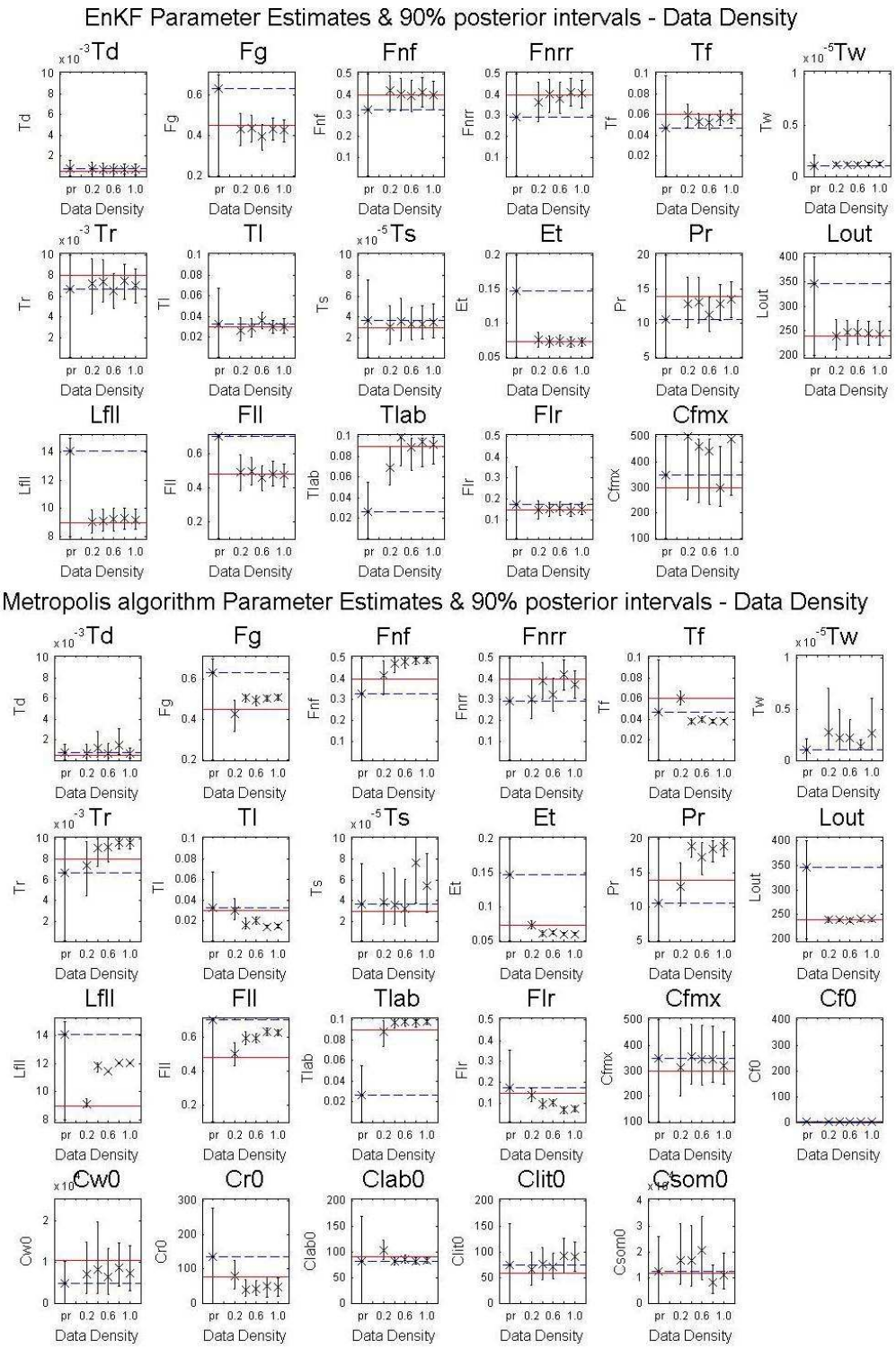
Finally, an interesting observation from the results is that, for the EnKF runs, the posterior uncertainty on the parameters does not appear to increase with an increase in observational uncertainty. This is intriguing and is discussed, along with the other observations made, in section 4.4



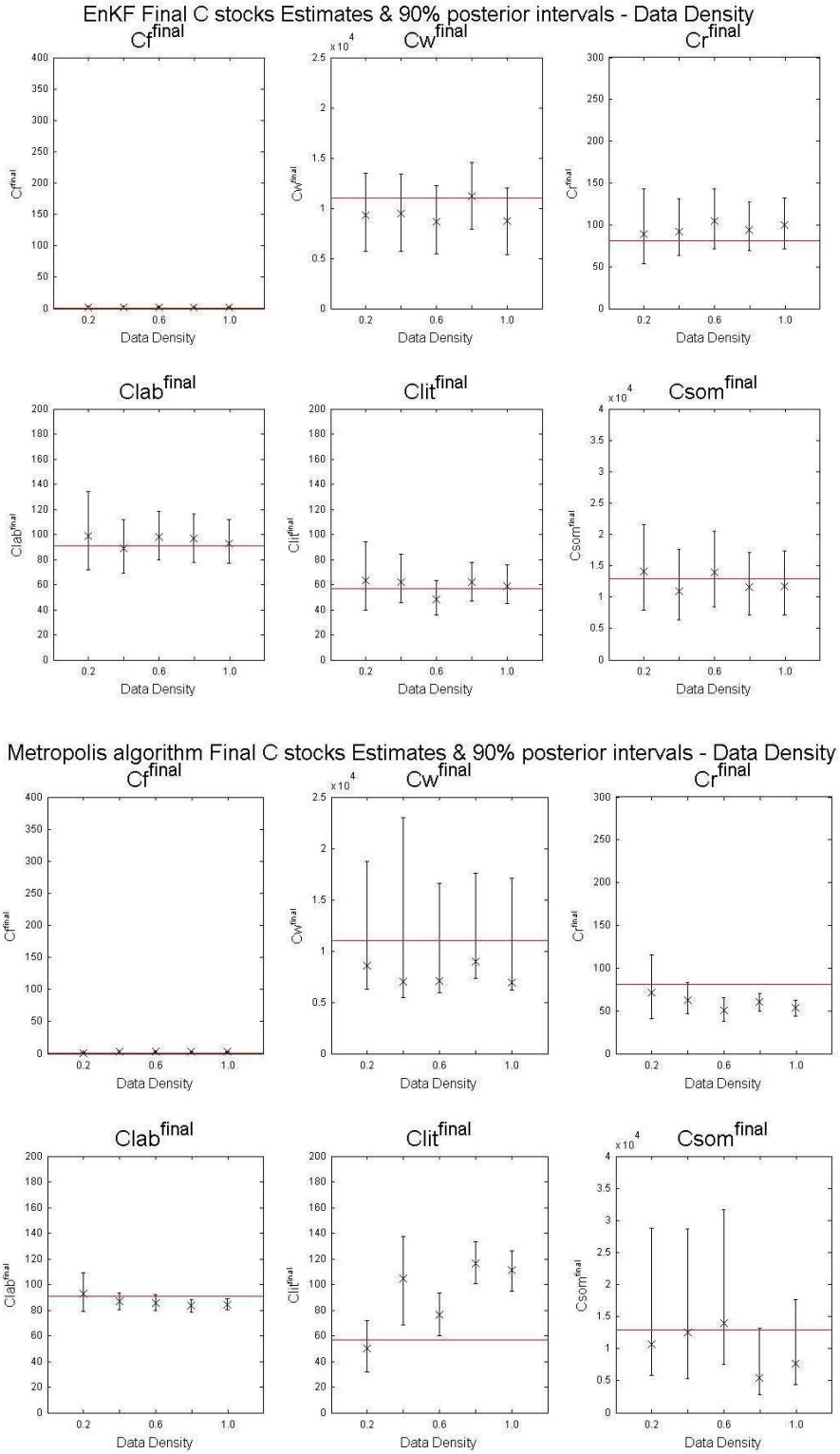
Figures 4.6a (upper panel) and 4.6b (lower panel): The effect of dataset length on parameter estimation, with upper and lower panel of plots denoting those from the EnKF and the Metropolis algorithm respectively. C_{f0} to C_{som0} are the 6 carbon pools required by the Metropolis algorithm, while the remaining 17 are the DALEC-D parameters. The cross represents the mean value of the posterior distribution with the error bars indicating the 5th and 95th percentiles. The red and blue lines are the true parameter value and the prior values, which vary for the initial carbon pools because each of those datasets start from different points in time. Note that the prior value is also denoted by 'pr' where the error bar represents the prior uncertainty.



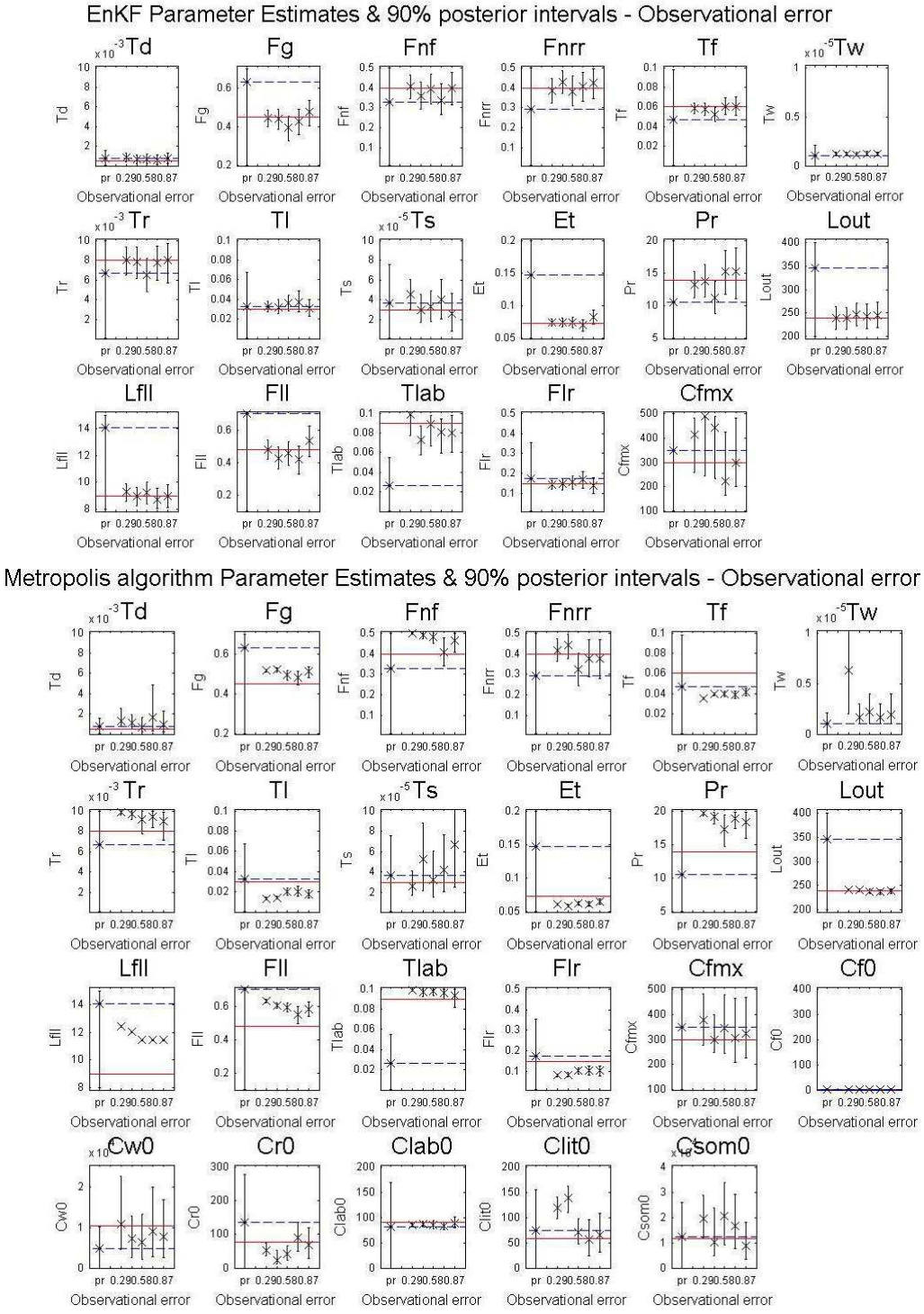
Figures 4.6c (upper panel) and 4.6d (lower panel): As for figures 4.6a and 4.6b respectively, but showing the effect of time-series length on final carbon stock estimation.



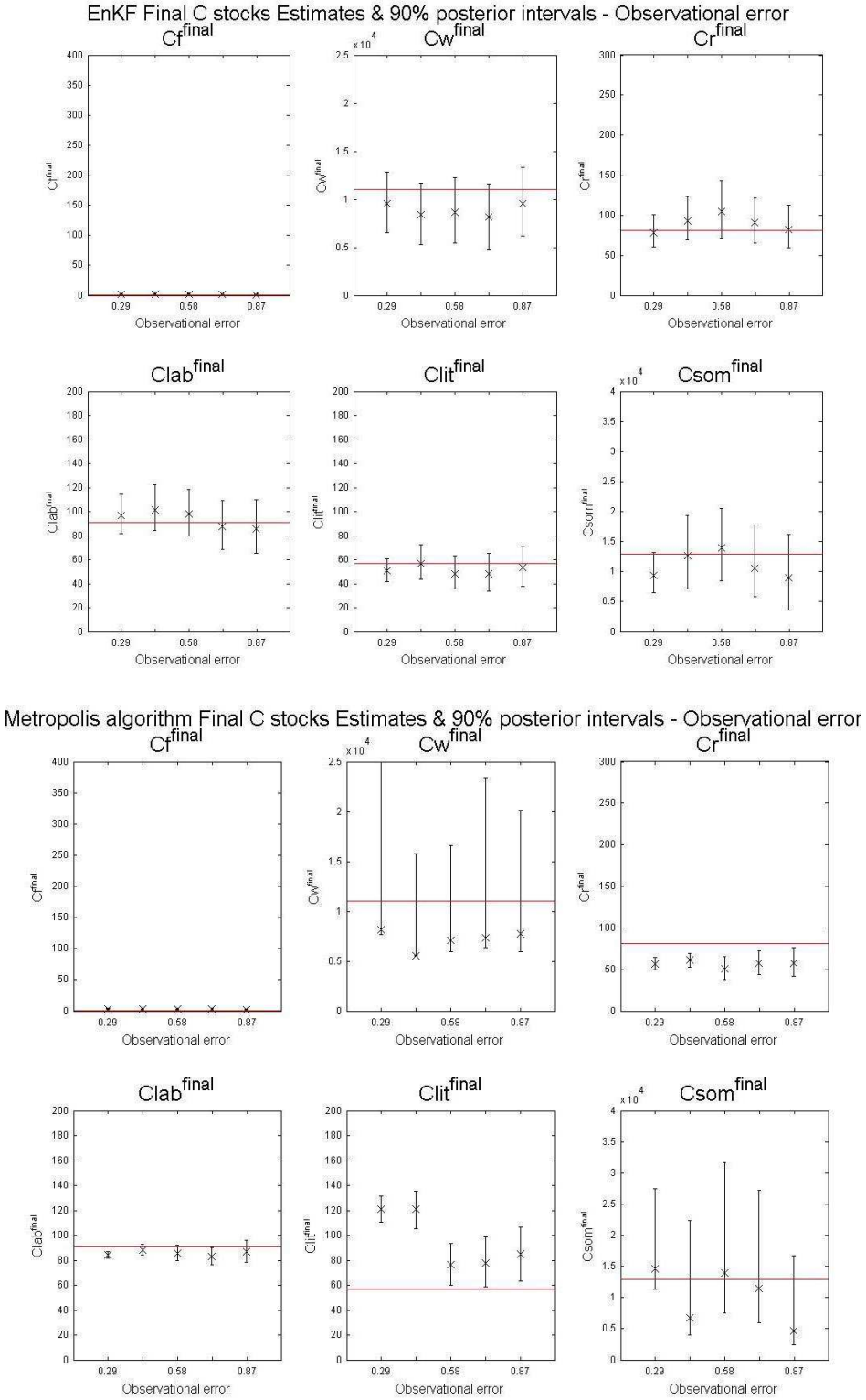
Figures 4.7a (upper panel) and 4.7b (lower panel): The effect of data density on parameter estimation, with upper and lower panel of plots denoting those from the EnKF and the Metropolis algorithm respectively. C_{f0} to C_{som0} are the 6 carbon pools required by the Metropolis algorithm, while the remaining 17 are the DALEC-D parameters. The cross represents the mean value of the posterior distribution with the error bars indicating the 5th and 95th percentiles. The red and blue lines are the true parameter value and the prior values.



Figures 4.7c (upper panel) and 4.7d (lower panel): As for figures 4.7a and 4.7b respectively, but showing the effect of data density out on final carbon stock estimation.



Figures 4.8a (upper panel) and 4.8b (lower panel): The effect of observational error on parameter estimation, with upper and lower panel of plots denoting those from the EnKF and the Metropolis algorithm respectively. C_{f0} to C_{som0} are the 6 carbon pools required by the Metropolis algorithm, while the remaining 17 are the DALEC-D parameters. The cross represents the mean value of the posterior distribution with the error bars indicating the 5th and 95th percentiles. The red and blue lines are the true parameter value and the prior values.



Figures 4.8c (upper panel) and 4.8d (lower panel): As for Figure 4.8a and 4.8b respectively, but showing the effect of observational error out on final carbon stock estimation.

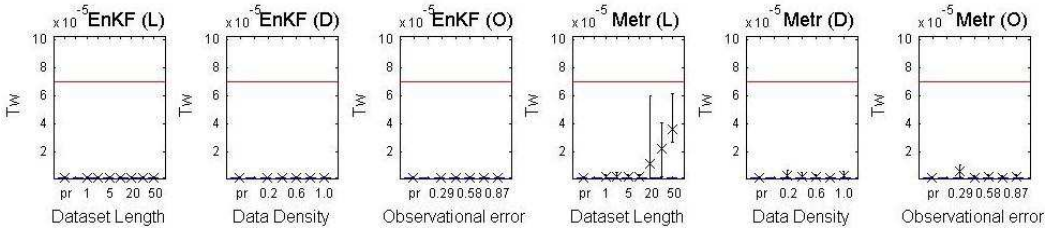
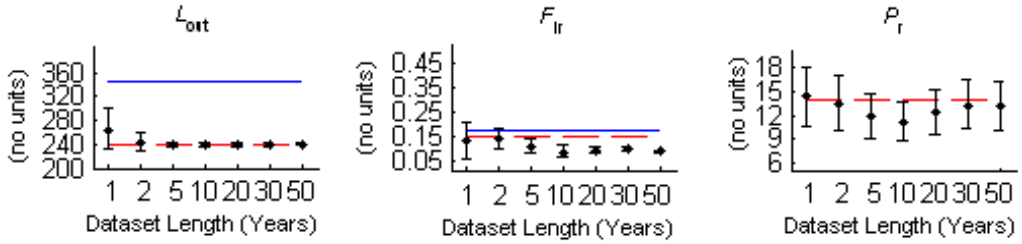


Figure 4.9 The parameter estimates for all 15 runs and for both DA schemes for the parameter Tw, using the actual upper bound from table 4.2. As with figures 4.6 – 4.8, the crosses and error denote the modal value and 5th to 95th percentile interval of the posterior distribution, while the red line is the true value of the parameter.



(a) (b) (c)

Figure 4.10: The effect of dataset length on parameter estimation for: (a) L_{out} (Growing degree day for leaf out), (b) F_{lr} (the fraction of labile carbon respired), and (c) P_r (Nitrogen use efficiency parameter for photosynthesis). (a) and (b) are estimates from the Metropolis algorithm, while (c) is an estimate from the EnKF (based on the mean of the distribution ensembles on the final time-step). For each dataset length, the dot represents the mean of the posterior distribution, while the error bounds represent the corresponding 5th and 95th percentiles (referred to as the 90th posterior interval). The true parameter values are indicated by red horizontal lines. Each of the Metropolis algorithm, the prior mean is represented by the blue line.

4.3.4 Accuracy of estimates of the parameters and final stocks

Here, we assess the accuracy of the parameter and stock estimates by two different ways. In the first we use compute the scaled absolute difference between the modal value of the posterior distribution of each parameter and the corresponding true value. Using this distance metric makes the results interpretable, but a downside is that the posterior intervals are not incorporated into the calculations, nor are correlations between the parameters. However, the posterior uncertainties are taken into account in subsection 4.3.5 when we consider the percentage of intervals that contain the true value. To overcome these shortcomings of this metric, an alternative one is also used to assess the distance of the parameters to the true values. An easily applicable one is the Mahalanobis distance metric. A noteworthy advantage of this is that the distances

(a)	T_d	F_g	F_{nf}	F_{nrr}	T_f	T_w	T_r	T_l	T_s	E_t	P_r	L_{out}	L_{fall}	F_{ll}	T_{lab}	F_{lr}	C_{fmax}
Length																	
Data density																	
Data noise																	

(b)	T_d	F_g	F_{nf}	F_{nrr}	T_f	T_w	T_r	T_l	T_s	E_t	P_r	L_{out}	L_{fall}	F_{ll}	T_{lab}	F_{lr}	C_{fmax}
Length																	
Data density																	
Data noise																	

(c)	$C_f^{(0)}$	$C_w^{(0)}$	$C_r^{(0)}$	$C_{lab}^{(0)}$	$C_{;it}^{(0)}$	$C_{som}^{(0)}$
Length						
Data density						
Data noise						

Table 4.3 Behaviour of the parameter estimates for (a) the EnKF, (b) the Metropolis algorithm, and for (c) the initial C pools for the Metropolis algorithm. The exact descriptions of the parameters are displayed in table 4.1.

(a)	C_f	C_w	C_r	C_{lab}	C_{lit}	C_{som}
Length						
Data density						
Data noise						

(b)	C_f	C_w	C_r	C_{lab}	C_{lit}	C_{som}
Length						
Data density						
Data noise						

Table 4.4 Behaviour of the final stock estimates for: (a) the EnKF, (b) the Metropolis algorithm.

are treated as vectors rather than point estimates. Also correlations between the parameters are taken into account. A downside to this metric is that its actual value is difficult to interpret. We use both distance metrics when assessing the accuracy of the parameters.

USING A SCALED ABSOLUTE AVERAGE DIFFERENCE METRIC

The accuracy of each parameter and final stock estimate, referred to as *A*, was measured by computing the average absolute difference between the parameter estimates and their corresponding true values, normalised by the prior range,

$$A = \frac{d}{R} \times 100 \tag{4.6}$$

where d is the absolute difference between the parameter estimate (the X in figure 4.11) and the corresponding true value (red line), and R is the difference between the smallest and largest possible values the parameter can take, which are listed in table 4.1. For example, for the Metropolis algorithm estimate of the F_{ll} parameter (Fraction of leaf loss transferred to litter) corresponding to the 1-year run (figure 4.11), this gave:

$$A = \frac{d}{R} \times 100 = \frac{|0.62 - 0.48|}{0.7 - 0.1} \times 100 = 23.3\%$$

Formula 4.6 was used to calculate the percentage difference for all parameters, since all had different units and radically different sizes. The average was then taken, so that comparisons between the 15 different runs could be made (e.g. as dataset length increases, are the parameter estimates on average getting closer to the truth?). For example, for the Metropolis algorithm, the run corresponding to the 1 year dataset (i.e. run 1 from table 4.2) had an average of 17%. This is displayed in the 2nd row of the 1st column of table 4.5a. The remaining percentages in table 4.5a are calculated similarly.

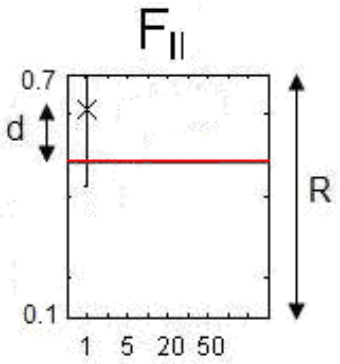


Figure 4.11 Plot of the Metropolis estimate of the F_{ll} parameter (Fraction of leaf loss transferred to litter). X = parameter estimate; Red line = truth; d = absolute difference between parameter estimate and the truth, for the 1 year run; R = range of possible parameter values.

One can see from table 4.5a that for all the runs, the EnKF parameter estimates are on average closer to the truth than those from the Metropolis algorithm. The overall average percentage difference between the DALEC-D parameters and the true values is around 11% for the EnKF, and around 17% for the Metropolis algorithm. The average percentage differences are also computed for the final stock estimates, where the truth refers to the true final stock value (table 4.5b). The EnKF again gave estimates closer to the truth for the most part, since all but three of the percentages in the EnKF rows were smaller than the corresponding percentages in the Metropolis algorithm rows. The overall average percentage differences between the final stock estimates and the true values are 5% for the EnKF and 9% for the Metropolis algorithm.

(i) Dataset Length

L (Years)	1	2	5	10	20	30	50	Average
EnKF (%)	17	13	13	11	10	9	11	12.0
Metropolis algorithm (%)	17	13	13	15	18	21	26	17.6

(ii) Data density

R_g	0.2	0.4	0.6	0.8	1	Average
EnKF (%)	11	9	11	11	7	9.8
Metropolis algorithm (%)	9	17	15	15	20	15.2

(iii) Dataset Noise

σ (gC m^{-2})	0.29	0.43	0.58	0.72	0.87	Average
EnKF (%)	9	10	11	11	10	10.2
Metropolis algorithm (%)	19	20	15	15	16	17.0

Table 4.5a The average percentage difference between the parameter estimates and the true parameter values.

(i) Dataset Length

L (Years)	1	2	5	10	20	30	50	Average
EnKF (%)	10	9	6	4	4	6	12	7.3
Metropolis algorithm (%)	12	8	7	7	13	11	14	10.3

(ii) Data density

R_g	0.2	0.4	0.6	0.8	1	Average
EnKF (%)	3	3	4	3	3	3.2
Metropolis algorithm (%)	4	8	7	7	11	7.4

(iii) Dataset Noise

σ (gC m^{-2})	0.29	0.43	0.58	0.72	0.87	Average
EnKF (%)	3	3	4	4	3	3.4
Metropolis algorithm (%)	10	13	7	7	7	8.8

Table 4.5b The average percentage difference between the final stock estimates and the true final stocks.

USING THE MAHALANOBIS DISTANCE METRIC

An alternative method for determining the accuracy of each parameter and final stock estimate is to compute the Mahalanobis distance (Mahalanobis, 1930) between a sample of parameter sets from the posterior distribution and the true parameter set. The metric is defined by the following formula (Mahalanobis, 1930):

$$D_{(i)}^2 = \left(p^{(i)} - p^{(t)} \right)^T C^{-1} \left(p^{(i)} - p^{(t)} \right)$$

Where $p^{(i)}$ is the *i*th sampled normalised parameter set from the posterior distribution, $p^{(t)}$ is the true parameter set also normalised, and C is the covariance matrix of the parameters. This results in N distances, i.e. $D_{(1)}$, $D_{(2)}$, …, $D_{(i)}$, …, $D_{(N)}$, where N is

(i) Dataset Length

L (Years)	1	2	5	10	20	30	50	Average
EnKF	34	28	36	36	39	39	38	**36**
Metropolis algorithm	45	78	204	216	241	311	413	**217**

(ii) Data density

R_g	0.2	0.4	0.6	0.8	1	Average
EnKF	32	36	36	37	38	**36**
Metropolis algorithm	144	212	216	285	265	**223**

(iii) Dataset Noise

σ (gC m^{-2})	0.29	0.43	0.58	0.72	0.87	Average
EnKF	38	38	36	35	34	**36**
Metropolis algorithm	437	381	216	173	139	**279**

Table 4.6a The Mahalanobis distances between the posterior parameter sets and the set of true parameter values.

(i) Dataset Length

L (Years)	1	2	5	10	20	30	50	Average
EnKF (%)	7	7	8	8	8	8	9	**8**
Metropolis algorithm (%)	5	6	7	18	32	32	64	**24**

(ii) Data density

R_g	0.2	0.4	0.6	0.8	1	Average
EnKF (%)	6	8	8	8	8	**8**
Metropolis algorithm (%)	5	11	18	19	30	**17**

(iii) Dataset Noise

σ (gC m^{-2})	0.29	0.43	0.58	0.72	0.87	Average
EnKF (%)	11	8	8	7	6	**8**
Metropolis algorithm (%)	38	28	18	14	14	**23**

Table 4.6b The Mahalanobis distances between the sets of final time-step C stocks (determined from posterior parameter sets) and the set of true parameter values.

the number of sampled parameter sets from the posterior distribution. In this chapter, *N* was set to 1000 for the Metropolis algorithm which was large enough to gain a good enough representative of the distribution of Mahalanobis distances between the parameter sets of the posterior distribution and the true parameter set. For the EnKF, *N* was fixed at 500 because this was the number of ensembles used. The median of this distribution of Mahalanobis distances was then calculated. This process was repeated for all 15 runs and for the two DA schemes and the results are displayed in tables 4.6a and 4.6b. We can see from tables 4.6a and 4.6b that overall the Mahalanobis distances between the parameter / stock vectors and the corresponding true vector remains relatively unchanged amongst the EnKF runs. However for the

Metropolis algorithm, we see that distance increases with increased information. The increase is much more than we see in tables 4.5a and 4.5b. For example as the length of the dataset increases from 1 to 50 years the Mahalanobis distance increases almost ten-fold from 45 to 413.

4.3.5 Consistency of the 90% posterior intervals with the truth

The percentage of parameters and final stocks whose 90% posterior interval contains the true value is displayed in tables 4.7 and 4.8. For example, the first number in the first row of the first column of table 4.7a means that 76% of the EnKF parameters for the 1-year run had 90% posterior intervals containing the truth (the higher the percentage, the more consistent the 90% posterior intervals are with the truth). For the parameters, the EnKF percentages were higher than those for the Metropolis algorithm in all but 4 cases. Taking the overall average, 88% of all the parameter estimates from the EnKF had 90% posterior intervals which contained the truth, as opposed to 40% for the Metropolis algorithm (or 46% if averaging over all 23 parameters). The Metropolis algorithm was better at estimating the initial C pools (3rd rows of table 4.7) than the parameters (2nd rows of table 4.7), with the initial C pools' percentages having an overall average of 62%. For the final stock estimates, the EnKF (rows 1 of table 4.8) was better than the Metropolis algorithm (rows 2 of table 4.8) in all but four cases. Taking the overall average, 87% of all the final stock estimates for the EnKF had 90% posterior intervals which contained the truth, as opposed to 53% for the Metropolis algorithm.

The reason why the percentages in tables 4.7 and 4.8 are a lot smaller for the Metropolis algorithm is that the posterior intervals are a lot narrower than for the EnKF. Understanding why they are narrower is discussed in section 4.4.

4.3.6 Comparing the sizes of the prior and posterior intervals

A key aspect of Data Assimilation is that the post-assimilation estimates of the quantities we are interested in have a smaller uncertainty that pre-assimilation. In the case of parameter estimation, it is essential that the posterior parameter estimates have a small uncertainty than those of the prior estimates. However when we use real data, it is difficult to know whether the resulting posterior uncertainty interval contains the

(a) Dataset Length

L (Years)	1	2	5	10	20	30	50	Average
EnKF (%)	76	76	82	82	94	88	88	**84**
Metropolis algorithm (%)	82	76	76	35	24	18	18	**47**
Metropolis initial C pools (%)	67	67	83	67	50	83	17	**62**

(b) Data density

R_g	0.2	0.4	0.6	0.8	1.0	Average
EnKF (%)	88	94	82	94	94	**91**
Metropolis (%)	88	35	35	35	18	**42**
Metropolis initial C pools (%)	83	50	67	67	67	**67**

(c) Dataset Noise

σ (gC m^{-2})	0.29	0.43	0.58	0.72	0.87	Average
EnKF (%)	88	88	82	94	94	**89**
Metropolis algorithm (%)	18	24	35	35	41	**31**
Metropolis initial C pools (%)	17	50	67	67	83	**57**

Table 4.7 The percentage of the 17 DALEC-D parameter estimates from the Metropolis and EnKF inversions that encompass the true parameter values within the 5th and 95th percentiles. The Metropolis algorithm treats the six initial carbon pools as parameters; these are shown with separate percentages.

(a) Dataset Length

L (Years)	1	2	5	10	20	30	50	Average
EnKF (%)	50	67	83	83	83	67	67	**71**
Metropolis (%)	83	100	100	50	17	17	17	**55**

(b) Data density

R_g	0.2	0.4	0.6	0.8	1	Average
EnKF (%)	100	100	83	100	100	**97**
Metropolis (%)	100	67	50	50	33	**60**

(c) Dataset Noise

σ (gC m^{-2})	0.29	0.43	0.58	0.72	0.87	Average
EnKF (%)	100	100	83	100	100	**97**
Metropolis (%)	33	50	50	50	33	**43**

Table 4.8 The percentage of final stock estimates from the Metropolis and EnKF inversions that encompass the 'true' final C stock values within the 5th and 95th percentiles.

true value (e.g. Richardson et al., 2010). By using synthetic data, we can check that the resulting reduction of uncertainty from carrying out DA does not result in us being over-confident with the parameter estimate. With this in mind, we deem a reduction in uncertainty post-DA credit worthy as long as this post-DA uncertainty interval still contains the truth.

For the results here a reduction in uncertainty, in some cases significant, from the prior to the posterior estimates was recorded for virtually all of the EnKF parameter estimates (see figures 4.6a, 4.7a and 4.8a). For the Metropolis algorithm, we can see from table 4.7 that only the runs which had the 1, 2 or 5 year datasets or had a data density of 0.2 (runs 1, 2, 3 and 8 from table 4.2) had the posterior uncertainty intervals for most of the parameters containing the truth. For these four runs, we see a reduction in uncertainty for the majority of the parameters whose posterior interval contained the truth.

These findings are discussed in subsection 4.4.5.

4.3.7 Parameter correlations

The correlations between the parameters in the posterior distribution was determined using 50,000 sampled parameter sets for the Metropolis algorithm and the 500 parameter sets from the 500 ensemble members. The correlations were highest amongst the Metropolis algorithm retrieved parameters, and the runs (see table 4.2) could be arranged into 4 groups according to level of correlation. Group 1 consists of runs 3 and 12 and showed the greatest correlations (panel (a) from figure 4.12) with an average correlation (in absolute terms) of around 0.15 but a 5^{th} to 95^{th} percentile range of around 0.7. Group 2 consists of by runs 4, 8 and 14 (panel (b) of figure 4.12) and the next highest correlations with a median of around 0.11 and a 5^{th} to 95^{th} percentile range of 0.6. Group 3 consisted of runs 6, 9, 10, 11, and 13 with the second lowest parameter correlations of around 0.08 on average and a 5^{th} to 95^{th} percentile range of 0.45 (panel (c) of figure 4.12). The remaining five runs made up group 4 (panel (d) of figure 4.12) and had the lower parameter correlations of around 0.06 on average and a 5^{th} to 95^{th} percentile range of roughly 0.4.

The correlations between the parameters from the posterior distribution of the EnKF runs were all very similar, with each run having a median absolute correlation of around 0.05 and a 5^{th} to 95^{th} percentile range of between 0.2 and 0.3. The spatial representation of the correlation matrix for each of the EnKF runs was therefore similar to panel (d) from figure 4.12.

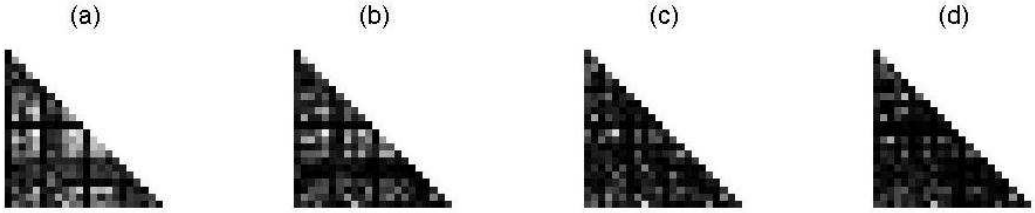


Figure 4.12 A graphical representation of the correlations between the parameters in the posterior distribution of the Metropolis algorithm. Panel (a) to (d) represent the parameter correlations from runs 3, 8, 11 and 15, which represent the four groups of runs which varied in the level of parameter correlations. Each box represents the correlations between two different parameters with the lighter and darker ones denoting the higher and lower correlations, respectively. The order of the boxes, both vertically and horizontally is the same as the order of the parameters in table 4.2.

4.3.8 Accuracy of the estimates of the iNEE forecasts

The parameter and final stock estimates from the Metropolis algorithm and EnKF for all 15 runs (table 4.2) were used to predict iNEE (cumulative NEE) 10 years into the future. To incorporate uncertainty into the forecasts, all 500 of the ensembles were used for the EnKF, and 500 sets of parameters and corresponding final stock estimates were selected at random from the posterior distribution for the Metropolis algorithm. Since the EnKF gave parameter and final stock estimates closer to the truth, and 90% posterior estimates which contained the truth more frequently, it might be expected that the EnKF would outperform the Metropolis algorithm in forecasting iNEE. This was not the case, as can be seen in figure 4.13. In this figure we can see that, with the exception of the 1-year run, the EnKF estimates are only marginally closer to the truth than the Metropolis ones, for all the runs. The root mean square error (RMSE) of the 15 EnKF forecast estimates is 647 gCm^{-2} (or 223 gCm^{-2} if run 1 is not included) whereas for the Metropolis algorithm estimate it is 773 gCm^{-2} (or 310 gCm^{-2} if run 1 is not included). Here the RMSE is given by:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{15}\left(f^{(i)} - f^{(t)}\right)^2}{15}}$$

with $f^{(i)}$ denoting the forecast estimate for the *i*th run (table 4.2), and $f^{(t)}$ is the true forecast value represented by the red line in figure 4.13. An alternative way of measuring the accuracy of the forecasts is to compute the RMSE of the difference between the NEE trajectory produced by each of the 500 parameter sets used in the forecast and the true NEE trajectory. This is shown in figure 4.14. Similar to the iNEE

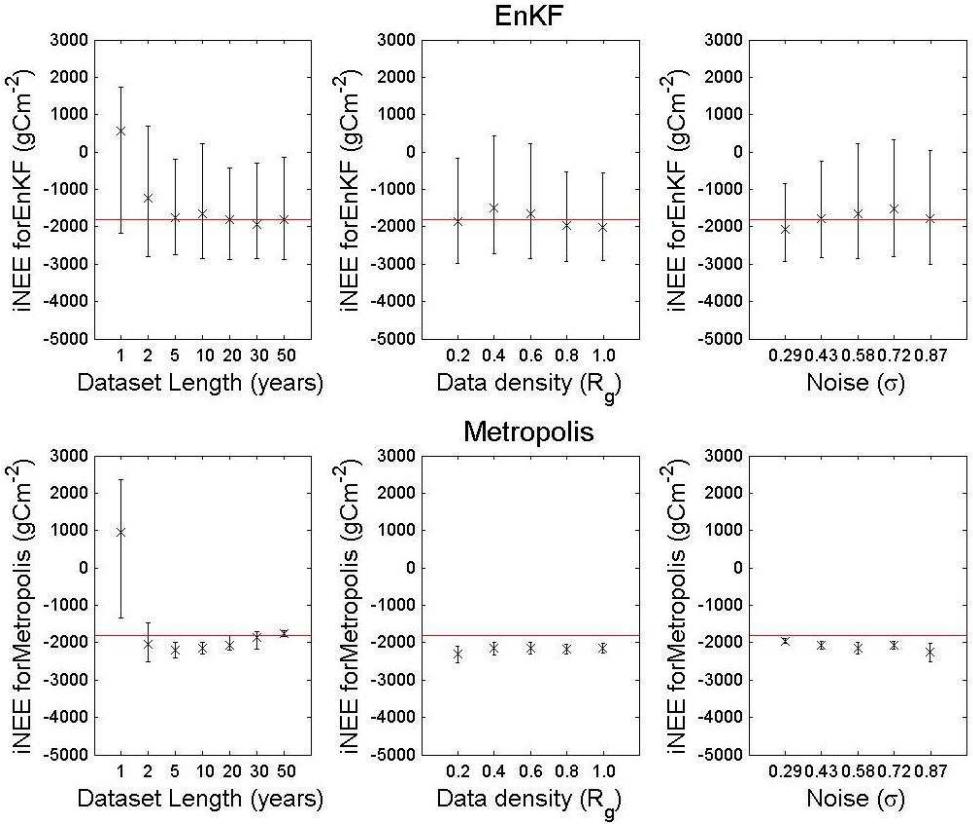


Figure 4.13 The predictions of cumulative NEE after 10 years for the EnKF and the Metropolis algorithm. The red horizontal line shows the truth.

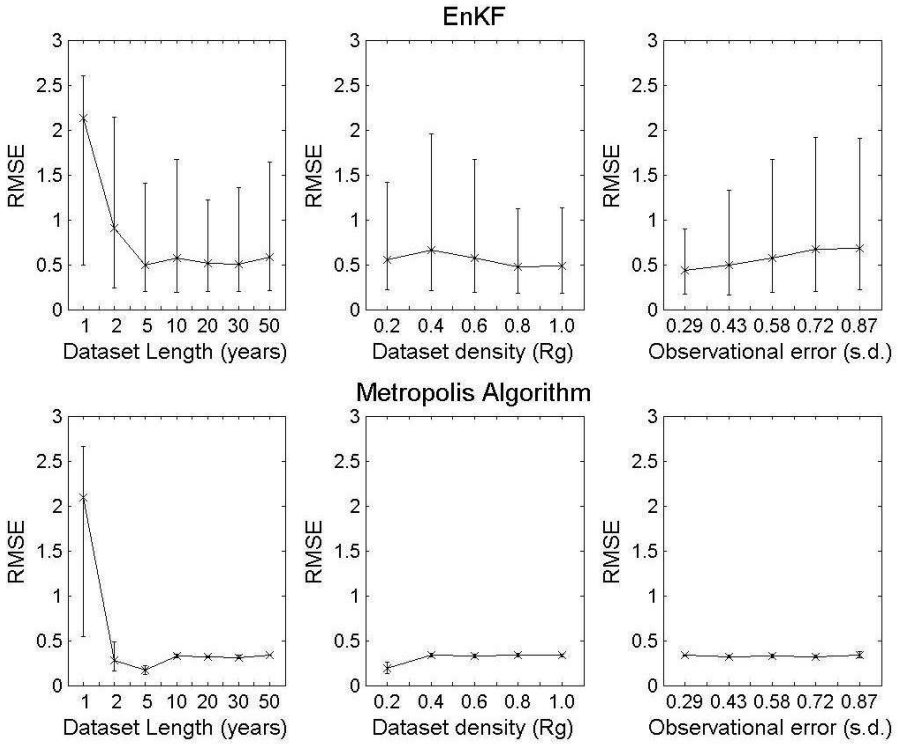


Figure 4.14 The RMSE (root mean squared error) of the predictions of the NEE trajectory after 10 years.

plots shown in figure 4.13, one can see that there is a sharp improvement in the accuracy of the forecasts between 1 and 2 years. This is probably because, in order to get the NEE trajectory of the forecast period to be close to the truth, you need at least 2 years of assimilated NEE observations in order to model the interannual variability. The 5^{th}-95^{th} percentile intervals are much wider for the EnKF than for the Metropolis algorithm. Conversely, the Metropolis algorithm's posterior intervals appear unrealistically narrow in terms of representing real world uncertainty and in fact very few contain the truth (though most are close); this is discussed further in section 4.4. In terms of trends, increasing the length of the dataset improves the estimates for both the Metropolis algorithm and EnKF. In addition, the Metropolis algorithm estimates of iNEE improve with a reduction in observation noise and with an increase in the density of observations. For the EnKF, the RMSE values reduce and/or percentile intervals get narrower while still containing the truth with a reduction in observational noise and with an increase in the density of observations; however this improvement is much less than the effect of increasing dataset length for both DA schemes.

4.3.9 Testing the robustness of the forecasts

In section 4.4, we discuss the possible reasons into how the Metropolis algorithm in general had poor parameter estimates and uncertainties, but its resulting forecasts were very accurate with small uncertainties. It could be argued that this is occurring because the Metropolis algorithm operates by finding parameter sets that give good fits of the model to the data; since the data is synthetic and is in a pseudo steady state, once good fitting parameter sets have been found they are likely to give good fits to the resulting forecasts. A way of disrupting this pseudo steady state is to impose an artificial drought (as an example of a natural disturbance) which will serve as a means of determining the robustness of the estimated parameter sets in still giving accurate forecasts and small uncertainties.

The drought was created to be severe and was imposed in years 5 and 6 of the forecast period. Since DALEC-D has no water component to it, this was achieved by changing two of the parameters of the ACM (Aggregated Canopy Model), the sub-model of DALEC-D which estimates Gross Primary Production from the meteorological data, which were responsible for stomatal conductance. These two parameters were the soil-water leaf potential different (phi_d) and the total plant-soil

hydraulic resistance (R_{tot}). In DALEC and DALEC-D, these two parameters are treated as constants and fixed at phi_d = -2 and R_{tot} = 1. In the water version of DALEC (Grace, 2005), the water dynamics of the model exists to estimate phi_d and R_{tot} for each time-step (i.e. they are not constants). Therefore, a drought can be simulated by fixing these two constants at values which would result from low precipitation and very low soil water content (i.e. the environmental conditions of a drought). Williams (2010) [personal communication] states that an artificial drought can be simulated in DALEC by changing the constants mentioned above to phi_d = -0.5 and R_{tot} = 3. In real-life, this change corresponds to the stomata in the leaf closing to reduce the amount of water transpiring, i.e. to conserve the water available to the plant.

The drought imposed forecasts are shown in figure 4.15. Surprisingly, the RMSE between the drought-imposed forecasts and to the drought-imposed truth was 584 gC/m² for the Metropolis algorithm (figure 4.15, top panel), whereas the original forecasts had an RMSE of 773 gC/m². The EnKF's forecasts showed a similar improvement when the drought conditions were imposed, with an RMSE of 496 gC/m² (figure 4.15, bottom panel) compared to an RMSE of 647 gC/m² for the original forecasts. These findings are discussed in section 4.4.

4.3.10 Other results included as part of the discussion

In this final subsection of the chapter 4 results, we include additional results to the original experiments and results to other modelling exercises. These were done to aid the discussion in section 4.4.

LAG PLOTS

In subsection 4.2.10, it is suggested that a potential reason why the parameter estimates and uncertainties between the two DA schemes may not agree is because the residuals between the modelled NEE and the NEE data in the cost function for the Metropolis algorithm may be autocorrelated. The autocorrelation in residuals for all 15 runs was determined for different lags, and is graphically represented by lag plots. Figure 4.16 shows the lag plots for three of the runs and represents the variations in the lag plots among the runs. For four of the runs, namely 1, 2, 3 and 8, there was close to zero autocorrelation (e.g. run 2, left panel of figure 4.16). For 10 of the remain-

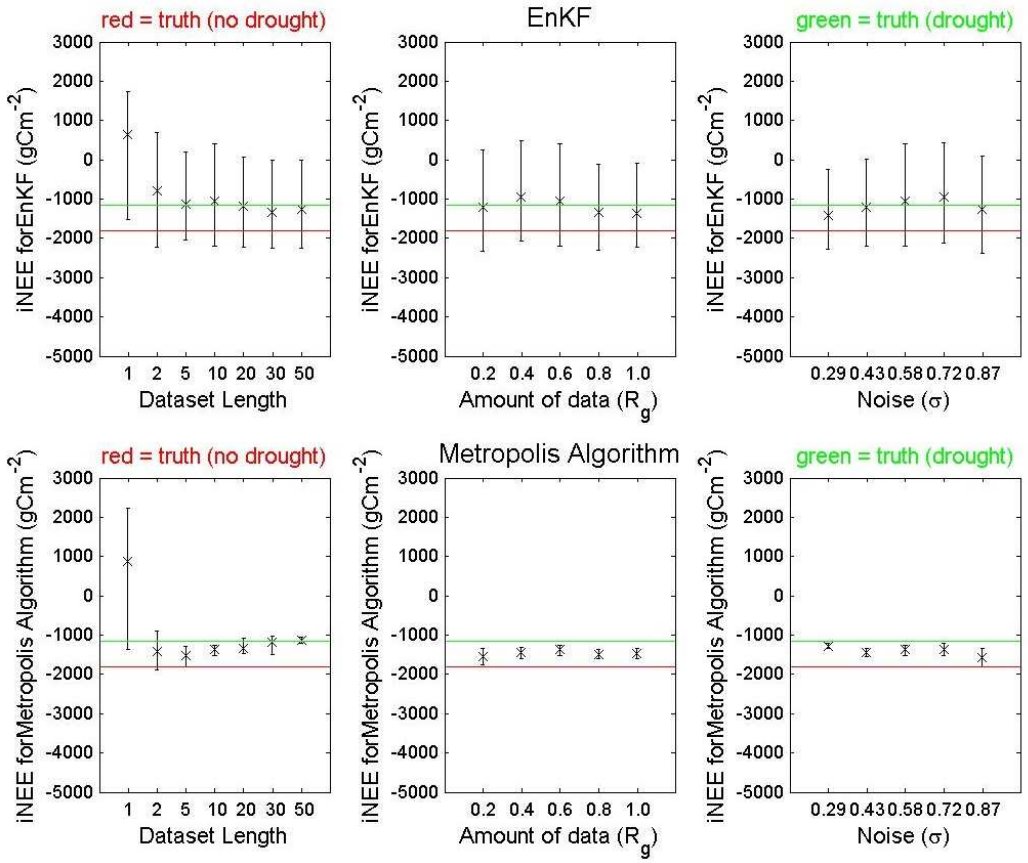


Figure 4.15 The forecast estimates and 5th – 95th percentile intervals of the EnKF and the Metropolis algorithm with a severe drought imposed in the central 2 years of the forecast period.

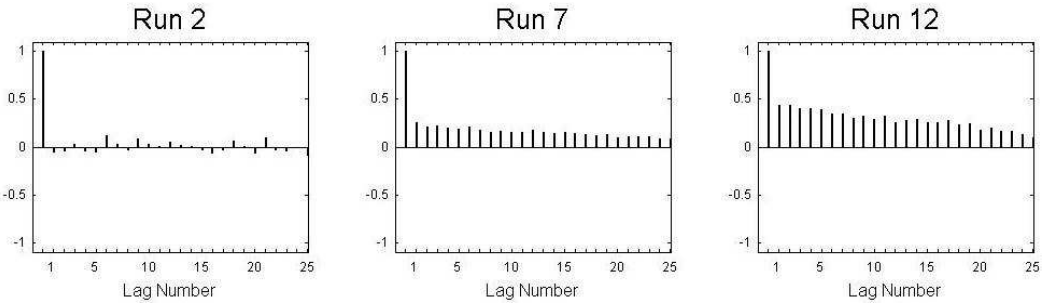


Figure 4.16 Plots showing the autocorrelation for lags going from 1 to 25 for runs 2 (left panel), 7 (central panel) and 12 (right panel).

ing 11 runs, there was consistently positive autocorrelation which started around 0.25 at lag 1 and very gradually dropped to between 0.15 and 0.2 after a lag of 25 (e.g. run 7, central panel of figure 4.16). For the final run, run 12 (right panel of figure 4.16), the autocorrelation was noticeably higher for all the lags starting at around 0.45 at lag 1 and dropping to around 0.2 at lag 20.

An interesting observation from these lag plots is that the four runs which showed very little autocorrelation in the residuals (i.e. runs 1, 2, 3 and 8) were the same ones where there was most agreement between the parameter estimates and uncertainties of the Metropolis algorithm and the EnKF. This is discussed further in section 4.4.

PLOTS OF THE RMSE BETWEEN THE MODELLED & OBSERVED/TRUE NEE

In addressing the hypotheses in the discussion section, one of the issues assessed was whether the fit of the post-DA modelled NEE to the NEE data improves with longer assimilated time-series, lower observational error and a greater density of data. This fit for each of the runs is determined by computing the root mean squared error (RMSE) of the differences between the modelled NEE and observed NEE. Thus the RMSE is computed for the difference between the modelled NEE to the true NEE. These plots are shown in figures 4.17 for the dataset length runs (runs 1 to 7), figure 4.18 for the dataset density runs (runs 8, 9, 4, 10, and 11) and figure 4.19 for the observational error runs (runs 12, 13, 4, 14 and 15).

What is interesting about these plots is that for figures 4.17 and 4.18, while the RMSE and the 5^{th} – 95^{th} percentile interval remain unchanged for the 'modelled vs observations' plots for the Metropolis algorithm (panel b) between the different dataset length and data density levels, we do not see this for the 'model vs true' plots (panel a) where the RMSE is noticeably less for runs 1, 2, 3 and 8, the only runs to not have autocorrelation in their cost function residuals. For figure 4.19b, unlike panel (b) from the previous two figures, the RMSE and 5th – 95th percentile interval changes for different observational error. In particular, it reduces as the observational error gets less. Extending the line joining the RMSE median values for the observational error runs (12, 13, 4, 14 and 15) as the observational error tends to zero, the RMSE appears to extend to a value between 0.3 and 0.4. This is confirmed when the RMSE and percentile interval was computed for an extra run where the observational error was 0.05 gCm^{-2}. One would expect the RMSE to tend to zero as the observational error tends to zero, so the fact that it does not is unusual. Given that the most of the runs which had autocorrelated residuals also had poor Metropolis algorithm parameter estimates, it is reasonable to assume that this unusual behaviour noted above about figure 4.19b is due the autocorrelated residuals.

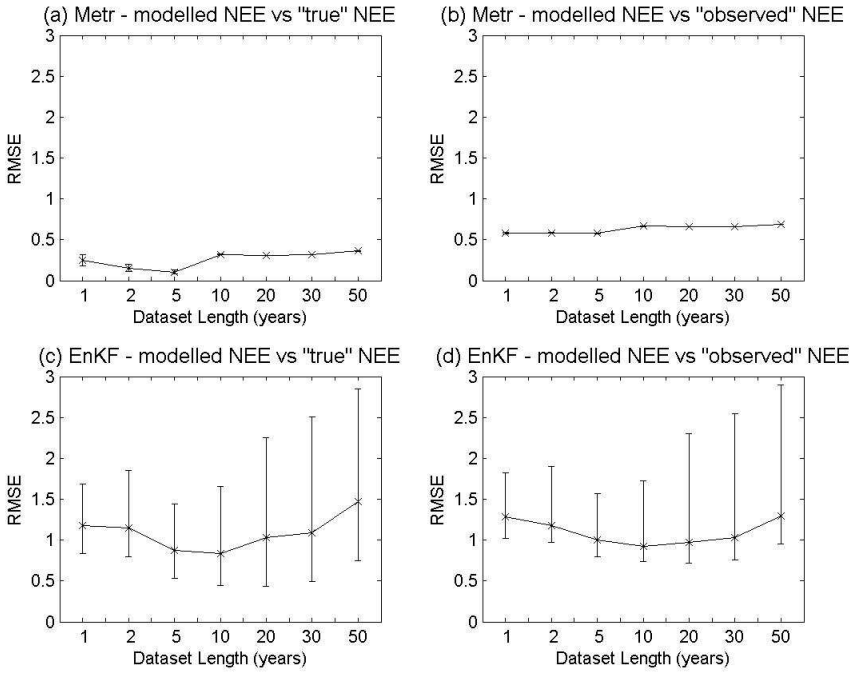


Figure 4.17 The mean, 5th and 95th percentile values of the RMSE between the modelled NEE and the true NEE [for (a) & (c)], and, the RMSE between the modelled NEE and observed NEE [for (b) and (d)] during the DA period corresponding to 500 randomly chosen parameter sets from the posterior distribution (for the Metropolis algorithm) or the parameter sets from all 500 ensembles (for the EnKF). These were calculated for the runs where dataset length varied.

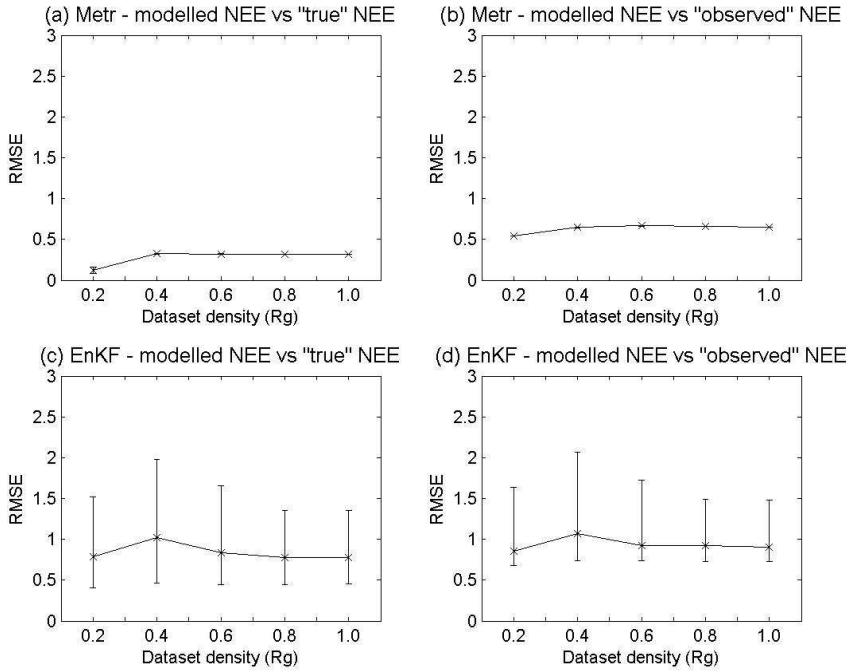


Figure 4.18 The same as figure 4.17 except the plots correspond to where data density varied (i.e. runs 8, 9, 4, 10, 11).

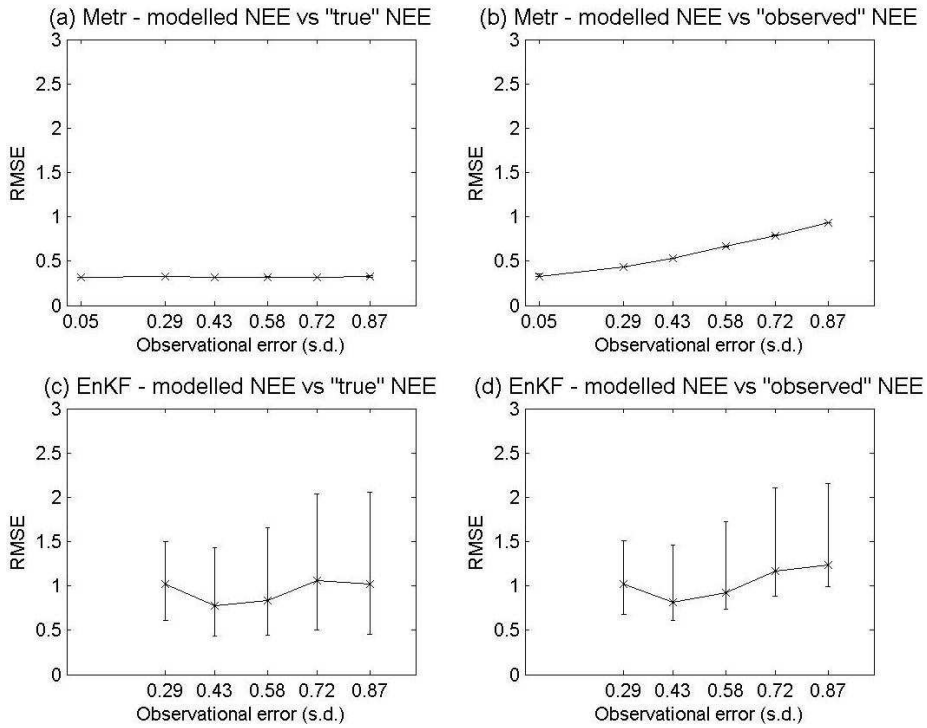


Figure 4.19 The same as figure 4.17 except the plots correspond to where observational error varied (i.e. runs 12, 13, 4, 14, 15). An additional run of the Metropolis algorithm was carried out with the same length and data density as these 5 runs but the observational error was sd=0.05gCm^{-2}.

ENTROPY

One way to assess the contribution of the observations is to determine how the entropy has changed from the prior to the posterior distribution of the parameters. This change in entropy is called Mutual Information (Rodgers, 2000). As entropy is a measure of uncertainty, for a Normal distribution it is connected to the variance in 1 dimension or the error covariance matrix otherwise. Eyre (1990) defines entropy and Mutual Information (MI) as:

$$\text{Entropy} = H(Z) = \int p(Z)\ln(Z)\,dZ = nln(2\pi e)^{1/2} + \frac{1}{2}\ln|\boldsymbol{C_z}|$$

$$\text{MI} = H(X) - H(X|Y) = \frac{1}{2}\ln|\boldsymbol{B}\boldsymbol{P}_a^{-1}|$$

where $\boldsymbol{C_z}$ is the error covariance matrix of a general multivariate random variable Z of dimension n, while matrices $\boldsymbol{B}$ and $\boldsymbol{P}_a$ are the error covariance matrices corresponding to the prior and posterior distributions X and X|Y. In the context of this chapter, X and Y are random variables corresponding to the prior parameter distribution and data respectively. Due to dataset length having most impact on the parameter estimates and due to the issues surrounding the Metropolis algorithms

affecting some of the runs, MI was computed for runs 1 to 7 for the parameters and distributions estimated using the EnKF only. As we are predominantly interested here in quantifying the impact of the observations on the parameters and their corresponding distributions, the entropy calculated here for the prior and posterior distributions, namely $H(X)$ and $H(X|Y)$, the augmented state vectors are shrunk to only include the parameter components.

A practical interpretation of entropy is that it gives a single value for the spread of the distribution, which is very useful for multivariate distributions where the spread of the distribution is quantified by the error covariance matrix which can be very large. Therefore the MI quantifies, again as a single number, the change in the spread of the posterior distribution relative to that of the prior distribution. The second column of table 4.9 below show MI for runs 1 to 7 which correspond to the runs where the dataset length varied from 1 year to 50 years. We can see that there is a marked increase in MI for runs 5 to 7 compared to runs 1 to 4. Since the entropy for the prior distribution of the parameters is fixed, by the formula for MI given above this means that the posterior distribution is narrower than the prior distribution to a greater extent for runs 5 to 7 compared to runs 1 to 4. In other words, the effect of assimilating a longer dataset is to cause a narrowing of the posterior distribution of the parameter space.

The third column of table 4.9 shows how far the mean vector of the parameter components only of the augmented state vector of the posterior distribution has moved compared to the mean vector of the prior distribution. To be explicit, if X^f and X^a refer to the mean vector of the prior (or forecast) and posterior (or analysis) distributions, and **B** is the error covariance corresponding to the prior, then the values the third column of table 4.9 are calculated using the following formula:

$$(X^a - X^f)^T \boldsymbol{B}^{-1} (X^a - X^f)$$

With the exception of the 20 year run, we can see in the third column of table 4.7 that as the length of the dataset increases, the mean vector of the parameter elements of the augmented state vector from the posterior distribution move further away from the mean vector from the prior distribution. These issues are briefly discussed in subsection 4.4.6.

Run	Mutual Information	'Posterior' minus 'Prior' mean vector
1 (1 year length)	15.53	1.60
2 (2 year length)	16.48	2.19
3 (5 year length)	17.34	3.02
4 (10 year length)	16.18	4.87
5 (20 year length)	21.64	0.29
6 (30 year length)	20.71	3.69
7 (50 year length)	19.58	4.82

Table 4.9 The mutual information of the prior and posterior parameter distributions (second column) and the location of the mean vector of the posterior parameter distribution standardised against the prior distribution (third column) for runs 1 to 7 (where dataset length varies).

4.4 Discussion

4.4.1 Summary of results and layout of discussion section

SUMMARY OF RESULTS

The length of the dataset is much more important than the density of the observations or the level of observational error when estimating the DALEC-D parameters based on NEE observations (subsection 4.3.1). Datasets of between 5 and 10 years in length are required in order to get good parameter estimates for the EnKF, and good forecast estimates for both DA schemes.

The REFLEX project found that different DA techniques can produce markedly different parameter estimates and uncertainties even when the same dataset and the same model were used (Fox et al., 2009). The results here showed similar behaviour. The EnKF produced markedly more accurate estimates of the parameters and final C stocks than the Metropolis algorithm. In addition, 88% of all the DALEC-D parameter estimates and 83% of all the 6 final C stock estimates from the EnKF had 90% posterior intervals which contained the truth. For the Metropolis algorithm, the corresponding percentages were 40% / 46% (if averaging over 17 / 23 parameters) and 53%. For the EnKF, the parameter and final C stock estimates were closer to the truth for longer datasets, smaller observational error and a greater density of observations,

though the effect of dataset length was greater than the other two factors. In contrast for the Metropolis algorithm, the parameter and final C stock estimates were in general closer to the truth for shorter datasets, larger observational error and a smaller density of observations (tables 4.3 and 4.4). When the sets of estimated parameter and final C stocks were used to forecast 10 years of cumulative NEE, the Metropolis algorithm produced narrower 5^{th}-95^{th} percentile intervals than the EnKF, though the estimates were similar (figure 4.13). For both DA schemes, the effect of dataset length on the forecasts was much greater than the other two factors.

LAYOUT OF DISCCUSION SECTION

The remainder of this section contains three parts: First of all, in subsections 4.4.2 – 4.4.4 we discuss why the EnKF was better at estimating parameters while the Metropolis algorithm gave more confident forecasts – particular focus is on trying to explain why the parameter estimates from the two DA schemes were so different for most of the 15 runs. Subsection 4.4.2 identifies, that for the runs where there are differences in the results between the two DA schemes, the global minimum is not being reached for the Metropolis algorithm runs; possible reasons why this is happening are also included. Subsections 4.4.3 and 4.4.4 then outline two key implications for not reaching the global minimum for these runs, in particular parameter equifinality and the autocorrelation of the residuals in the cost function.

In the second part (subsections 4.4.5 and 4.4.6), we first compare our results with our expectations of them. We then consider more general issues surrounding the results, which include: (i) the significance of the unchanging posterior uncertainties for varying observational error; (ii) whether we see a reduction in the posterior uncertainties on the parameter estimates compared to the prior uncertainties; (ii) trying to explain the differences in the posterior uncertainties on the parameter estimates between the two DA schemes; (iv) determining the robustness of the forecasts, particularly those from the Metropolis algorithm; (v) assessing the entropy of the results.

Finally the third part (subsection 4.4.7) addresses the hypotheses in terms of whether these are likely to be correct or not based on the evidence from the results.

4.4.2 The reason for the differences in parameter estimates and uncertainties between the two DA schemes

SITUATIONS WHERE THE TWO DA SCHEMES GIVE SIMILAR RESULTS

Before try to explain the differences, it is first worthwhile outlining where there are similarities in the results. For runs 1, 2, 3 (which correspond to the dataset length being 1 year, 2 years and 5 years) and run 8 (where the dataset density was 0.2) we find that the Metropolis algorithm gives very similar results to the EnKF. Table 4.10 summarises this agreement among the two DA schemes for these four runs (see figures 4.6a and 4.7a) and also for the REFLEX (figure 2.11) where the same two DA schemes were used. Of the 17 DALEC-D parameters, there are between 10 and 14 parameters whose posterior estimates agreed well among the two schemes, and between 14 and 15 parameters where there was good agreement with the width of the uncertainty intervals. The dataset used in REFLEX is most similar to the one from run 2, as it had similar data density and observational error characteristics and was on length two years. In REFLEX we can see that there are 13 parameters where there is good agreement in parameter estimates amongst the two DA schemes, and 13 agree well in terms of the uncertainty width. We would not necessarily expect there to be exactly the same numbers of parameters where there was agreement between the parameter estimates and uncertainties since the data used was based on a different site to what is used here. However, we nonetheless see a similar number of parameters agreeing to those from run 2 in particular, which is encouraging.

Furthermore for runs 1 to 3 (where dataset length increased from 1 year to 2 years to 5 years) we see from the bottom panel of figure 4.6a more of a consistent pattern of the parameter estimates moving closer to the true value as the dataset length increases, as was seen with the EnKF runs (table 4.4a), compared to when all of the dataset length runs are considered. Furthermore, runs 1, 2, 3 and 8 have the highest percentages (76% - 88%) of parameters where the posterior uncertainty interval contains the true values compared to the remaining 11 runs where the percentages range from 18% to 41%.

The question therefore remains as to why these four runs result in more consistent parameter estimates and uncertainties between the two DA schemes, are more closely matched to what is hypothesised, and are more consistent with the true parameter values. The likely reason is that for these four runs, there is close to zero

	Parameter estimates			Uncertainty interval width	
	Good agreement	Fairly good agreement		Good agreement	Fairly good agreement
Run 1	10	6		14	3
Run 2	12	4		15	2
Run 3	13	3		14	1
Run 8	14	2		14	1
REFLEX	13	3		13	4

Table 4.10 The numbers of parameters where there is good or fairly good agreement of the parameter estimates and widths of the uncertainty intervals between the Metropolis algorithm's parameter estimates and those determined from the EnKF.

autocorrelation in the residuals of the cost function used for the Metropolis algorithm as can be seen in the lag plot for run 2 as an example (left panel of figure 4.16). In contrast, the autocorrelation is significantly higher and remains the same sign for the different lags of the remaining runs as can be seen in the lag plots for runs 7 and 12 (central and right panel of figure 4.16). This is discussed in greater detail in subsection 4.4.4.

EVIDENCE FOR THE REMAINING 11 METROPOLIS ALGORITHM RUNS NOT AT THE GLOBAL MINIMUM

The most obvious reason why the Metropolis algorithm's parameter estimates for the 11 'bad' runs were mostly far from the true values may have been because the chains in each of these runs are not at the global minimum of the parameter space which the algorithm was searching. Evidence for this can be found by determining the value of the cost function (the mismatch between the modelled and observed NEE) for a sample of parameter sets at the minimum where the algorithm converged to and also the value of the cost function corresponding to the true parameter set; this is shown in figure 4.20.

We can see from figure 4.20 that for runs 1, 2 and 3, the mode of the distribution of posterior cost function values was lower than that of true parameter set (left plot of figure 4.20). This is extremely interesting because it means that the true parameter set is not at the global minimum (although it does lie within the 5^{th}-95^{th} percentile range).This suggests that the model or the algorithm is sensitive to the noisy

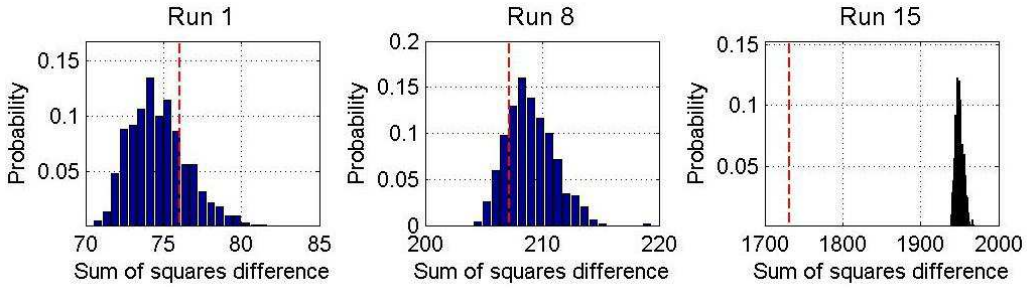


Figure 4.20 The distribution of sum of squared difference values between the modelled NEE and the observed NEE for runs 1, 8 and 15 of the Metropolis algorithm (table 4.2). For each run, the distribution was constructed from 500 sum of squared difference values, where the modelled NEE was obtained from 500 randomly selected parameter sets from the posterior distribution. The vertical red dotted line is the sum of squared difference value between the true NEE (i.e. the NEE trajectory obtained from a forward run of DALEC-D using the true parameter set).

observations, i.e. the problem is ill-conditioned. By this, we mean that although the algorithm has converged to the global minimum, this is not where the true parameter set is. This means that it is impossible for the modal value of the posterior distribution to be the same as the true parameter values. This is probably caused by the data being too noisy and too sparse thus extracting the 'true NEE' is impossible, although we can get close since the true parameter set is at least in the posterior distribution. For one of the runs (8), the mode was close to the cost function value of the true parameter set (central plot of figure 4.20).

For the remaining 11 runs, the 'bad runs' where we observed the significant differences in the parameter results between the two DA schemes, the value of the cost function for the true parameter set was significantly less than any of the cost function values for the parameter sets in the posterior distribution (e.g. run 15, right panel of figure 4.20). This means that the global minimum is not being reached and so the Metropolis algorithm is not functioning as well as it should. This is very unlikely to be the result of any implementation issues as this has been thorough investigated and checked (see subsection 2.3.3 and 2.3.4). It is certainly not a convergence issue, as convergence was ensured by the Gelman criteria (subsection 4.3.2) and by plotting the cost function against the different dimensions of the posterior parameter space (subsection 4.4.2) where the we observed a flat line strongly indicating that the minimum had been reached. The only feasible explanation that has been explored in the results is that the residuals of the modelled NEE and the NEE data in the likelihood function are autocorrelated. This is discussed further in the

subsection 4.4.4. However, because of this unresolved issue with these 11 runs, it is difficult to draw any conclusions from them because of the consistent strange and poor behaviour resulting from them. Therefore, it is decided these 11 runs, for the Metropolis algorithm runs only, are to be removed from the conclusions to this chapter. This is discussed further in the subsection 4.4.4.

POSSIBLE REASONS WHY THE 11 METROPOLIS ALGORITHM RUNS ARE NOT AT THE GLOBAL MINIMUM

We next discuss four possible reasons for lack of convergence to the global minimum for these 11 runs from the Metropolis algorithm: (i) an implementation issue; (ii) the Metropolis algorithm is solving a slightly different optimization problem to the EnKF; (iii) the type of step-size used; (iv) the initialisation of the algorithm from the same locations rather than random locations.

A possible factor which may be causing or partly responsible for the divergence in estimates of the parameters and uncertainties of the Metropolis algorithm from the truth is that the algorithm has not been implemented correctly. There are two reasons why there is no evidence that this is true. First of all, if there was an implementation issue, we would expect the results from all 15 runs to be different and for the other runs not to reach the global minimum. The fact that the results are very similar between the two DA schemes and the Metropolis algorithm reaches the global minimum for 4 of the runs suggests that other reasons are causing the differences in the remaining runs (subsection 2.3.3). Secondly, a test run were carried out in chapter 2 using both DA algorithms. In this test run, the data and prior distribution were prescribed very low noise and for both DA schemes the vast majority of the estimated parameters were very close to the true values, with the posterior uncertainty narrow but containing the truth.

If two DA schemes are being used to estimate parameters and their distributions but the setups in each case is slightly different, it could be argued that the DA schemes are solving different optimization problems. We now show that this is not true. Both DA algorithms are based on Bayes' theorem, although they operate in different ways (i.e. the EnKF is sequential and the Metropolis algorithm is a batch method) and so under the same assumptions both should result in the same or approximately the same posterior distributions. In this chapter there are two minor

differences in the setup of the two algorithms. First of all the log-Normal distribution is used for the prior for the Metropolis algorithm in order to ensure the distribution does not cover non-physical parameter values such as negative values. It also means that all the log-Normalised parameters have the same mean and variance. In contrast the prior distribution used in the EnKF is Normal. Despite the different prior distribution for the model parameters, it is important to note that the log-Normalised prior parameters used in the Metropolis algorithm follow a Normal distribution and have the same prior variance, relative to the prior mean, to the Normal prior distribution used for the parameters in the EnKF. Therefore we are still in effect solving the same optimization problem. The other minor difference in the setup is that the Metropolis algorithm assumes no model error but the EnKF incorporates a very small model error. As stated in subsection 4.2.10, 13 out of the 15 runs used in this chapter assimilate datasets with at least 60% of daily data present, and so there are never large enough gaps between data points for the model error to grow to anything beyond insignificant. Therefore despite the two DA schemes having slightly different setups, there is no evidence that the resulting differences in the posterior distributions would be very small if any. Evidence of this can be seen when the parameter estimates and uncertainties from the REFLEX project, where these exact two DA methods with the same setups, were used. The relevant plot can be found in subsection 2.3.4, where more detail is given along with subsection 4.2.10.

A further possible reason for the lack of convergence to the global minimum could be the type of distribution used. The proposal distribution used in the Metropolis algorithm for this setup is discrete, as opposed to being continuous as it the norm with a more traditional setup of the algorithm. The chains of the Metropolis algorithm were also all initialised from the same location in the parameter space rather different locations, and this may have resulted in convergence to a local minimum rather than a global one. However, as with the other possible reasons, a counter-argument to these final two possibilities is why convergence to the global minimum was still reached for four of the runs. These issues are other types of implementation issues, and it is unusual that if these are the reasons why they are not causing the other four runs to not convergence to the global minimum.

Given these four possible reasons, the cause of the lack of convergence to the global minimum for these remaining eleven runs is unknown. An interesting exercise

would be to repeat all 15 of the Metropolis algorithm using a more traditional version of the algorithm as outlined at the start of subsection 2.2.8.

EFFECTS OF THE METROPOLIS ALGORITHM NOT BEING AT THE GLOBAL MINIMUM: EQUIFINALITY AND AUTOCORRELATION OF THE RESIDUALS IN THE COST FUNCTION

Equifinality occurs when different parameter sets result in equally good fits of the model outputs and data. It is known problem and has been investigated or identified as a potential reason for causing differences in parameter estimates and uncertainties using batch DA schemes such as the Metropolis algorithm, as highlighted in Tang & Zhang (2008), Williams et al. (2009), Medley et al. (2005) and Luo et al. (2011). This is investigated in detail in relation to the results from this chapter in subsection 4.4.3.

Although the noise added to the true NEE data is guaranteed to be independent in time, there is no such guarantee for the residuals between the modelled NEE and the NEE. An assumption of the cost function used for the Metropolis algorithm is that the residuals are not autocorrelated, therefore if they are this could result in biased parameter estimates and uncertainties. If the residuals are autocorrelated for the likelihood function used in the EnKF, this does not matter because unlike the likelihood in the Metropolis algorithm which involves all the data, the likelihood for the EnKF only involves the data at one time point. In the results of this chapter the residuals were found to be autocorrelated in the 11 of the 15 runs, the same 11 runs where there were considerable differences in the parameter results between the two DA schemes, compared to the remaining 4 runs where the parameter results were very similar. Therefore, as a result of the lack of convergence to the global minimum, this is potentially the cause of the divergence in parameter estimates between the EnKF and Metropolis algorithm as observed in the results. This is discussed further in subsection 4.4.4.

In subsections 4.4.3 and 4.4.4 that now follow, we go into more detail of the evidence and implications of equifinality and autocorrelation of the residuals in the cost function.

4.4.3 Equifinality

The results showed that the Metropolis algorithm was far worse than the EnKF at est-

imating parameter and the final C stocks, but gave more accurate forecasts. The most likely reason for this is that the structure of the model introduces "equifinality" (Tang and Zhang, 2008; Williams et al., 2009). This is where a set of parameters that is different to the true parameter set can give a near optimal fit to data. There is strong evidence of this affecting the Metropolis algorithm's results because, although the corresponding parameter estimates deteriorate in most cases with longer/more dense/less noisy datasets, the fits of the modelled NEE to the observed NEE during the DA period remained more or less unchanged, with narrow percentile intervals (figure 4.21), although the RMSE is noticeably less for runs 1, 2, 3 and 8 (table 4.2).

EQUIFINALITY AND THE METROPOLIS ALGORITHM

One reason why equifinality is occurring is because the output of the model (i.e. NEE) may be insensitive to changes in a number of the parameters, and in fact only a select few have significant influence. We tested this by performing sensitivity analysis (SA) on the DALEC-D model to see how sensitive NEE accumulated over 1 year was to changes in the parameters. The details of this sensitivity analysis are not included here because the output used in the sensitivity analysis was chosen to be the root mean squared error (RMSE) between the modelled NEE and the true NEE, since the output could only be scalar. This was perhaps a bad choice for the output, and an arguably better one would have been the iNEE, i.e. NEE aggregated over all the time points.

Despite this, the results to this sensitivity analysis did suggest that there are some parameters that have the largest influence on the modelled NEE, and that there are other 'redundant' parameters that have little effect. If this is true, then as stated in the previous subsection, the equifinality observed in the results may in part be due to two things:

(i) For the parameters that have the largest influence on the modelled NEE, the value the Metropolis algorithm chooses is based on which gives the best fit of the modelled NEE to the observed NEE.

(ii) For the redundant parameters, they can take any value and they will have little effect on the trajectory of modelled NEE at any iteration of the Metropolis algorithm. This is because when the Metropolis algorithm searches the parameter space for the region resulting in optimal fits to the NEE observations, there is little information in the NEE observations to decide what the optimal values of these re-

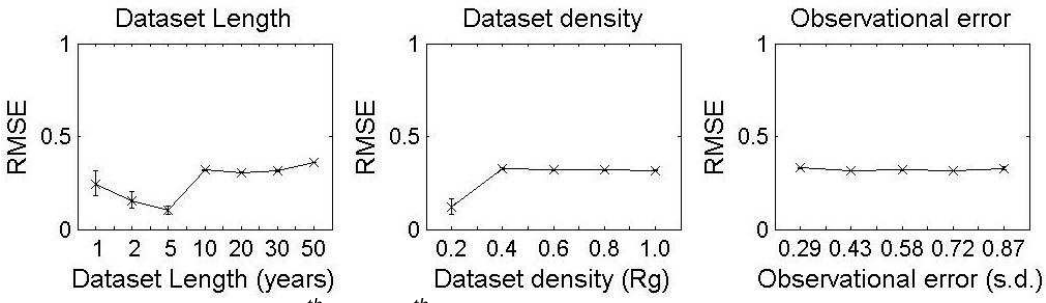


Figure 4.21. The mean, 5th and 95th percentile values of the RMSE (root mean squared error) between the modelled NEE during the DA period corresponding to 500 randomly chosen parameter sets from the posterior distribution of the Metropolis algorithm and the observed NEE.

dundant parameters should be. Therefore, the choice of these parameters is based on how correlated they are with the other parameters which have more influence on the modelled NEE.

Inspecting the parameter correlations for the posterior distributions, shown visually for four runs in figure 4.12, we see high correlations between many pairs of parameters. Therefore, if equifinality is the cause of the differences in Metropolis algorithm's parameter estimates and uncertainties to the EnKF's ones, then it is very reasonable to suggest that the posterior parameter correlations may be influencing the estimations of these 'redundant' parameters. Despite these reasonable arguments there is insufficient evidence from the results to support them, therefore we are not able to explain the causes of equifinality here. However, what is undisputed however is that it is happening and we see this clearly in the results.

HOW DOES EQUIFINALITY RESULT IN GOOD FORECASTS?

One important question is: how does equifinality result in the Metropolis algorithm producing good forecasts? Since the meteorological data has no extreme events (e.g. droughts) and the atmospheric CO_2 levels are constant throughout the 50 year DA period and 10 year forecast period, the trajectory of the true NEE is very periodic. Although the observations have noise and gaps, the periodicity of the true NEE trajectory is preserved. This periodic nature is continued during the forecast period. Therefore, once the Metropolis algorithm has found an optimal fit to these observations, it has essentially found an optimal fit to the true NEE. Thus, it is likely that a set of estimated parameters which resulted in a good fit to the NEE observations during the DA period will also result in the modelled NEE being close to the true NEE

during the forecast period.

In figure 4.21, we can also see that the 5^{th} – 95^{th} percentile intervals are very narrow for all 15 runs. Thus, it is perhaps expected that the percentile intervals would also be narrow for the distribution of values of RMSE of the modelled versus the true NEE trajectory during the forecast period, as can be seen in figure 4.14. Therefore, it is no surprise that the distributions of iNEE forecasts for the Metropolis algorithm is narrow (figure 4.13). The exception to this though is the 1 year dataset run, which has a very large percentile interval. As explained in subsection 4.3.4, this is probably because in order to get the NEE trajectory of the forecast period to be close to the truth, you need at least 2 years of assimilated NEE observations in order to get the interannual variability correct.

Finally, it is important to make the distinction between the uncertainty estimate represented by the percentile intervals of the forecasts and real-world representation of uncertainty. As explained above, it is perfectly reasonable that the percentile intervals for the Metropolis algorithm are narrow, but this does not mean that the actual uncertainty intervals we might expect for such a forecast would be as narrow. For example in run 3, the 10 year forecast estimate is approximately -2200 +/- 200 gC/m², i.e. the percentile interval is approximately +/- 10%. Considering 5 years of assimilated data is used to make a 10 year forecast, the interval on this forecast does appear unrealistically narrow. This over-confidence of the Metropolis algorithm is a limitation of the algorithm.

THE ENKF

Equifinality due to the structure of the model can also affect the EnKF parameter estimates, but its impact is likely to be less than for the Metropolis algorithm. This is because the EnKF does not try to find the best fit of modelled NEE to the observed NEE for the entire dataset, as the Metropolis algorithm does. Instead it compares modelled NEE with the observed NEE one time-step at a time. This explains why most of the parameter estimates improve when assimilating longer, more dense and more accurate datasets.

An interesting observation is that the EnKF's parameter estimates have moderately narrow posterior intervals relative to the parameter estimates, yet the 5^{th} – 95^{th} percentile intervals of its forecasts are a lot wider than the Metropolis algorithm's.

This is because using the estimated parameters (even though some come from assimilating very long datasets) to forecast so far into the future (10 years) is likely to result in large spread. If the forecast period was much shorter, say 1 year, the intervals would be a lot narrower. Figure 4.23 shows iNEE annual forecasts for each year of the 10 year forecast period, for run 4 (table 4.2) which had a dataset length of 10 years, observational noise of 0.58 gC/m², and a data density of $R_g = 0.6$. In the 10 year forecast plots (figure 4.13), the length of the percentile interval for run 4 is around 3000 gC/m², which is 300gC/m² on average per year. In the top panel of figure 4.22, we can see that the length of the percentile interval is around 150 gC/m² for the 1st year of the forecast period. For subsequent years, the interval gets wider and wider, which is what would be expected since the forecasts are further into the future. This also happens for the Metropolis algorithm (lower panel of figure 4.22), but much less so than the EnKF. Therefore, when one compares the 1 year forecast of the EnKF with that of the Metropolis algorithm, the difference in the size of the percentile intervals is not as great as for the 10 year forecast (figure 4.13).

Even when comparing the 1 year forecasts, there is still a large difference is percentile interval sizes between the two DA schemes. Rather than being due to the EnKF interval size being too large, it is more likely that the Metropolis algorithm's interval size is too narrow. As explained, this is because the Metropolis algorithm minimises the difference between the observed and modelled NEE; therefore the RMSE of the difference between the true and modelled NEE of the parameter sets in the posterior distribution is very narrow (figure 4.21). Hence the distributions of iNEE forecasts are also narrow (figure 4.13). As stated earlier, this does not mean that the narrow distribution of the forecast estimates from the Metropolis algorithm represents the real-world estimate of the uncertainty of such a forecast.

LIMITING THE EFFECTS OF EQUIFINALITY

Although equifinality amongst the Metropolis algorithm runs was only observed as the 11 runs which it is has been decided should be excluded from this chapter, it is still worthwhile briefly explaining how equifinality can be avoided. Equifinality is a problem because for the Metropolis algorithm, parameter sets other than the true parameter set can produce optimal fits of modelled NEE to the observed NEE. The sensitivity analysis results indicated that it is likely that equifinality may be in part due

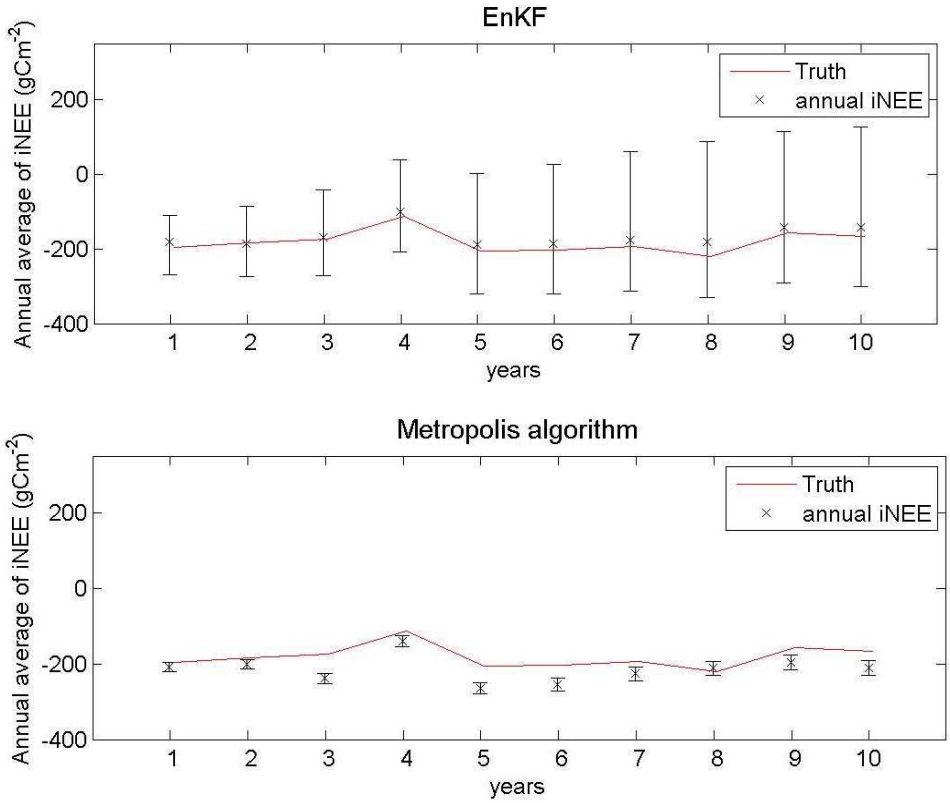


Figure 4.22 The forecast iNEE for run 4 of the EnKF (upper panel) and the Metropolis algorithm (lower panel) (dataset length=10 years, Observational error=0.58gC/m², data density (R_g)=0.6), calculated on a year by year basis rather than over 10 years. The lower and upper bounds of the error bars are the 5th and 95th percentiles, and the cross is the mean. The red line is the true iNEE.

to redundant parameters in the model which have minimal effect on the modelled NEE. One way of minimising the effects of equifinality is by constraining more of the processes of the model. This can be done by changing the structure of the model to impose more constraints on the model equations. Additionally, assimilating other types of observations is likely to be beneficial. Richardson et al. (2010) used a variant of the Metropolis algorithm to show that when more types of observations are assimilated, the uncertainties of the parameters reduce and the predicted states match more closely to the observations.

4.4.4 Autocorrelation of residuals

In the results (subsection 4.3.10), plots of the autocorrelation function (ACF) or lag plots were constructed for each of the 15 runs to determine the autocorrelation of the

residuals in the cost function used in the Metropolis algorithm. It was found that runs 1, 2, 3 and 8 (table 4.2), corresponding to runs where the dataset length was 1, 2 and 5 years and where the data density was 0.2, had close to zero autocorrelation for the different lags (e.g. run 2, left panel of figure 4.16). For the remaining 11 runs, non-insignificant and consistent autocorrelation was discovered (e.g. runs 7 and 12, central and right panels of figure 4.16). Although it is difficult to determine the effect of autocorrelated residuals, a number of patterns have emerged in the results. For each of these 11 runs from the Metropolis algorithm, it was found that:

(i) Most of the parameter estimates were far from the truth and very few of the posterior uncertainty intervals contained the truth. The converse was true for the other 4 runs.

(ii) Most of the posterior parameter estimates were very different to the EnKF's ones. The converse was true with the other 4 runs.

(iii) The modelled NEE using the parameter sets from the posterior distribution fitted the NEE data worse, compared to other 4 runs (figure 4.22).

(iv) The true parameter set had a cost function value which was significantly lower than those computed from the parameter sets in the posterior distribution. The true parameter set was contained in the posterior distribution for the other 4 runs.

While it is difficult to say for definite that these four poor features of the results in these 11 runs are linked to the autocorrelation of the residuals, it seems very unlikely that these effects occurred by chance. Above all, the poor behaviour of the results from the Metropolis algorithm for these 11 runs is of great concern and makes it difficult to draw any meaningful conclusions from these runs. Therefore, as stated at the end of the previous subsection, it is decided that these 11 runs should be excluded from any conclusions to this chapter and in particular when addressing the hypotheses.

Autocorrelation in the residuals in the cost function when using the Metropolis algorithm has been identified by Vrugt et al. (2005), and Feyen et al. (2007) suggested a transformation of the residuals in order to reduce or eliminate the this autocorrelation. An interesting study involving the Metropolis algorithm to estimate the parameters of the CASA model also using synthetic data (MacBean et al., 2010) used the following likelihood function:

$$P(D|f(\Theta)) = f(\Theta) - D = N(0,\sigma^2)$$

where D represents the data, $f(\Theta)$ is the model output using a particular set of param-

-eters Θ. The likelihood is then determined as a random sample from the Normal distribution with mean zero and a variance σ^2 calculated from the variance of the residuals. A key advantage of this approach is that it is a simple but effective way of eliminating the autocorrelation of the residuals in the likelihood function.

4.4.5 How the results compare with our expectations of them

In subsection 4.2.10, expectations of the results in this chapter are described. In summary, it is stated there that although the two DA schemes are setup in slightly different ways, this was not expected to cause anything beyond insignificant differences in the posterior parameter distributions. This was demonstrated by referring to the REFLEX study (Fox et al., 2009) where the same two DA algorithms used in this chapter were used there and gave very comparable parameter estimates and uncertainties. As a result of this, we expected the runs corresponding to the shorter datasets to have similar results between the algorithms because the data used in REFLEX covered only 2 years of daily data. This turned out to be true. For the runs corresponding to the dataset lengths of 1, 2 and 5 years (runs 1, 2 and 3 from table 4.2), the parameter estimates and uncertainties were very similar for both schemes as described at the start of subsection 4.4.2.

For the remaining runs, i.e. where the dataset length was higher and where the observational error and dataset density varied, it is stated in subsection 4.2.10 that it is unclear what to expect from the results. Instead, possible factors were considered that may result in the results between the two DA schemes being different. The first was equifinality occurring when using the Metropolis algorithm, and certainly evidence was seen in the results where the biased parameter estimates and over-confident uncertainty intervals resulted in good fits of the modelled NEE to the NEE data during the DA period but also during the forecast period. However, it has been difficult to understand why this is happening for certain runs and not others. The other main potential reason given for the differences was one of the assumptions being broken, such as autocorrelation of the residuals in the cost function. For 11 of the remaining 12 runs from table 4.2, it was found that the cost function residuals as used in Metropolis algorithm were indeed autocorrelated for a number of lags. It was for these 11 runs, that the only differences in the results between the two DA schemes were observed, as outlined in subsection 4.4.4. Although no link has been provided

between the autocorrelation of the residuals in these runs and the poor results for the Metropolis algorithm, it seems unlikely that this happened by chance.

Further evidence of the autocorrelated residuals resulting in poor parameter estimates and uncertainties can be found when the result of the final run, namely run 8 is considered. In run 8 the data density is 0.2, in other words the data is most thinly spread in the dataset. As a result of this it is less likely we would see temporal correlations of the residuals because the data are so sporadic. This is exactly what occurs. Like runs 1, 2 and 3, run 8 also had close to zero autocorrelation in the residuals of the cost function used in the Metropolis algorithm. As with runs 1 – 3, we then also see that the Metropolis algorithm's results are very comparable to those from the EnKF.

As stated in the previous subsection the 11 runs, where we saw poor results from the Metropolis algorithm, have now been removed from the conclusions of this chapter and when addressing the hypotheses. This is because it is difficult to draw conclusions from poor results, which appear to be due to a violation of the assumptions of the Metropolis algorithm.

4.4.6 Other relevant issues

UNCHANGING POSTERIOR UNCERTAINTIES FOR VARYING OBSERVATIONAL ERROR

As stated at the end of subsection 4.3.3, the results from the EnKF runs indicate that the posterior uncertainty remains unchanged with increases in the observational error. This is intriguing and potentially important because it suggests that assimilating very noisy data is just as good as assimilating non-noisy data. A practical application of this is the florescent data retrieved from the GOSAT satellite (Guanter et al., 2012). It is hoped that this data will be assimilated into a model, however there has been concern over this due to the excessively noisy nature of the signal. The results from this thesis suggest that despite this noisy data, assimilating it into a model will still result in near optimal solutions to parameters and model states.

POSTERIOR PARAMETER ESTIMATES AND UNCERTAINTIES COMPARED TO THE PRIOR

In subsection 4.3.6 it was found that most of the parameter estimates from the posterior

distributions of all of the EnKF runs had a reduced uncertainty interval, whilst still containing the truth, compared to the prior uncertainty intervals. For the Metropolis algorithm, the same reduction in uncertainty was found in the runs corresponding to the dataset length of 1, 2 and 5 years and a data density of 0.2 (runs 1, 2, 3 and 8 from table 4.2). These reductions in uncertainty in the parameter estimates are encouraging to see but also key for DA to be of worth doing.

For many of the parameters, the reductions in uncertainties were very large (e.g. F_g, F_{nf} from table 4.6) but this is partly due to the prior uncertainty being large, sometimes covering the entire parameter range. However for other parameters the prior uncertainty was narrower, but this narrowness is not necessarily a good thing. For example for T_{lab} (figure 4.6a), the prior uncertainty interval does not contain the true value. One could argue that in reality, as we do not know the true value, the prior uncertainty on the parameter estimates may be over-confident since our prior knowledge about certain parameters may be limited. For T_{lab}, another factor which influences whether the prior uncertainty interval contains the truth is the value of the prior estimate. For experiments like this where synthetic data is used, the modal value of the prior distribution is determined by randomly choosing a parameter set from a multivariate distribution with the true parameter set as the modal value and the same variance used for the prior. While theoretically this method of selecting the prior modal values is adequate, in reality the choice would be based on prior knowledge rather than randomness.

THE DIFFERENCES IN UNCERTAINTY ESTIMATES BETWEEN THE TWO DA SCHEMES

One feature of the results was that the huge differences in uncertainty estimates of the parameters between the two DA schemes for many of the runs. However, these differences only existed in the 11 runs stated on the previous subsection. Since these 11 runs from Metropolis algorithm have now been excluded from the analysis and conclusions to this chapter, we instead focus on the remaining 4 runs. For these runs, which correspond to the dataset length of 1, 2 and 5 years and a data density of 0.2 (runs 1, 2, 3 and 8 from table 4.2), there is very good agreement among the uncertainties of the posterior parameter estimates. This can be seen from figures 4.6 and 4.7, and in table 4.10 we can see that good agreement between the posterior

uncertainty estimates of the two DA schemes occurred for between 14 and 15 parameters. In REFLEX, where exactly the same two DA schemes were used using exactly the same setups, this good agreement occurred for 13 of the parameters. This feature of the results then is as we would expect.

ROBUSTNESS OF FORECASTS

In subsection 4.3.9, it was found that the forecasts using both DA schemes were robust against a synthetic drought that was imposed during the forecast period. Moreover the fit of the drought-induced forecasts to the truth was actually improved for both DA schemes, although in only a minor way. This is encouraging particularly given the poor parameter estimates and uncertainties for 11 of the Metropolis algorithm's runs. However, this highlights the need for caution when using solely using forecasts as a measure as a measure of algorithm performance as other studies have done (e.g. Richardson et al., 2010).

ENTROPY

In subsection 4.3.10, it was shown that as dataset length increases the entropy or uncertainty of the posterior distribution of the parameters, using the EnKF, becomes narrower. In other words the data in the longer datasets are having more of an impact on the posterior distribution than the shorter datasets. This can be seen in table 4.9 which shows the Mutual information increasing as dataset length increases (if the mutual information is zero this implies that there has been no reduction in the entropy of the posterior compared to the prior distribution). It is unsurprising that this reduction in entropy is happening as the dataset length increases – with more data being assimilated it is perfectly reasonable to expect the uncertainty on the post-DA estimates of the parameters to decrease. What would be interesting would be to see if this reduction in entropy would also occur for the Metropolis algorithm. It is also interesting to note from figure 4.9 that the Mutual information is highest for the 20 year run, not for the 50 year run as might be expected. However, the MI is only slightly less for the 30 and 50 year runs compared to that of the 20 year run. Therefore, with the information that we have, a reasonable explanation is that the MI begins to level off at the 20 year run onwards. The slight reduction in MI from the 30 and 50 year runs could be then just to the stochastic nature of the EnKF. For example,

if we were to repeat the runs we may indeed find that the highest MI value is for the 30 or 50 year run.

In subsection 4.3.10, it was also shown that the mean parameter vector of the posterior distribution moved further away from the mean vector corresponding to the prior distribution. Again, this indicates that the data in the longer datasets are having more of an impact on the location of the posterior mean of the parameters than the data from the shorter datasets. The exception to this rule was for the 20 year dataset – the mean vector of the parameters for the posterior distribution was very close to that of the prior. This is likely to be an anomaly but an interesting extension to this problem would be to understand why this is happening.

Overall therefore, more information being assimilated results in parameter estimates having a lower entropy in the posterior distribution compared to the prior and a posterior mean vector that is further from the prior mean.

4.4.7 Hypotheses

In addressing these hypotheses, an improvement in the parameter and stock estimates and uncertainties was determined for an increase in dataset length (H1 and H2), an increase in data density (H3), a reduction in observational error (H4) was defined as either the parameter/stock estimate moving closer to the true value or the corresponding uncertainty interval on that estimate becoming narrower while still containing the truth. The same measure of improvement was used in addressing the hypothesis on the forecasts (H5). Details of this measure are given at the start of subsection 4.3.3.

H1: LONGER DATASETS GIVE IMPROVED PARAMETER AND STOCK ESTIMATES

The plots of the EnKF estimates of the parameter and stocks and corresponding uncertainties are displayed in figures 4.6a and 4.6c and summarised in table 4.3a and table 4.4a. These show that as dataset length increases these estimates improve for 12 of the 17 DALEC-D parameters and 4 of the 6 stocks. For 3 of the remaining parameters and 1 of the remaining stocks, there is no improvement with an increase in dataset length, while for the final parameter and final stock there is deterioration in the estimates with dataset length.

For the Metropolis algorithm 4 of the 7 dataset length runs, which corresponded to dataset lengths of 10, 20, 30 and 50 years, were not included as part of the conclusion and in addressing the hypotheses because the results from these runs are likely to be unreliable most likely resulting from violation of one of the assumptions of the Metropolis algorithm (see subsections 4.4.3 to 4.4.5). However for the runs 1 to 3, corresponding to dataset lengths of 1, 2 and 5 years, the results are valid and the corresponding plots of the parameter and stock estimates and uncertainties is given in figures 4.6b and 4.6d. From these plots, we see that for 16 of the 23 parameters and 3 of the 6 stocks there is an improvement in the parameter and stock estimates with dataset length. For 5 of the remaining parameter and all of the 3 remaining stocks, there is no improvement with an increase in dataset length, while for the final 2 parameters we see deterioration.

In conclusion, there is strong evidence from the results that this hypothesis is true. Overall, we find that in most parameters and stocks, longer datasets result in improved parameter and stock estimates for both DA schemes. However, it is important to note that the longest dataset for the Metropolis algorithm on which this evidence is based is 5 years, whereas for the EnKF it is 50 years.

H2: SLOWLY VARYING C POOLS AND THEIR CONTROLLING PARAMETERS ARE ONLY CONSTRAINED BY LONGER DATASETS

This hypothesis is based on the idea that the estimates of the parameters and final stocks relating to the slowly varying pools (T_w, T_s, C_w and C_{som}) would move closer to their true values or the uncertainty intervals become narrow while containing the true value with the assimilation of longer datasets, as these parameters and final stocks only have a large impact on the model over long time-scales. Since the Metropolis algorithm runs corresponding to the dataset length of 10 to 50 years are excluded from the conclusions to this chapter, this hypothesis cannot be answered for the Metropolis algorithm. However it is interesting to note from figure 4.9 the Metropolis algorithm estimate for the T_w parameter moves closer to the true value as dataset length increases up to 50 years.

For the EnKF, figure 4.6c shows that the only one of these slowly varying parameter and final C stocks whose estimate is improved under the longest dataset length is C_w. For the other three, figures 4.6a and 4.6c show that T_w remains

unchanged with dataset length, whereas T_s and C_{som} are further from the true value for the longer dataset. The possible reason for this is that these slowly varying parameters and final stocks need a far longer datasets to be assimilated since they operate over large time-scales. We see evidence of this from inspecting the T_w subplot in figure 4.6a. As dataset length moves from 1 year to 50 years, we see a very slight move towards to the true value which can be seen when this subplot is plotted with a larger upper bound on the y-axis (figure 4.9).

However as mentioned above, we see a deterioration of the EnKF estimates of T_s and C_{som} with an increase in dataset length, which is not very encouraging. One explanation for this is that the C_{som} pool changes more slowly that the C_w pool, since the true values of the turnover rates of the two pools are 3×10^{-5} and 7×10^{-5} respectively. Therefore while the T_w parameter appears to require longer datasets for it to be well constrained, T_s and C_{som} require datasets several orders of magnitude greater. The probable reason that the T_s and C_{som} estimates are closer to the truth for the shorter datasets is due to the prior, which happens to be very closer to the true value. Since the shorter datasets are very unlikely to have any effect on the posterior distributions of T_s and C_{som}, it is reasonable to assume that T_s reverts to the prior distributions, causing the final value of C_{som} to also do the same.

In conclusion, we find that there no evidence that this hypothesis is true for the EnKF probably because there is not enough information even in the 50 year dataset to constrain the slowly varying parameters and final stock estimates. For the Metropolis algorithm, it is not possible to test this hypothesis.

H3: DATASETS WITH A GREATER DENSITY OF OBSERVATIONS GIVE IMPROVED PARAMETER AND STOCK ESTIMATES

Figures 4.7a and 4.7c shows that 13 out of the 17 parameters and 4 out of the 6 final C stocks improved with an increase in the density of the data in the dataset, and for the remaining parameters and stocks there was no improvement. These findings are summarised in table 4.3a. Unlike an increase in dataset length, the improvement here was mainly due to a reduction in the posterior uncertainty intervals. The actual parameter and stock estimates tended to be close to the true value and remain unchanged for all or most levels of the data density. What it means is that by reducing the density of observations to 20%, we only have a $1/5^{th}$ of the information

yet it is still possible to estimate the parameters just as well (for the EnKF) and forecast just as well (both methods) as when we have all the information. One explanation for this is because the denser datasets do not carry as much extra information as less dense ones do, since the post-EnKF trajectories of the model states of each type of dataset density would have similar paths. However it is worth bearing in mind two things:

(i) If instead of the gaps being spread out across the dataset, we removed all the data for a particular season (e.g. summer), this is more likely to have a greater impact on the estimates.

(ii) If we decreased the decreased the density of observations beyond 20%, e.g. 10%, 5%, 1%, …, then at some point the parameter estimates would likely deteriorate.

For the Metropolis algorithm, since four of the five runs where the data density varied are excluded from the conclusions of this chapter it is inconclusive whether increasing the data density would improve the corresponding estimates of the parameters and final C stocks. What is interesting though is that for the one run which was not excluded here, namely run 8 which correspond to the data density of 0.2, 18 out of the 23 parameters and 4 of the 6 final C stocks are very close to the true value. For the remaining parameters and final C stocks, the Metropolis algorithm estimates are fairly close to the truth. Given that for the 4 Metropolis algorithm runs not excluded from the conclusions to this chapter agree well with the EnKF's estimates, it is reasonable to assume that if the remaining 4 Metropolis algorithm runs had not been excluded with no autocorrelation in the cost function residuals, we might expect the Metropolis algorithms parameter and final C stock estimates to improve with dataset density in the same as the EnKF estimates do. However, the results for the Metropolis for the most part remain inconclusive so it is not possible to make that judgement.

In conclusion, we find that there is strong evidence that this hypothesis is true for the EnKF, since an increase in dataset density results in the majority of EnKF parameter and final C stock estimates improving. It is inconclusive about whether this hypothesis is true for the Metropolis algorithm, although it is encouraging that most of the parameter and final C stock estimates for the lowest data density level were close to the truth with the majority of the uncertainty intervals containing the truth.

H4: DATASETS WITH A SMALLER OBSERVATIONAL UNCERTAINTY GIVE IMPROVED PARAMETER AND STOCK ESTIMATES

Figures 4.8a and 4.8c and table 4.3a shows that as the observational error decreases, there is an improvement in the EnKF estimates of 14 of the 17 parameters and 4 of the 6 final C stocks, with the remaining parameters and final C stocks showing no improvement. As with changes in data density, the improvement observed in figures 4.8a and 4.8c is mainly due to reduction in the uncertainty intervals of the parameter and final C stock estimates, with the actual estimates remaining close to the truth and relatively unchanged for the different levels of observational error. As with H3, this can be explained by realising that as long as the general trend of true NEE trajectory is preserved, the amount of noise added will yield similar sets of estimates. It is unclear if this result would hold if real observations were used where the error on the observations will not only have random components but potentially systematic biases as well.

For the Metropolis algorithm, all the runs here are excluded from the conclusions to this chapter, so it is inconclusive about whether there would be an improvement. However as commented in the H3 assessment above, given that the parameter and final C stock estimates agreed well among the two DA schemes, it could be assumed that if the cost function residuals were not correlated we would get a similar improvement in the estimates with a reduction in observational error, as seen with the EnKF ones.

In conclusion, there is strong evidence that this hypothesis is true for the EnKF, since a reduction in observational error leads to the majority of EnKF parameter and final C stock estimates improving. It is inconclusive about whether this hypothesis is true for the Metropolis algorithm,

H5: DATASETS OF GREATER LENGTH, GREATER DATA DENSITY AND OF SMALLER OBSERVATIONAL ERROR RESULT IN MORE ACCURATE POST-DA FORECASTS

The top panel of figure 4.13 shows that an increase in dataset length and data density and a reduction in observational error results in an improvement in the forecasts for the EnKF. As with the parameter plots the improvements were characterised only by a reduction in the forecast uncertainty intervals for the data density and observational

error levels. For the dataset length runs, the forecasts also moved closer to the truth. For the Metropolis algorithm forecasts, an improvement is evident with an increase in dataset length up to 5 years, but beyond this and for the other factors any other improvements are inconclusive due to 11 of the runs being excluded from the conclusion to this chapter.

In conclusion, there is very strong evidence that this hypothesis is true for the EnKF, but only partial evidence for the Metropolis algorithm as a result of the excluded runs.

4.5 Conclusion

The length of the dataset is most important when using the EnKF to estimate the DALEC-D parameters based on NEE observations, and when using both DA schemes to make forecasts of cumulative NEE. Varying the density of the observations or the level of observational error has less impact in terms of the estimates but we do see a slight reduction in the uncertainty which is encouraging. Datasets of at least 5 years in length are required in order to get good parameter estimates for the EnKF, and good forecast estimates for both DA schemes.

The Metropolis algorithm and the EnKF gave markedly different estimates of the parameters and final C stocks. The EnKF produced accurate estimates, with a high percentage of 90% posterior intervals that contained the truth (88% for the parameters and 87% for the final C stocks). The estimates improved noticeably with longer datasets, and only slightly for denser and less noisy datasets. The Metropolis algorithm estimates were on average further away from the truth, had a much lower percentage of 90% posterior intervals that contained the truth (40% for the parameters and 53% for the final C stocks), and in general deteriorated with longer, denser and less noisy datasets. When the parameter estimates were used to forecast cumulative NEE 10 years into the future, both schemes produced forecasts close to the truth for runs 5 years or more in length. However the widths of the corresponding uncertainty intervals were very different for the two schemes for 11 of the runs, with the Metropolis algorithm producing very narrow ones, and the EnKF having much wider ones. For the remaining 4 runs, namely runs 1, 2, 3 and 8, the parameter and final C stock estimates were very similar to those of the EnKF.

It was found that DALEC-D is susceptible to equifinality, so that a parameter set which is not close to the true parameter set can produce an NEE trajectory which is close to the true trajectory. This means that NEE produced by DALEC-D may be insensitive to changes in at least some of its parameters. The Metropolis algorithm is particularly affected by equifinality due to the algorithm trying different parameter sets to find a good fit to the observations, which results in the trajectory of forecasted NEE being close to the truth as well. Equifinality can also affect the EnKF, but its impact is likely to be less. To test the robustness of the Metropolis algorithm's forecasts, a severe drought was imposed in the middle 2 years of the forecast period. Surprisingly, the forecasts remained very accurate. Despite this, the Metropolis algorithm's forecasts have uncertainty intervals that appear unrealistically narrow in terms of representing real-life uncertainty. However it is not known if real data had been used the Metropolis algorithm's forecasts would also have such narrow percentile intervals which contained (or almost contained) the truth. Further work assimilating more realistic datasets (e.g. modelling systematic error) is therefore necessary to determine whether the Metropolis algorithm's forecasts would remain close to the truth and whether its percentile intervals would still be narrow.

More important than equifinality it was found that for 11 of the runs, the residuals in the cost function used for the Metropolis algorithm were autocorrelated violating one of the assumptions of the Metropolis algorithm. It was for these same 11 runs that the parameter estimates were not only very different to the EnKF's but they were also far from the truth with the uncertainty intervals rarely containing the truth either. Perhaps more importantly is was found that the true parameter set had a small cost function value than any of the values corresponding to the parameter sets in the posterior distributions. For these reasons, these 11 runs are excluded from the conclusions of this chapter. Despite this, this autocorrelation of the residuals was an interesting finding and serves as a warning to other studies were the Metropolis algorithm is used with synthetic data.

Both equifinality and the autocorrelated residuals are probably the result of the 11 'bad' runs failing to converge to the global minimum.

Chapter 5
Assimilating satellite observations of Leaf Area Index into a terrestrial carbon model

Abstract

Leaf area index (LAI) is an essential climate variable strongly linked with vegetation growth, the water cycle and the energy balance (Sellers, 1997; IPPC, 2007). Thus it is important to be able to measure it continuously and globally, which can only be achieved using satellites. The MODIS sensor on NASA's TERRA satellite offers such spatially and temporally frequent measurements of LAI. In this chapter we assess whether assimilating the latest version of the MODIS LAI observations (referred to as collection 5) into a simple terrestrial ecosystem model improved the estimates of Net Ecosystem Exchange (NEE) when compared to ground observations. The study was carried out using MODIS LAI data from three Evergreen sites, namely Loobos (Netherlands), Oregon (U.S.A.) and Tharandt (Germany).

Due to some of the lower valued MODIS LAI observations having very small prescribed uncertainties, this resulted in the post-assimilation LAI trajectory being dragged down to values close to zero during the growing season of at least one of the years for two of the sites. As a result, the post-assimilation NEE fitted the observed NEE much worse than the NEE from a run with no Data Assimilation, particularly during the growing season. The MODIS LAI data have another characteristics which are unrealistic, namely they are very noisy in time, varying by around 2–4 m^2/m^2 in

only 1 or 2 weeks in some instances. To overcome these limitations, the MODIS LAI observations were processed by smoothing them and increasing the observational uncertainty for small observations. These changes resulted in a significant improvement in the fit of the post-assimilation NEE estimates to the observed NEE. Assimilating the processed MODIS LAI also resulted in a slightly improved fit of NEE compared to the NEE from the run without DA. The resulting fit of the modelled to the observed GPP was substantially improved after assimilating the MODIS LAI. This strongly suggests that assimilating MODIS LAI along with total respiration (Re) data (since NEE = Re – GPP) would further improve the fit of modelled to the observed NEE. The results also indicated that the temporal variability of the observed NEE could be modelled better if water related observations (e.g. soil moisture) were also assimilated.

5.1 Introduction

Leaf Area Index (LAI) is an essential climate variable as it is an indicator of the health of vegetation, and is used to estimate evapotranspiration and the accumulation of carbon from photosynthesis (Sellers, 1997). In this chapter we indicate it is important to know, how it is used in models, how it is measured from the ground and space, what the advantages and limitations are to these two approaches, and why these two approaches do not always agree. We overcome some of the limitations of satellite-based LAI measurements when using such data to estimate the states of an ecosystem model in a Data Assimilation (DA) scheme, and we test whether this helps to improve the estimates of Net Ecosystem Exchange.

5.1.1 Background and Motivation

LAI (m^2/m^2) is defined as 'the one-sided green leaf area per unit ground area, in broadleaf canopies, and as the projected needle leaf area in coniferous canopies' (Myneni et al., 2002). It has a strong influence on: (i) the interception of solar radiation, (ii) photosynthesis, (iii) transpiration, (iv) respiration, and (v) water interception (Breda, 2003).

(i) & (ii) If LAI is high there is more foliage (i.e. leaves, needles) meaning the total surface leaf area will be larger, resulting in more interception of solar radiation and more photosynthesis (Clark et al., 2008). However, the more the leaves clump together the less impact LAI has on radiation interception and photosynthesis. This is because clumping of the leaves means that solar radiation is less able to penetrate to the entire surface of each leaf (Breda, 2003). Thus, solar radiation interception and photosynthesis can be inferred from LAI as long as there is not excessive clumping.

(iii) Transpiration also depends on LAI because transpiration is proportional to GPP (i.e. the amount of C accumulated from photosynthesis); this is because the stomata guard cells that open to ingest CO_2 from the atmosphere release water molecules though these same guard cells (Gower et al., 1999). However, leaf clumping reduces the impact of LAI on transpiration because of the reduction in GPP due to clumping (Gower et al., 1999).

(iv) If the surface of the leaves are horizontal and not greatly clumped together, then autotrophic respiration can also depend on LAI since it can be treated as being proportional to GPP; for example, in the DALEC model roughly 40-50% of GPP is respired autotrophically (Williams et al., 2005). However if the leaves were for example oriented vertically, very little photosynthesis would likely be taking place but the leaves would continue to respire autotrophically.

(v) The amount of water intercepted can also depend on LAI because the larger the surface area of the leaves (i.e. the larger the LAI value) the more water will be intercepted. However, if the leaf angle is orientated vertically and/or there is a lot of clumping, water interception may not necessarily increase as LAI increases.

As a result, LAI is a key climate variable for estimating the growth of vegetation and for understanding key components of the water cycle and the energy balance. However, as noted above in points (i) to (v), clumping and leaf angle can have an impact when using LAI to infer other quantities of interest such as GPP or water inception. A number of models take this into account leaf angle when using LAI to estimate related variables such as GPP; for example the BETHY model (Knorr, 2000) assumes a uniform distribution of leaf angle among the foliage on the trees. However all of the models surveyed in this thesis, listed in subsection 1.2.3, do not assume any clumping in the foliage; although all of the more complex ones divide the canopy into multiple layers. Most also use Beer's law (Law & Wang, 1994) to describe the inception of solar radiation within each of the canopy layers. For example, in the SPA model (Williams et al., 1997) the implementation of Beer's law used, as described by Amthor (1994), accounts for leaf level and forest floor interception and reflection of incoming solar radiation. Therefore, although clumping is not specifically assumed, it is taken into account by incorporating different layers which absorb and reflect radiation by different amounts. We discuss issues related to clumping, leaf angle and multiple canopy layers in relation to the DALEC model, as used in this thesis, in the next subsection.

Once the issues of leaf angle, clumping and multiple canopy layers have been taken into account, then as stated above LAI is an important climate variable for predicting and estimating vegetation growth and health. In particular photosynthesis and respiration of C, helps us to understand how much CO_2 is being taken out of the atmosphere by plants. If we wish to accurately forecast atmospheric CO_2 we must

know with greater confidence how much is removed by plants. This loss of CO_2 from the atmosphere by plants is often referred to as Net Ecosystem Exchange (NEE), which is total respiration of C from the plant and soil minus Gross Primary Production or GPP (the amount of C amassed from photosynthesis). Therefore, accurate estimates of LAI should improve estimates of NEE, as we now explain in the context of carbon models.

5.1.2 LAI in models

A number of Dynamic Vegetation Models use LAI. SDGVM is an example of one designed to calculate the carbon fluxes and pools in the biosphere – Woodward et al. (1995), Woodward & Lomas (2004). It includes a number of biophysical processes which include those linked to evapotranspiration, hydrology, nitrogen cycling, and disturbances. Unlike smaller models like DALEC which can only be used for evergreen and deciduous forests, SDGVM can operate in many more types of vegetation including specific types of evergreen or deciduous trees, e.g. broadleaf evergreen and needleleaf evergreen. In total it has 14 plant functional types (PFTs). Each PFT requires 25 parameters so there are 350 parameters involved; this compares with DALEC which uses only 11 parameters for the Evergreen and 17 for the Deciduous versions. The purpose of SDGVM is to estimate carbon fluxes and pools globally under a changing climate. Temporal changes in LAI are an important way in which the model responds to this and LAI is determined, using the notation of Woodward & Lomas (2004), by:

$$\text{LAI} = \min\begin{cases} NPP \times (1 \times a_s \times a_r) \times SLA \\ NPP \times a_r \times W_u \\ \dfrac{P \times E}{V \times g} \end{cases}$$

where NPP = Net Primary Production, a_s and a_r are the allocations of stem (or wood) and root carbon, SLA is specific leaf area (i.e. ratio of leaf area to dry weight), W_u is the water uptake by roots, P is precipitation, E is evapotranspiration, V is vapour pressure deficit of the air, and g is the stomatal and boundary layer conductance. The vegetation dynamics is also more complex in SDGVM, with it having a mechanism

for modelling how a plant expands across a spatial scale, for example mid-latitude deciduous trees cannot grow in the tundra. Like DALEC, SGDVM is a point-based model, but unlike DALEC it can provide point estimates simultaneously across the globe, with a resolution specified by the user. The Lund-Potsdam-Jena or LPJ model (Site et al., 2003) has a similar structure to SDGVM, and although most of the processes are common to both of the models, their approaches can be fundamentally differently (Kantzaz, E., 2012, [personal communication]). As a result of this, while their estimates of Net Biome Production or NBP (=-NEE+disturbances) may be similar, other model outputs (such fire estimates or LAI) may be completely different.

In DALEC (Williams et al., 2005) – the model used in this thesis – C_f is determined only by $NPP \times (1 \times a_s \times a_r)$, using the notation above, with SLA used to scale it to LAI. In other words, the soil moisture content or precipitation and evapotranspiration amounts do not influence LAI. In Williams et al. (2005) Leaf Mass per Area or LMA (units g/m²) which is the reciprocal of SLA (Larcher, 2003) is used as the scaling factor. In other words, LAI is related to foliar Carbon (C_f) by the relation: LAI = $C_f \div$ LMA. SLA (or its reciprocal LMA) is a site-specific constant, which can be measured in the field by direct means by dividing its area by its mass (Meziane & Shipley, 2001). Measurements of SLA are also stored, and can be extracted for modelling purposes, from trait databases like TRY (Kattge et al., 2011).

In DALEC, C_f is one of the model states and so its value varies with time. The rate of change in C_f is given by:

$$\frac{dC_f}{dt} = (1 - F_g)F_{nf} \times GPP$$

where F_g and F_{nf} are parameters which refer to the fraction of GPP respired autotrophically and the fraction of the remainder which is allocated to foliage. In the deciduous version of DALEC, the change in C_f includes two terms which represent the amount leaving this pool to go to the labile Carbon (C_{lab}) pool (during the growing season) and the amount coming into this pool from C_{lab} which occurs during the Spring bloom. C_f is part of a feedback loop in the model since, as its value increases so does the value of GPP, since more leaves equate to more photosynthesis. A complete description of DALEC, including schematics of both versions, is given in chapter 2.

It is important to note that DALEC does not assume foliar clumping, and fixes the angle of the leaves so that they are always all facing the sun. It also only has one canopy layer. Therefore as a result of not making these assumptions, it could be argued that DALEC is too simplistic and an unrealistic representation of the real world. As a counter-argument the Aggregated Canopy Model (ACM), the sub-model of DALEC which computes the daily GPP estimates, was built as a highly computationally inexpensive version of the SPA model (Williams et al., 1997). Williams et al. (1997) found that under a broad range of conditions, the ACM was able to capture close to 100% of the model behaviour of SPA whilst being many orders of magnitude faster. This accurate representation was tested in a Deciduous forest site in North East USA, an tundra site in the Arctic and a range of forest types across the OTTER (Oregon Transect Ecosystem Research) transect in Oregon, running from coastal Sitka spruce to high-plateau mountain juniper. However, it is important to emphasise that the LAI data inferred from satellites (but also ground LAI data too) is 'true LAI' because clumping in the multiple canopies has been accounted for. However in DALEC, the ACM which computes GPP is a big leaf and so assumes no clumping; therefore the model will not get the interception properties of solar radiation correct. As a result the modelled LAI is not strictly speaking true LAI. Hence, there is a different interpretation of LAI between what is measured and what is modelled. This limitation to the model could be fixed by including a clumping factor in the calculations.

5.1.3 Measuring LAI from the ground and from space

MEASURING LAI AT GROUND-LEVEL

Ground measurements of LAI can be made directly or, more commonly, by indirect methods. Direct methods involve stripping the leaves off a tree or collection of trees and feeding each leaf through a scanner to determine its one-sided surface area (Gower et al., 1999). Direct methods therefore give very accurate measurements of LAI. However such methods are very rarely used because: (i) they are destructive, involving the removal of the leaves of a tree or trees; (ii) they are very time-consuming, as a tree can contain many thousands of leaves. As a result, indirect methods are normally used. There are many different approaches, but the most common involves estimating LAI by measuring the gap fraction of the canopy using

optical techniques (e.g. fish-eye photography) from different zenith angles. This involves inverting the modified Beer-Lambert equation (MacFarlane et al., 2007):

$$P(\theta) = \exp\left(-\frac{G(\theta)\Omega(\theta)(L+W)}{\cos(\theta)}\right)$$

where θ is the zenith angle, the angle at which a light beam is directed; $P(\theta)$ is the canopy gap fraction; $G(\theta)$ is the canopy extinction coefficient, defined as the fraction of leaf area that is projected onto a plane perpendicular to θ (Gower et al., 1999); $\Omega(\theta)$ is the clumping index (which equals l when the foliage is randomly distributed and is less than 1 when it becomes more clumped). Finally L and W represent leaf area index (LAI) and woody area index (WAI), the latter of which is similar to LAI but represents the area of woody material rather than leaves. Measurements of the canopy gap fraction are required to estimate plant area index or PAI (L+W) and $G(\theta)$ (which is generally unknown). To determine LAI, we also need to determine the ratio of WAI to PAI. A variety of different instruments are available to measure the canopy gap fraction but a common one is the Li-Cor LAI Plant Canopy Analyzer (Gower et al., 1999, Hyer & Goetz, 2004).

Indirect methods are much less time-consuming than direct methods and have become the standard way of measuring LAI at ground-level. However, they have two major drawbacks:

(1) The difficulty of estimating the errors due to incorrectly accounting for the clumping of the foliage in the canopy and this can arise when $\Omega(\theta)$ is incorrectly estimated. For example, $\Omega(\theta)$ is sometimes set to 1, to represent no clumping, which results in an under-estimation of LAI.

(2) The difficulty in distinguishing woody biomass from foliar biomass, which arises if no information is available to determine the ratio of WAI to PAI. As a result, measurements of PAI are sometimes used as a proxy for LAI; this often leads to LAI being overestimated. Gower et al. (1999) found, using LAI and PAI data from 13 different species of evergreen trees, that LAI is greater than PAI by 12% on average, with the smallest and largest increases being 3% and 34% respectively.

If $\Omega(\theta)$ and the ratio and WAI to PAI can be correctly estimated, these indirect ground measurements of LAI can compare well with direct ones (MacFarlane et al,

2007) when used in small scale (e.g. site level) studies. In-situ measurements are routinely made only at a limited number of sites worldwide and typically few are made annually per site (e.g. 10-20). Therefore, the only way to obtain frequent global measurements is to measure LAI from space.

INFERRING LAI FROM SPACE

Since 2000, LAI data inferred from space have been made using the MODIS sensor on NASA's TERRA/AQUA satellite. Data are available from the Oak Ridge National Laboratory website (http://daac.ornl.gov/cgi-bin/MODIS/GR_col5_1/mod_viz.html), every 8 days across the globe for every 1km×1km land area. The fraction of Absorbed Photosynthetically Active Radiation (fAPAR) can also be inferred from MODIS, which is closely related to LAI (Law & Waring, 1994). Collection 5 LAI data or products (as they are commonly known) are derived from the latest version of the algorithm used to convert light reflectance (what the sensor measures) into LAI, and was first released in 2007. As stated in subsection 5.1.1, LAI is a dependent of many different factors, such as light interception, photosynthesis, transpiration, respiration and water interception, but also leaf angle and clumping; these relations are all incorporated into the algorithm. The algorithms used to making indirect in-situ ground and satellite measurements of LAI are based on different types of observations. For in situ methods measurements of gap fraction are made, therefore LAI is based on the leaf area to a large extent. However, measurements from space are reflections of solar radiation, and these relate to the greenness of the vegetation, i.e. how photosynthetically efficient the leaves are, and not simply to the number of leaves (Myeni et al., 1997). In other words, two sites with the same number of leaves but different satellite LAI values because the photosynthetic properties of the trees at the two sites may be different. Thus the satellite LAI is more of a radiation measure, hence it is potentially a mismatch to vegetation model and to ground based observations. This is important because, while satellites are useful in that they offer frequent coverage of the entire earth, what the satellite sensor actually measures is very different what is measured from the ground. Thus, direct comparison of the two types of LAI measurements need to be made with caution. Furthermore LAI measurements made from space have a different resolution to those made on ground. In situ measurements are made at a resolution of 10-30 m^2, whereas from space they

are typically at a resolution of 0.5 – 1 km^2 – see Yang et al. (2006), Bacour et al. (2006) and Baret et al. (2007). Repeated in situ measurements are sometimes made to give an idea of the spatial distribution of LAI over a 0.5 – 1 km^2 grid cell (De Kauwe et al., 2011). In such circumstances, direct comparisons of satellite and ground measurements can be valid from a resolution point of view.

Figure 5.1, taken from Sea et al. (2011) and Ganguly et al. (2008), shows the temporal profile of LAI from four sites with varying vegetation types, including two savannah forests in Australia and Zambia (panels a and c, respectively), and a grassland and deciduous forest in the USA (panels b and d, respectively). For panel a, data from two versions of MODIS LAI are shown, referred to as collections 4 and 5 (or MC4 and MC5) as well as ground LAI data. Panels b to d show LAI from MODIS collection and another satellite sensor called AVHRR (Holban, 1986), as well as ground LAI data. Figure 5.1a is discussed in the next subsection, but for the other subplots it is interesting to note that there is in general good agreement between the MODIS and AVHRR LAI, with both also comparing well with the ground based LAI data. For all the plots, it is worth noting to note that for these four sites the temporal trajectories of LAI are cyclic in nature. This is expected due to varying rates of growth of the foliage during the different seasons.

As well as MODIS and AVHRR, LAI is currently inferred from space by a number of other different satellite sensors, including MERIS (Bacour et al., 2006) and CYCLOPES (Baret et al., 2007), but various validation exercises suggest that MODIS LAI products compares best with ground-based measurements than other LAI satellite products. We review these validation exercises next.

VALIDATION OF SATELLITE MEASUREMENTS OF LAI

Garrigues et al. (2008) compared the performances of LAI derived from four different satellite sensors, namely MODIS, ECOCLIMAP, GLOBCARBON and CYCLOPES, using ground-based observations of LAI at 41 different sites across the globe. The sites varied in terms of land-use (forests, croplands, grasses, etc…) with LAI values varying from 1 to around 6 m^2/m^2. Comparisons were made between 2001 and 2004, and used the internationally recognised VALERI framework (Morisette et al., 2006) which was established to promote consistency in studies where satellite observations

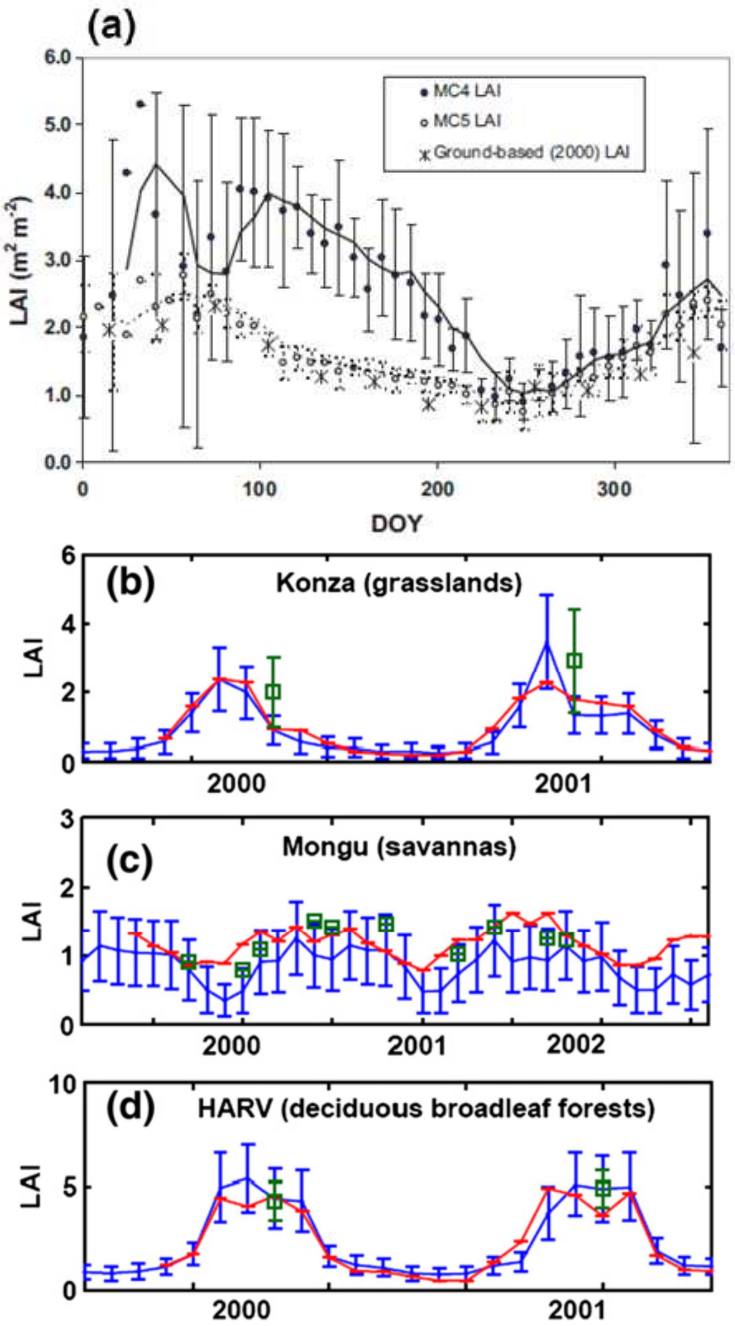


Figure 5.1 The temporal profiles of LAI for four different vegetation types which include a savannah woodland in Howard Springs, in the Northern territory of Australia (panel a), grasslands at the Konza site in Kanzas, USA (panel b), a savannah at the Mongu site in Zambia (panel c), and a deciduous broadleaf forest at the Harvard forest in the USA (panel d). Panel (a), taken from Sea et al. (2011), shows the seasonal variation of LAI for 2000 to 2006 with the lines corresponding to 24 day moving averages of mean values for 2000 to 2006 and the error bars representing +/- 1 standard deviation; MC4 and MC5 refer to collections 4 and 5 of the LAI data from the MODIS sensor of the TERRA satellite. For panels (b) to (d), taken from Ganguly et al. (2008), the blue, red and green lines/points correspond to AVHRR, MODIS and ground LAI with the error bars refer to the standard deviation associated with the data.

of land variables (such as LAI) are validated against ground observations. The authors found that the MODIS and CYCLOPES LAI data most realistically depicted the spatial variation of LAI at the continental scale and the temporal variation over most vegetation types. Despite good overall agreement with most vegetation types, the MODIS LAI tended to over-estimate the ground-based LAI for Evergreen needle leaf forests (the forest type used in this study) by around 1.5 m^2/m^2. However, a more recent version of the algorithm that converts the reflectance from MODIS to LAI is available (collection 5), whereas the previous version (collection 4) was used in Garrigues et al. (2008). Collection 5 provides a substantial improvement in the quality of LAI derived from MODIS. It also showed a better accuracy than the CYCLOPES LAI when compared to ground LAI data. Fang et al. (2012) found that MODIS collection 5 LAI produced a bias of -0.13 m^2/m^2 when compared to ground measurements of LAI, whereas the bias for CYCLOPES LAI was -0.65 m^2/m^2.

Other authors have studied the relation between MODIS and ground based estimates of LAI De Kauwe et al. (2011) compared MODIS collection 5 LAI with ground-based measurements using the VALERI protocol, scaled to the same resolution as the MODIS data. This study was carried out using observations from May 2006 for a site covering an area of 11km×11km in Oregon, U.S.. The authors found very good agreement (a positive bias of around 0.1 m^2/m^2), although the MODIS LAI tended to under-estimate values of the ground observations greater than 2 m^2/m^2 by around 1 m^2/m^2 on average. Moreover the spatial distribution of MODIS LAI broadly matched that of ground measurements, and collection 5 MODIS LAI offered a significant improvement on collection 4 temporally and spatially. One shortcoming of this study was that Plant Area Index (PAI) was used as an approximation to LAI for the ground measurements, and the differences between PAI and LAI were not properly quantified. As stated previously, Gower et al. (1999) found that that PAI is higher than LAI by 12% on average. Hence, although the MODIS LAI tended to under-estimate ground-based PAI for high values, it is likely to be estimating ground-based LAI more accurately. Although as a consequence, lower values of LAI may be fractionally over-estimated.

Sea et al. (2011) compared MODIS collection 5 LAI to ground based LAI during September 2008 at a set of sites in central northern Australia, which mainly consisted of Evergreen, Eucalyptus and Aracia forests. The VALERI procotol was

also used for this study. The fraction of LAI to PAI was estimated using the CAN_EYE software package (Baret & Weiss, 2004). They found very good agreement between the MODIS and ground-based LAI measurements, with a positive bias of roughly $0.1m^2/m^2$. In particular, the authors found that the most recent version of the MODIS LAI products (collection 5 or MC5) compared much for favourably with the previous version (collection 4 or MC4). This can be seen in figure 5.1a, where the MC5 LAI is not only much closer to the ground LAI data compared to the MC4 LAI but there is far less inter-annual variation, represented in figure 5.1 by the smaller error bars, which is also more consistent with the ground LAI data. It is worth noting here though that the ground LAI measured here was 2 or less for all the sites.

McColl et al. (2011) compared MODIS LAI to ground-based LAI scaled up to the same resolution as MODIS in an 84,000 km^2 area of New South Wales in Australia, during November 2006 and January, February and September 2010. The region has a diverse climate ranging from semi-arid to alpine, and the land was used mainly for grazing and crops. The MODIS LAI observations showed good agreement with ground-based measurements, with a mean difference of 0.35 m^2/m^2 and a standard deviation of 0.82 m^2/m^2. Although this was a useful study, the LAI values were predominantly small (less than 2 m^2/m^2) due to the type of land-use. Also, this validation study did not adopt the VALERI protocol.

These studies suggest the collection 5 MODIS LAI observations offer a realistic representation of the spatial and temporal variability of LAI, and compare well with ground observations. Further reasons for using MODIS LAI are:

(i) It offers a long-term time-series of LAI data, since 2000. In contrast CYCLOPES LAI, which as stated above is the regarded along with MODIS as being most accurate when compared to ground observations (Garrigues et al., 2008), is only available up to 2003 (Weiss et al., 2007). Therefore even if there was a stronger argument to choose CYCLOPES LAI over MODIS LAI, the gain in knowledge from this chapter could not be applied for those who wishes to use LAI after 2003.

(ii) With the exception of the MODIS LAI, it was impossible to be able to access the LAI from other satellite products such as CYCLOPES LAI, ECOCLIMAP LAI, GLOBCARBON LAI. Therefore even if there was a preference to use

another LAI dataset, it is not possible or at least very difficult to be able to access the actual data.

(iii) MODIS LAI has been used in a lot of research, and perhaps more than the other LAI satellite products, which may be for the reasons given in (i) and (ii) above. In the critical review of assimilating satellite LAI/FAPAR data into models within subsection 5.1.4, there are references to five studies which have used MODIS LAI are given, but there are many more such as Leuning et al. (2008), Li et al. (2009) and Clark et al. (2008). Therefore, even if there was a more compelling reason to use another type of LAI dataset, the fact that it is used by the research community in a non-insignificant way means that there is great value in the knowledge gained from this chapter to those who have or will use the MODIS LAI data for their research.

5.1.4 Assimilating satellite observations of LAI

Assimilating observations into models is better than simply comparing model estimates with observations of LAI because we are able to incorporate information on how confident we are that the observations and corresponding model estimates are correct (see chapter 2). These are referred to as the observational and model uncertainty and are often expressed as probability distributions, for example a Normal distribution with zero mean and a specified variance. However DA is only of real use if it results in an improvement of the model estimates of a quantity of interest (in this case NEE) against independent observations of that quantity.

In this subsection we focus on the best way to assimilate satellite observations of LAI. In particular, we address:

(i) Whether other studies performed DA using LAI observations from MODIS or other satellites, and if this resulted in an improvement to the estimate of interest (e.g. NEE) when compared to independent observations.

(ii) What model and DA scheme should be used.

(iii) The problems which have not been addressed when assimilating satellite LAI observations, and how these problems should be overcome.

(i) CRITICAL REVIEW OF ASSIMILATING SATELLITE LAI/FAPAR/REFLECTANCE DATA INTO MODELS

DA studies involving fAPAR measurements are also included here as fAPAR, like LAI, is an indicator of healthy vegetation (Kaminski, 2011) and is used to estimate Gross Primary Production (GPP) from satellite data. It is also a much more stable quantity to estimate because it is closer to what the satellite actually measures, i.e. reflection of solar radiation.

Spadavecchia et al. (2008) used an EnKF to estimate the parameters and states of a version of the DALEC model which included water dynamics. Data used in the DA process included ground observations of NEE and LAI and MODIS collection 4 LAI data, all collected during 2000-02. It was difficult to determine if assimilating the MODIS LAI improved estimates of NEE since the ground observations of NEE were also assimilated. The post-assimilation LAI was found to significantly underestimate the MODIS LAI, though this was probably due to the large uncertainties prescribed to the MODIS LAI, resulting in them having little influence. A weakness of the study did not say how the MODIS LAI uncertainties were determined, and did not justify the 10% error value ascribed to the ground LAI observations.

Jarlan et al. (2008) described the assimilation of MODIS LAI collection 4 was assimilated into the CTESSEL model (the land surface model of the ECMWF model), using a 2D-Variational DA scheme, as part of the GEOLAND project. This was found to improve the model's estimate of LAI and aboveground biomass when compared with ground observations, and to lead to significant improvements in the simulated phenology of the vegetation. The authors also concluded that LAI is an important driver for estimating NEE. However, a limitation with this study is that, since collection 4 of the MODIS LAI measurements was used, no estimate of the uncertainty was available, as opposed to the collection 5 MODIS LAI (used in this chapter) which includes standard deviation values. Therefore the uncertainties had to be estimated, and the authors treated them as fixed at 1 m^2/m^2.

In another study, Demantry et al. (2007) assimilated 2 years (2000-01) of MODIS LAI into the ORCHIDEE dynamic global vegetation model. This resulted in shortening the modelled growing season and consequently reduced the estimates of GPP and NPP by 5% and 3% respectively. There was a 25% improvement in estimating GPP, due to the assimilation of MODIS LAI, when compared to in situ

GPP measurements. Again, because collection 4 of MODIS LAI was used, the uncertainties had to be estimated, but, unlike Jarlan et al. (2007), the authors here used the quality control (QC) values (which give an indicator of the retrieval quality, e.g. if there were clouds present). This is erroneous as even values with high QC values (i.e. retrieval is of a high quality) may still have large uncertainties. The DA method employed was very simple, and was given by: LAI(c) = QC×LAI(d) + (1–QC)×LAI(m), where LAI(c), LAI(d) and LAI(m) stand for the corrected, the data and the model estimates of LAI. Finally, although assimilating MODIS LAI improved the GPP estimates using in situ measurements, there was an enormous difference in scale between the model grid cells (10^4 km²) and the in-situ GPP measurements (1 km²). This is wrong because it assumes that there is no correlation between neighbouring pixels of ground-based GPP measurements.

Dente et al. (2008) assimilated LAI from the ASAR and MERIS sensors on the ENVISAT satellite into a crop model, using a variational DA scheme, in order to improve estimates of parameters. The authors found a 25% reduction in the errors of wheat yield estimates when LAI was assimilated, compared to when no assimilation took place.

Sus et al. (2012) assimilated MODIS 250m vegetation index (VI) data into SPAc, a crop version of the SPA model (Sus et al., 2010), using the Ensemble Kalman filter. The study was unique in that it was the first to assimilate remote sensing time series data of field patches into a crop model simulating agroecosystem C exchange, where 'agroecosystem' refers to agricultural ecosystems excluding managed grasslands as defined by Kucharik & Twine (2007). Before being assimilated into the crop model, the modelled LAI were converted to VI using an empirical relationship involving an exponential function. In total, MODIS VI data from 104 field patches from 2000-2006 were assimilated, with the flux tower positioned in one of the fields providing a validation of the post-DA modelled NEE. The purpose of the study was (i) to provide a framework for estimating cumulative NBP and other useful quantities such as sowing dates and crop yields over a wide area, (ii) to also quantify the spatial variability of NEE, and (iii) to quantify the cumulative NEE over the entire period For (i), there was reasonably good agreement between the post-DA NEE and the observed NEE from the flux tower in terms of magnitude and seasonality, with the post-DA sowing dates also agreeing well with figures reported by farmers; the

agreement with observations for crop yields was also significantly better when the MODIS data were assimilated compared to when they were not. For (ii), averaged over the 104 pixels, the annual cumulative NBP was positive for soy-bean crops ($13 gCm^{-2}yr^{-1}$) and negative for maze ($-58 gCm^{-2}yr^{-1}$). Finally for (iii), the spatial variability was found to range from around 7% to 18%.

Doraiwamy et al. (2004) used MODIS surface reflectance data to estimate the parameters of a radiative transfer model for multiple 250m by 250m pixels, which in turn was used to predict LAI on a daily basis for each of the pixels. The modelled LAI was aggregated for all of the tiles to provide pseudo data in order to calibrate a coarser resolution scale for the ARS climate crop model (Muchow et al., 1990). Not only did the simulated LAI, from the radiative transfer model and calibrated with the MODIS data, fit the ground based LAI very well but the crop yield predictions of the crop model were improved from using the MODIS satellite data.

Knorr et al. (2010) estimated the parameters of BETHY (a terrestrial biosphere model) by assimilating two years of fAPAR measurements for seven sites simultaneously from the MERIS sensor of ENVISAT using a 3D-variational scheme. The estimated parameter set was used for all seven sites, which include boreal, temperate, humid‐tropical, and semiarid climates. They authors found that not only did the uncertainty estimates for 10 of the 38 optimized parameters were each reduced by at least 10%, but using the estimate parameter set, the model was able to reproduce the observed fAPAR measurements for the six sites where ground measurements were available. However, the modelled fAPAR, using the estimated parameter set, also under-estimated the observed fAPAR during the growing parts of the year for three of the sites investigated, including Loobos, one the sites used in this chapter.

Kaminski et al. (2011) carried out an almost identical study to Knorr et al. (2010), using the same model and DA scheme, but assimilating ground-based measurements of atmospheric CO_2 as well as MERIS fAPAR. The authors found that assimilating both streams of data improved the fit of the modelled NPP to ground measurements, more than just assimilating fAPAR ones. However, others believe that assimilating more than one type of observation can be wrong. Rayner (2010) assimilated in situ measurements of fAPAR and flux measurements of CO_2 and evapotranspiration, and found that while assimilating fAPAR improved the fit of the modelled to the observed NEE, when the fluxes were also assimilated the fit was not

as good. He concluded that the two data types conflicted with one another, but also recognised the model did not represent the processes well and this may have biased the results. This problem may also been do with the mis-specification of the uncertainties of either or both observations being assimilated, or the accuracy of the observations being used for validation.

An interesting study by Quaife et al. (2007) assimilated bidirectional reflectance factor (BDRF) MODIS data into the DALEC model. BDRF was used because is much closer to what the satellite sensor actually measures compared to satellite measurements of LAI, hence it is much easier to quantify the observational error. The authors found a better fit to the NEE in situ observations for a test site in Oregon, after assimilating BDRF, but for the wrong reasons. GPP and total respiration (Re) were over-estimated by approximately the same amounts, and due NEE being the difference between GPP and Re, the estimates of NEE were not over-estimated. The process of assimilating BDRF into DALEC involves the observational operator (which relates the observation to the model state) becoming very complex, since it is based on a radiative transfer model. Most ecosystem modellers who want to use satellite observations are likely to prefer the directly relevant LAI product.

(ii) THE MODEL AND DA SCHEME

In this chapter we use a simple model, namely the Data Assimilation Linked ECosystem (DALEC) model (Williams et al., 2005), rather than a complex one such as SDGVM. This is because most complex models were not designed for DA; we require a model which is simple enough for computationally efficient DA but complex enough to represent the biophysics and essential plant processes (Disney & Quegan, 2007). We employ DALEC here as it was created to be used in collaboration with a DA scheme. It was first used in Williams et al. (2005) who used an Ensemble Kalman Filter (EnKF) DA scheme and later in Fox et al. (2009) where a number of different DA schemes (including the EnKF, the Metropolis algorithm and the Genetic algorithm) were employed. DA schemes fall into two categories: (i) state estimation and (ii) parameter estimation (see chapter 2 for details of the difference between the two). For this chapter, state estimation was performed because we are interested in how assimilating MODIS LAI observations affects modeled NEE. Therefore we require the parameters to be predetermined and a description about how this was

achieved is outlined in subsection 5.2.2. For state estimation, four commonly used DA schemes are the Kalman Filter (KF), 4D-Var, the EnKF, and the Particle filter (PF). The KF cannot be used here as it requires the model to be linear (see chapter 2 for definition) whereas DALEC is non-linear. 4D-Var is not used as it requires the model to be differentiated which, although possible for DALEC, is complicated. The remaining two, the EnKF and the PF, can both operate with a non-linear model. We adopt the EnKF here due to its wide use, particularly in terrestrial ecosystem modelling (e.g. Williams et al., 2005, Quaife et al., 2007) and its ease of implementation.

For state estimation, the EnKF works by correcting the trajectories of each of the model states when an observation becomes available. In essence the trajectory will move closer to the observation if the uncertainty measure specified for the observation is less than that of the corresponding model state. The EnKF carries out parallel model runs, called ensembles, which all get updated when observations become available; at any one time point, the spread of these ensembles gives an indication of the uncertainty of the model state vector before and after assimilation of the observation. A vector of error terms (or noise terms as they are sometimes called), generated randomly from the error distribution for the model, is added to the state vector of each ensemble after the execution of the model (at each time point) in order to incorporate model uncertainty. A vector of noise terms is also added to the observations for a similar reason. More specific details of how the EnKF operates are given in subsection 3.2.3. A potential limitation of the EnKF is its assumption that the vector of noise terms for the model and observations should be generated from the Normal distribution. If the model is non-linear then we would find that this noise vector is transformed to another distribution once inside the model (Quaife, 2011 [personal communication]). However the EnKF forces the distribution of the noise vector to be Normal again after observations have been assimilated. Another limitation of the EnKF is to do with its ensemble aspect. Due to the Normal distribution of model noise terms, the trajectory of the ensembles can go to areas of the state space which are physically impossible. Since the DALEC model has only two simple non-linear components, it is unlikely that these problems will be encountered (Quaife, 2010 [personal communication]).

(iii) PROBLEMS WHEN ASSIMILATING MODIS LAI OBSERVATIONS

When we perform DA to do state estimation using satellite observations of LAI (e.g. from MODIS), there are various issues relating to the actual observations which have not been addressed in any previous DA study:

(a) For many sites the MODIS LAI observations are very noisy, i.e. observations can vary by around 2–4m^2/m^2 in only 1 or 2 weeks, which is unrealistic.

(b) For MODIS collection 5 LAI data, a measure of uncertainty (represented by a standard deviation value) accompanies each LAI observation. For small values of MODIS LAI (i.e. <1 m^2/m^2), this s.d. value can be unrealistically small. In addition, the errors are assumed to be symmetric, which is quite likely to be false.

A further issue is that for earth observation products of LAI only 'green' LAI is targeted, rather than total LAI which is what is assumed when determining LAI from ground measurements and when using it models. In other words, if the foliage is non-green, this may not be registered or completely registered as foliage. In this way, the EO interpretation of LAI is how photosynthetically active the leaf rather than the actual area of the leaf, resulting in the actual LAI being under-estimated. This is an issue which all EO products of LAI suffer from, and is a limitation. Validation with ground based LAI data is therefore essential to ensure that the EO data is trustworthy.

Another point worth making here concern issues of spatial scale. Unless the site is homogeneous the effective LAI, defined as LAI uncorrected for clumping or the amount of woody material present (De Kauwe et al., 2011), will be lower than the true LAI. Therefore it is important to relate the scale of the LAI data used to what is expected of the NEE data. NEE is only measured at one point for a particular site, and this is usually near the centre of the central pixel of the 49 1km by 1km pixels where MODIS LAI is given for each site. However, due to the existence of a flux footprint (see the third to last paragraph of the subsection 5.2.4), which would be expected to cover most of the 1km by 1km pixel where the flux tower is located, the scale of the MODIS LAI data approximately corresponds to the scale of the ground based NEE measurements. However, the analysis of this chapter would be improved if NEE was formally scaled up to the same spatial scale as the MODIS LAI. Thus, this is a limitation of this work, but it also a limitation of every single other published study where EO products like LAI have been used to make inferences on other quantities (such as NEE) involving comparison with ground based data of a different measure-

ment scale to the EO data.

Returning to points (a) and (b) above, we investigated whether modifying the MODIS LAI data in order to overcome these limitations improved the model fit of the post-DA NEE to in-situ measurement of NEE. With respect to (a)-(b) above, the MODIS LAI data were processed as follows:

(a) To reduce the temporal noise of the LAI while maintaining its signal, the MODIS LAI observations were smoothed. We adopted a simple 7 point moving average, rather than a more complex smoothing technique so as the method could be easily applied to other situations.

(b) We increased the small s.d. values of the MODIS LAI by an amount depending on the size of the s.d. value to begin with;

In making these changes, the aim was to make the MODIS LAI data and uncertainties more believable, whilst trying to preserve the features of the MODIS LAI as far as possible.

5.1.5 Aims and layout of chapter

This chapter answers the following questions: (1) Does assimilating MODIS LAI observations improve the fit between the modelled and observed NEE? (2) Does processing the LAI dataset to make it more realistic before assimilating it into DALEC improve the fit of the modelled to the observed NEE? In the methodology section (5.2) we describe the model (DALEC), the EnKF scheme used, the three sites used and why they were chosen, and how the MODIS LAI and ground NEE datasets were obtained for each of the sites. Details are also given on how the MODIS LAI was processed and how changes were quantified in terms of the fit of the resulting trajectory of post-assimilation NEE to the trajectory of ground NEE observations. The results are displayed in the section 5.3, and discussed in section 5.4. The conclusions are given in section 5.5.

5.2 Methodology

5.2.1 Site Selection

We focussed on evergreen sites to avoid the extra implications in both model and data

when dealing with deciduous forests (for example the inclusion of a labile carbon pool), and to more clearly bring out the issues in using LAI.

There are 163 sites across the world classified as evergreen forests where MODIS LAI and ground NEE observations are both available (ORNL, 2001). A subset was chosen under several criteria:

(i) Specific site information was required for each site in order for the model to operate, e.g. Leaf Mass per Area, Foliar Nitrogen and the initial conditions of the C pools;

(ii) Calibrated values of the parameters (which are site-specific) are required, perhaps from past studies;

(iii) The MODIS LAI values from the different sites should ideally exhibit different characteristics, e.g. different shape of the LAI trajectories, different amplitudes, etc …;

(iv) Ideally the sites would come from different evergreen biomes. Three sites matched the first three criteria, namely Loobos (Netherlands), the Metrolius young ponderosa Pine, Oregon (hereafter referred to Oregon) (U.S.) and Tharandt (Germany). The fourth criterion was partially met because two are from Europe and one is from the western part of the USA. Figure 5.2 shows ground photos of the each of the sites.

The Loobos site (52˚10' N, 5˚44'E) is located approximately 70km east of Amsterdam in the Netherlands. It is almost flat and is at an elevation of 52m. The forest is classified as an Evergreen Needleleaf forest, with the dominant tree species being Scots Pine (Pinus sylvestris). It is a mature forest, around 80 years old in 2000. In 2000 there were 620 trees per hectare with a mean tree diameter at breast height of 27cm, and a mean height of 16.2m. The forest has a Mediterranean climate with a mean air temperature of 10˚C and an annual precipitation of 786mm.

The Oregon site is located in a Research Natural Area (44˚26’N, 121˚34’W) roughly in the centre of Oregon state, U.S.A. The area is nearly flat and at elevation of approximately 1200m. The forest is classified as an Evergreen Needleleaf forest with with the dominant tree species being Ponderosa pine (Pinus ponderosa). The site was completely cleared in 1978, but it has grown back since then. In 2002, there were 431 trees per hectare, with a mean diameter at breast height of 11.3 cm and mean height 4.3m. The site is in a semiarid region that experiences warm, dry summers and wet,

Figure 5.2 Photos of the three sites: Loobos (left), Oregon (centre), and Tharandt (right). All photos were obtained from http://fluxnet.ornl.gov/.

cool winters.

The forest at the Tharandt site (50˚58'N, 13˚34'E) is located 20km SW of Dresden, Germany. The topography is gently sloped and is at an elevation of 380m. The forest is classified as an Evergreen Needleleaf forest with the dominant tree species being Norway Spruce (Picea abies). It is a mature forest of around 110 years old in 2000. In 2000 there were 480 trees per hectare with a mean tree diameter at breast height of 33cm, and a mean height 26m. The forest has a Mediterranean climate with a mean air temperature of 7.5˚C and an annual precipitation of 820mm.

For Loobos and Tharandt, the MODIS LAI data varied between $1m^2/m^2$ and $4m^2/m^2$ during the winter and summer months and were in general very noisy. For Oregon, the trajectory of MODIS LAI data was fairly smooth varying between around $1m^2/m^2$ and $2m^2/m^2$. The MODIS data, along with the in situ data, are described in greater detail subsection 5.2.4. The parameter values available for the three sites (Williams et al., 2005; Fox et al., 2009) were calibrated against data from 2000-2002, and so these years were the obvious choice of time period to adopt for the study.

5.2.2 The DALEC Model

The evergreen DALEC model (see chapter 2 for full description) is a simple box model that tracks the journey of carbon from photosynthesis through to C pools and respiration fluxes, working on a daily time-step. Figure 5.1 shows it in diagrammatic form. It has five carbon pools namely C_f (foliar), C_r (fine roots), C_w (woody biomass), C_{lit} (litter), and $C_{som/cwd}$ (soil organic matter & coarse woody debris). There is an additional pseudo-pool for GPP (Gross Primary Production), which represents carbon accumulated from photosynthesis. There are a total of ten fluxes which connect the C

pools and also provide a way for C to get into the system (via GPP) and out of it (via respiration). The values of the eleven parameters and five initial conditions are fixed and shown in table 5.1.

For Loobos and Tharandt, the parameters and initial conditions were estimated from site inventories and from local knowledge of the site (Fox et al., 2009). To assess the realism of these values for the Loobos site we note that, for 7 of the 11 parameters, the values used in this chapter agree very well with the means of estimates obtained by the participants of the REFLEX project (Fox et al., 2009). For the remaining 4, the values used here are well within the range of the estimates from REFLEX. In REFLEX, Tharandt was chosen because it was considered to be very similar to Loobos in terms of climate and vegetation, therefore the same parameter values as used in Loobos were used here too.

For Oregon the parameters were taken from Williams et al. (2005), who estimated them by nesting the EnKF within an optimization. The authors state:

> *This routine varied the unknown parameters and initial conditions to find the values that minimized the sum of squared differences of the innovations (i.e. the difference between model forecast/prediction and observations) for all available observations. This approach, in effect, undertakes numerous implementations of the EnKF with varied parameters and initial conditions, and identifies in which implementation the predictions require the minimum correction. This implementation is then assumed to have the optimal parameter set and initial conditions.*

This same parameter set was used by Quaife et al. (2007), who assimilated MODIS satellite data of surface reflectance to improve modeled estimates of NEE, GPP and total respiration, which was then compared against ground based measurements of these three data types. This chapter is different in that MODIS satellite LAI data are assimilated, but it is similar to Quaife et al. (2007) in that the same parameter set from Williams et al. (2005) is also used and the post-DA NEE states are validated against ground data. Thus for the same reasoning as Quaife et al. (2007) we deem it acceptable to use the parameter set estimated by Williams et al. (2005) for the Oregon site. A further important point to make here is that DALEC assumes one layer of canopy, whereas in reality the trees will contain multiple canopies. At the end of subsection 5.1.2 it is stated that in one sense this does not matter because the ACM gives very acc-

Symbol	Description	Units	Value (Loobos | Tharandt)		Value (Oregon)
T_d	Litter decomposition rate parameter	day^{-1}	4.41E-06		4.40E-06
F_g	Fraction of GPP respired autotrophically	-	0.52		0.47
F_{nf}	Fraction of GPP allocated to foliage	-	0.29		0.31
F_{nrr}	Fraction of GPP allocated to roots	-	0.41		0.43
T_r	Turn over rate of foliage	day^{-1}	0.0028		0.0027
T_w	Turn over rate of wood	day^{-1}	2.06E-06		2.06E-06
T_r	Turn over rate of roots	day^{-1}	0.003		0.00248
T_l	Turn over rate of litter	day^{-1}	0.02		0.0228
T_s	Turnover rate of C_{SOM} pool.	day^{-1}	2.65E-06		2.65E-06
E_t	Exponential temperature dependent rate parameter	-	0.0693		0.0693
P_r	Nitrogen use efficiency parameter	-	7.4		2.155
$C_{f,0}$ (Foliage)	Foliar carbon pool	gCm^{-2}	110	110	58
$C_{w,0}$ (Wood)	Wood carbon pool	gCm^{-2}	9200	12400	770
$C_{r,0}$ (Roots)	Fine root carbon pool	gCm^{-2}	180	102	102
$C_{lit,0}$ (Litter)	Fresh foliar and fine root litter carbon pool	gCm^{-2}	93	40	40
$C_{SOM,0}$ (Humus)	Soil organic matter & coarse woody debris carbon pool	gCm^{-2}	11000	9700	9897

Table 5.1 The description of the parameters and C pools of the DALEC model, together with the values used for the individual sites. Loobos and Vielsalm used the same parameters but the initial values of the C pools were different.

urate coarse scale estimates of GPP from the SPA model which does include multiple canopy layers. Nonetheless, the overall behaviour and response of the understory will be different to that of the canopy not just in terms of photosynthesis but also plant growth, respiration, and movement of carbon. This is a limitation of the DALEC model, but arguably not one which may bias the overall approximations of vegetation dynamics. For example, it has been shown that it can still give accurate NEE estimates at the spatial scale of a site (i.e. 1km by 1km area) (Williams et al., 2005).

5.2.3 The Ensemble Kalman Filter (EnKF)

The Ensemble Kalman Filter (EnKF) is a sequential data assimilation algorithm which like other DA algorithms is based on Bayes' theorem (Bayes, 1763). For terrestrial ecosystem models the EnKF has been successfully implemented for state and parameter estimation (Williams et al., 2005, Quaife et al., 2007, Fox et al., 2009). The formulation of the EnKF used in this study is described in detail by Evensen (2003).

The complete description of the EnKF is given in chapter 2, as so is not repeated here. However there are some details which need to be described here which are specific to the implementation of the EnKF for this chapter. These are the number of ensemble members to use and the specifying the uncertainty on the LAI from a forward run of the model (no DA).

For the number of ensemble members, we first note that the error covariance matrix of the distribution of ensembles at the analysis step of the EnKF converges to the error covariance matrix used in the Kalman filter (Evensen, 2009). Therefore it is desirable to have the number of ensemble members as high as possible, however not so high that the EnKF becomes too computationally expensive, which is a particular issue when the EnKF is used for large models. Two previous studies which used the EnKF with DALEC (Williams et al., 2005 and Hill et al., 2012) used N=200 and N=500, but for this chapter we use N=1500. The reason for this is that the model is simple and the relatively short dataset is assimilated, so taking N=1500 takes very little extra time and has the advantage that the distribution of ensembles more accurately resembles a Normal distribution.

As we are performing DA, as well as specifying the uncertainty associated with the LAI data being assimilated (see next subsection), it is also necessary to specify the uncertainty on the modelled LAI. For consistency with Fox et al. (2009) and Williams et al. (2005) which assimilated LAI data into the DALEC model using the EnKF, this chapter uses the same uncertainty on the modelled LAI as was used in those two studies; that is a standard deviation of 20% of the model LAI.

5.2.4 The MODIS LAI observations and ground NEE dataset

MODIS AND GROUND LAI OBSERVATIONS

The MODIS LAI product is given as pixels of size 1km×1km, and are given over the same area every 8 days. The data can be accessed publicly from the NASA MODIS website (http://daac.ornl.gov/cgi-bin/MODIS/GR_col5_1/mod_viz.html) and is labelled as MOD15A2 (Terra). The MODIS LAI is derived from an algorithm (hereafter referred to as the main algorithm) which uses a three-dimensional radiative transfer model (RTM). The RTM has surface reflectance of solar radiation (or simply 'reflectance') as its output and uses various biophysical parameters, such as LAI, as inputs. The RTM is inverted to find the parameters (LAI, etc…) that produce the best

fit between the modelled and observed reflectance (Knyazikhin et al., 1998, Myneni et al., 2002). To avoid multiple solutions (i.e. if there are multiple values of LAI, which produce near optimal fits of modelled to observed reflectance) a biome classification map is also used in retrieving LAI.

An empirical back-up algorithm is used if the main algorithm fails, which can result from:

(i) Low quality reflectance data: the best fit of the observed to modelled reflectance is calculated using an RMSE value which must be below a certain threshold, otherwise the data are considered low quality (Privette et al., 2002);

(ii) Algorithmic error, for example relating to geolocation, calibration or failures in the cloud mask (De Kauwe et al., 2011).

The back-up algorithm estimates LAI using a look-up table of vegetation indices for which LAI values are specified. However Yang et al. (2006) argue that this back-up algorithm results in poor quality estimates of LAI. For this reason it was not used in this chapter, and when the main algorithm failed we assume no MODIS LAI observation. Even when the main algorithm was used, the quality of the LAI retrieval was reduced slightly when clouds were present, so these observations were also removed.

For each of the three sites, the MODIS LAI observations were in principle available every eight days during 2000-2002 with the caveats above and subject no cloud problems. Figure 5.3 shows the MODIS LAI data for the three sites, along with the ground LAI data. Several interesting observations can be made from these three time series plots:

(1) The MODIS LAI observations appear higher for Loobos than for Tharandt, which might be surprising given that the sites are at the same latitude and have a similar climate. This can be seen by the observations peaking at a slightly higher level for Loobos, and around half of the Loobos observations are 2 m^2/m^2 or, whereas this is true for only a quarter at Tharandt. A possible reason for this might be that the leaves of the Norway Spruce trees at Tharandt are not as photosynthetically active than the Scots pine trees at Loobos (Myeni et al., 1997). This is because what MODIS actually measures is closely related to the level of photosynthetic activity of the foliage.

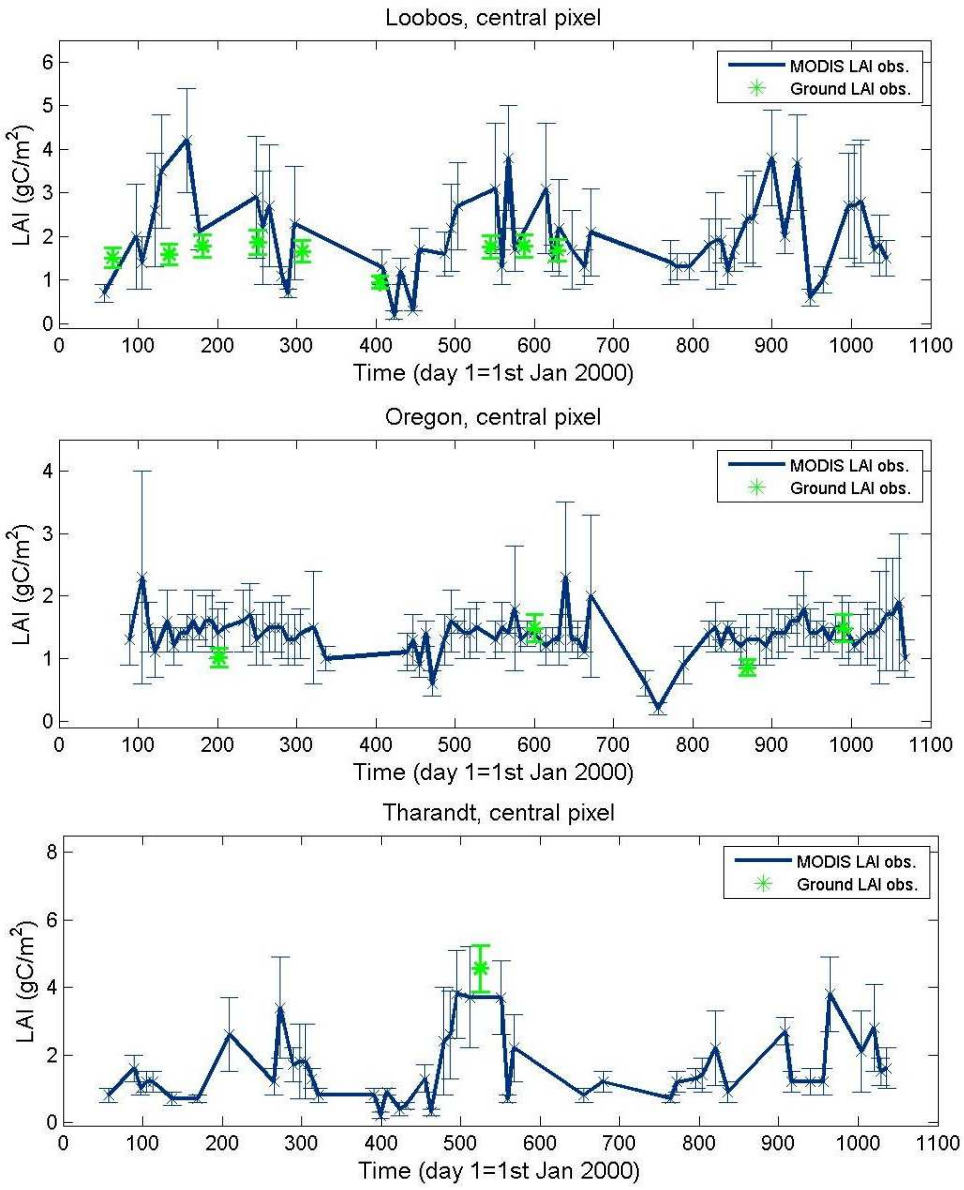


Figure 5.3 Time series of the MODIS LAI collections data for the three sites, Loobos, Oregon and Tharandt. The ground LAI measurements are also shown for comparison. MODIS LAI is available every 8 days but only those were included where the main algorithm in the Radiative Transfer model was used and where there was sufficiently low cloud cover (see the first part of this subsection). The error bars correspond to +/- s.d.. The s.d. values accompanied the MODIS LAI data, and the s.d. of the ground LAI was determined as a percentage, in this case 10%, of each data point, as done in Williams et al. (2005) and Fox et al. (2009).

(2) At day 500, where the only ground LAI data is available for Tharandt, the MODIS LAI data is different to this ground observation. However the difference is not great and the ground LAI data point is within the uncertainty interval of the neighbouring MODIS data.

(3) The LAI at Oregon generally lower (1-2 m^2/m^2) than at the other two sites (1-4 m^2/m^2)? This is due to the forest at Oregon being younger (22 years in 2000) than at the other two sites (80 and 110 years in 2000). The trees also average 4.3m in height, which is significantly lower than the trees at the other two sites.

(4) The peak to peak amplitude of the MODIS LAI data is much larger than that of the ground LAI for Loobos, where there are most ground data available compared to the other sites. This is likely to be caused by the fact that ground LAI data would include only trees and perhaps shrubs, but the MODIS data infers LAI for all vegetation types including bare ground and so a larger range in the MODIS LAI is expected.

(5) The MODIS LAI at Oregon is very flat at around 1.5 m^2/m^2 during most of the 3 year period, whereas the ground LAI vary between 1 and 1.5 m^2/m^2. This might be because, although there are fewer leaves during Spring, they are more photosynthetically active than during summer.

(6) There are few MODIS LAI observations during the winter months, for all of the sites, probably because there is more likelihood of the reflective data being low quality such as more cloud (Roy et al., 2006).

The effects of assimilating the MODIS LAI observations into the model were quantified by comparing the modelled NEE, after assimilating the MODIS LAI data, with the ground observations of NEE made at the same location as the MODIS and ground LAI measurements. This comparison was possible due to the existence of a flux footprint; since the wind is the carrier of CO_2, the CO_2 measured at the flux tower can originate from several kilometres away. Hence the NEE, or the CO_2 vertical flux, measured at the site is representative of NEE across an area sometimes several kilometres upwind from where the flux tower is located. The size and of the flux footprint can be inferred from the wind direction and wind speed, available from the fluxnet database.

MODIS LAI data from two adjacent 1km×1km pixels were also assimilated and the resulting post-assimilation NEE was also compared with same ground NEE measurements, even though these were not in the same pixel as the MODIS LAI. However the two extra pixels were chosen to be contained in the flux footprint at the time the measurements were made. Figure 5.4 shows the location of these pixels (blue) relative to the central pixels (red). For Loobos, the fetch measured 1.5–3.0 km and the

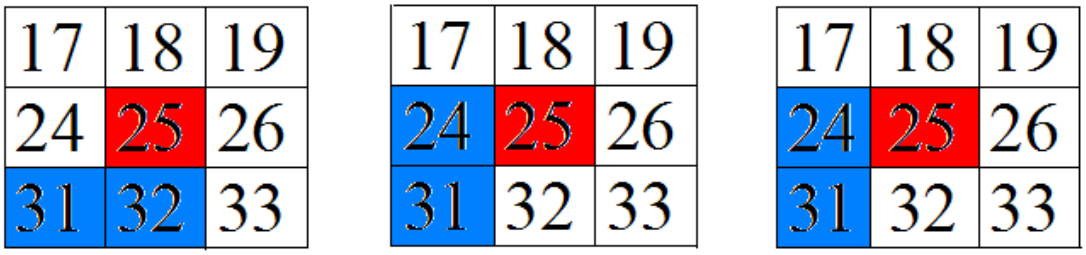


Figure 5.4 For each of the sites, MODIS LAI observations were used from the central pixel (red) and two neighbouring pixels (blue). The numbers relate to the numbering from 1 to 49 given for the MODIS LAI data at each of the 49 pixels at a given time point (Knyazikhinet et al., 1998, Myeni et al., 2002). The 9 pixels here are the central ones given in the 7 by 7 array of the 49 pixels.

wind direction was from the SW direction at the time the measurements were made, and since the flux tower is in the SE corner of the central pixel, pixels 31 and 32 were chosen. Oregon had a fetch of approximately 1-2km from the West/SW direction, thus pixels 24 and 31 were chosen. Tharandt had a fetch of 0.5km from the SW direction thus, given the flux tower is fairly close to the West of the central pixel, pixels 24 and 31 were chosen. These measurements of the flux footprint distance and wind direction were made at the same time as the LAI and NEE data. Note that the numbering of the pixels used in figure 5.4 is not 1 to 9 because the MODIS data is available for 49 pixels covering an area of 7km×7km, with pixel 25 as the central pixel.

GROUND MEASUREMENTS

Half-hourly measurements of NEE were provided from eddy-covariance instrumentation on flux towers at each of the three sites. The in situ LAI measurements were made by field scientists based at the sites using gap fraction methods as described in subsection 5.1.3; in particular the ground LAI data here only include foliage and do not include woody area index. These data were obtained from the Ameriflux database for Oregon (http://public.ornl.gov/ameriflux/Site_Info/siteInfo.cfm?KEYID= us.metoli us_yng.01) and from the GHG Europe database (http://www.europe-fluxdata.eu/newtcdc2/GHG-Europe_home/login.aspx for Loobos and Tharandt. The NEE measurements were converted to daily data by converting them from the units used for the eddy-covariance NEE measurements (µmol/m²/s) to those used in DALEC, but at the daily time-scale (gC/m²/day). When the half-hourly

measurements were not made (due to, for example, an instrument error or power failure) or the quality was not good enough, for example an insect interfered with the instrumentation, biasing the data). The measurements were estimated (or gap-filled using common terminology) using a statistical model. For each day the mean of all 48 half-hourly measurements was used to represent the daily measurement, as long as at least 75% of the measurements were not gap-filled. The daily in situ NEE and LAI data are also available from the Carbon Fusion website (http://www.geos.ed.ac.uk/carbonfusion/Reflex.html).

5.2.5 Processing the MODIS LAI datasets

As stated at the end of subsection 5.1.4 there are two main issues with the MODIS LAI data: (i) the data are noisy, (ii) the s.d. (error bar) values are unrealistically small, particularly for small LAI values. These two problems are clearly evident in figure 5.3.

For (i) we can see that at Loobos and Tharandt, MODIS LAI can change dramatically (1–2 m^2/m^2) in the space of only a week or two. Unrealistically small error bars are evident at all sites but particularly Loobos and Tharandt; for example, the error bar on the MODIS LAI observation near day 425 of Loobos.

To overcome these problems with the MODIS LAI we performed the following operations: smoothing the noisy observations and increasing the s.d. values (where necessary) for the small LAI values; decreasing the peak-to-peak amplitude of the MODIS LAI time series to be no more than 2 m^2/m^2; performing both operations.

OTHER STUDIES WHICH HAVE PROCESSED LAI DATA INFERRED FROM SPACE

The claim that the satellite data need modifying is backed up other studies, although all of these studies correspond only to smoothing the data. One of the first attempts to do this was Jonsson & Eklundh (2004), who developed the TIMESAT software for smoothing data. This is done by a choice of different algorithms, such as least square fitting, Savitzky-Golay filtering, and the use of harmonic functions and asymmetric Gaussian functions. All of these approaches broadly speaking achieve the same result, but individual approaches may be more applicable to particular datasets with specific needs about how the data should be processed. For example the Savitzky-Golay

filtering can capture subtle and rapid changes in the time series but at the cost of being sensitive to noise (Gao et al., 2008). All of the approaches weight the data according to some measure of uncertainty, for example the amount of cloud present which can cause bias and less certain estimation of the satellite product. The TIMESAT software have been used by a number of studies, most notably by Gao et al. (2008) whose authors include the team from Boston University responsible for producing the MODIS LAI/FaPAR products. Gao et al. (2008) used assymetric functions to smooth MODIS collection 4 LAI data with the weighting determined by the Quality Control (QC) information, where the QC information was based on for example whether the back-up algorithm was used in the radiative transfer model (RTM) or there was too much cloud cover for each data point. The authors found good agreement with the smoothed data points versus the highest quality unsmoothed data. Yuan et al. (2011) did a vertically identical investigation to Gao et al. (2008), apart from specific details such as the site, but used MODIS collection 5 LAI data. Interestingly, Yuan et al. (2011) found that the resulting smoothed LAI data agreed better with ground based LAI data, compared to the unprocessed data.

Quaife & Lewis (2010) offered an interesting alternative way of smoothing modelled surface reflectance variability, through inverting linear bi-directional surface reflectance (BDRF) models using MODIS satellite data, by imposing a smoothness constraint using Lagrangian multipliers. This is not yet fully operational, but offers a novel way of producing more regularized satellite products.

While there has been an interest in smoothing satellite products, there is been no research found on validating the prescribed data uncertainties, such as those which accompany the MODIS collection 5 LAI data, nor on the size of the peak to peak amplitude of the LAI signal. However accurate data uncertainties is essential where the data is used for data assimilation – certainly having extremely low uncertainties as observed with the MODIS data in figure 5.3 is unrealistic and will have a strong impact on the post-assimilation model estimates. In terms of the need to reduce the peak to peak amplitude of the MODIS LAI signal, we can clearly see that the ground LAI data show a low peak to peak amplitude as we might expect for an Evergreen forest (Alton, 2010 [personal communication]), whereas the amplitude of the MODIS LAI is noticeably higher. These arguments form the basis for the need for assessing

the impact of modifying of the MODIS LAI in terms of the s.d. values and the peak to peak amplitude of its signal.

CREATING THE PROCESSED DATASET

Figure 5.5 shows the result of this processing for the MODIS LAI for Loobos. An exact description is now given of how each of the processed datasets was created, i.e. how the smoothing, scaling of the LAI data and scaling of the uncertainties of the LAI data was carried out.

(i) The MODIS LAI data were modified in two ways: First a 5-point moving average was applied to the entire dataset which caused the trajectory to be smoothed. Although others have used a moving average to smooth the MODIS LAI data (e.g. Sea et al., 2011), it could be argued that a more sophisticated smoothing method should be employed here such as the Savitzky-Golay filtering as used in TIMESAT (Jonsson & Eklundh, 2004). However given the sufficient amount of MODIS LAI data used here, except for during the winter months, it was deemed appropriate that a moving average was adequate here. This has two main advantages over more sophisticated techniques: first it is simple to implement and therefore easy to replicate by others, and second it gives more or less the same result as if other smoothing methods had been used instead. Furthermore, the purpose of the smoothing done here was not for the results of this chapter to be implemented on a larger scale. Instead, we wish to demonstrate the biases introduced to the post-assimilation model states, particularly NEE, when some sort of processing is not implemented.

(ii) Next, if the s.d. value of a MODIS observation was less than $0.5 m^2/m^2$ it was changed to $0.5 m^2/m^2$. If the s.d. was greater than or equal to $0.5 m^2/m^2$ and less than $1 m^2/m^2$ it was set to $\min(1.5 \times \text{s.d.}, 1) m^2/m^2$, and if it was greater than $1 m^2/m^2$ it remained unchanged. The rationale for increasing the s.d. values in this way was simplicity, and it is noted that there are other perhaps more sensible ways increasing the s.d. values in a less ad-hoc way. This may have been preferable, but in truth the purpose of this transformation was not to provide 'accurate' estimates of the small s.d. values, but rather to observe the effect of increasing these s.d. values.

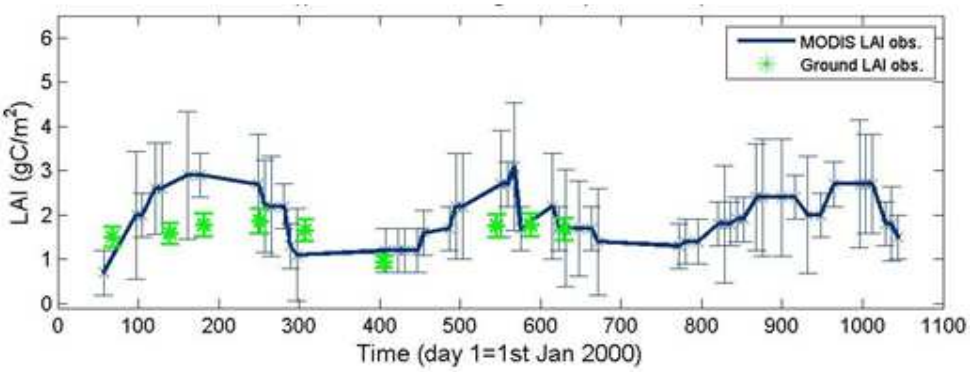


Figure 5.5 The processed MODIS LAI data for the central pixel of the Loobos site. The data was smoothed and the s.d. values for the smaller LAI values were increased.

It is worth noting here that the processing moves the data some way from their original positions and so it could be questionable whether these data should be used for DA. However, studies in the past which have processed data like this before also (for example Gu et al., 2006, Gao et al., 2008) have resulted in the same thing. Furthermore in the case of smoothing the data, it is important to state that the data points which have been moved away the furthest are predominantly those which were very close to zero due the dominant winter months of the year, and it is questionable whether this is actually true due the vegetation being evergreen. Finally, as we will show in the results, we find that this processing results in the post-assimilation modelled NEE being closer to the ground NEE observations in many cases.

The original MODIS LAI dataset and the processed datasets are hereafter referred to as:

'LAI-orig' to denote the original set of observations.

'LAI-smsd' to denote the dataset where the LAI was smoothed and the small s.d. values

5.2.6 Experimental Design

In total, 16 experiments (executions of the EnKF) were performed for each site. These are outlined in table 5.2. Pixel refers to whether the MODIS LAI from the central or one of the adjacent pixels was used (figure 5.4). The remaining columns correspond to what MODIS LAI dataset was assimilated, whether it was the unprocessed one (i.e. LAI-orig), or the processed one. For experiment 1, no dataset was assimilated, i.e. a forward run was carried out. For experiments 8 to 10 the MODIS LAI-smsd dataset

Experiment	Pixel	LAI-orig	LAI-smsd	LAI-ampl
1	None			
2	Central	Yes		
3	Adjacent 1	Yes		
4	Adjacent 2	Yes		
5	Central		Yes	
6	Adjacent 1		Yes	
7	Adjacent 2		Yes	
8	Central		Yes	Yes
9	Adjacent 1		Yes	Yes
10	Adjacent 2		Yes	Yes

Table 5.2 The list of executions of the EnKF carried out for each site, based on the unprocessed and processed MODIS LAI datasets which were assimilated. 'Adjacent1' and 'Adjacent2' refer the pixels adjacent to the central pixels (figure 5.4) whose MODIS LAI data were used.

and the ground NEE observations were assimilated for the three pixels. This demonstrated the assimilation of more than one dataset using the EnKF, but more importantly to helped understand what the modelled LAI should look like since the NEE being assimilated is also used to validate the post-DA NEE.

5.2.7 Assessing the results

The effect of assimilating each of the datasets was determined by computing the median and 95% tolerance interval (i.e. an interval with lower and upper bounds given by the 2.5^{th} and 97.5^{th} percentiles) of the absolute differences between the post-DA NEE and the observed NEE. To determine whether assimilating the unprocessed MODIS LAI observations improved the estimates of NEE, the medians and lengths of the tolerance intervals from experiments 2 – 4 were compared with those of experiment 1. Improvements in the NEE estimates occurred if there was reduction in the median and/or the length of the tolerance interval. Similarly, to determine whether assimilating processed MODIS LAI observations gave better estimates of NEE than assimilating the unprocessed ones (i.e. MODIS LAI orig), the medians and tolerance intervals for experiments 5 – 10 were compared with those of experiment 2 – 4.

The distribution of absolute differences of the post-DA and observed NEE was calculated for periods of 4 months in length rather than the whole 3 year time period. This is because the assimilation of the LAI observations has more influence on the post-DA NEE during the growing season of the year. Therefore the biggest improvements in the estimates of NEE from assimilating MODIS LAI are more likely to be seen during the May-August time periods. This can be seen clearly in figure 5.7, whose top two panels show the time series of LAI and NEE before and after assimilating the MODIS LAI data for the central pixel of Loobos. The third and fourth panels show the absolute differences for all nine of the 4-montly periods between the modelled NEE (red=before assimilation, black = after assimilation) and the observed NEE. Lastly the bottom panel summarises the information in the third and fourth panels as medians and 95% tolerance intervals for the four monthly periods during 2000-02.

To complement the sets of plots of the medians and percentile intervals, we also plot the trajectories of the LAI and NEE. As an example, we show such a plot after assimilating the unprocessed MODIS LAI (figure 5.6), which is the same as the top two panels of figure 5.9a but in an uncondensed form.

Included in the assessment of the importance of assimilating MODIS LAI we also assess how the uncertainty on the modelled states reduces after assimilation compared to the pre-assimilation modelled states. We do this by plotting the lengths of the lengths of the 66% tolerance intervals (equivalent to +/- 1 s.d.) corresponding to the pre-DA model states (i.e. the forward run) and post-DA states as two separate time series. We should see that the time series of tolerance intervals for the post-DA modelled NEE is consistently less than that of the forward run.

5.3 Results

5.3.1 Checking the assumption of Normality

One of the assumptions of the EnKF is that the distribution of ensembles at any time step (i.e. on any day) is Normal. This is checked by choosing nine of the days from the three year period at random and, and then for each of these days plotting the distribution of ensembles for each of the states was plotted. As an example, figure 5.8

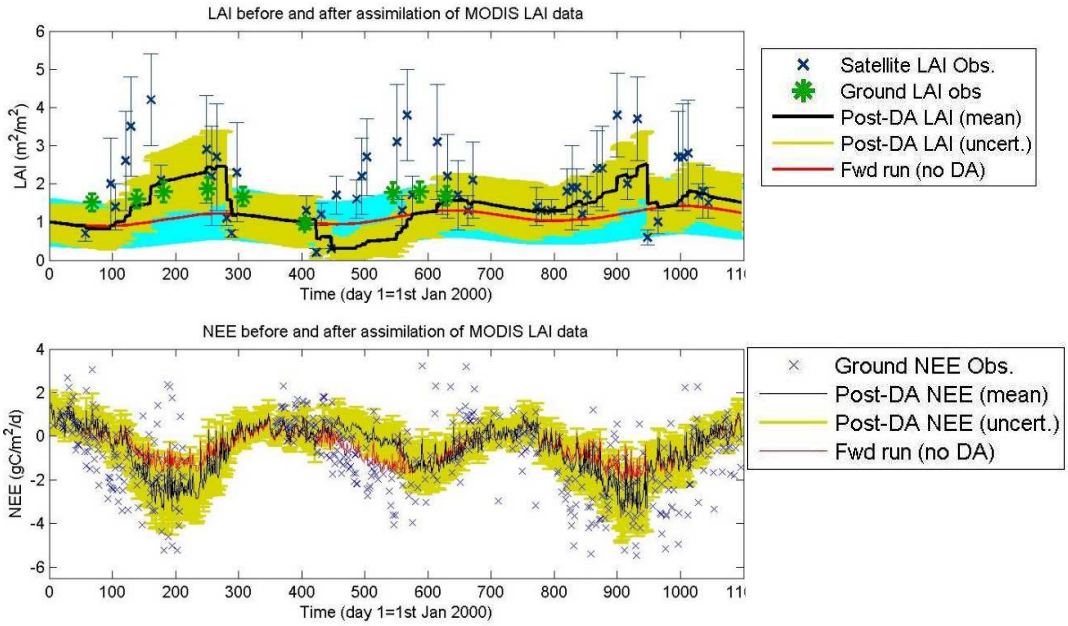


Figure 5.6 The two panels show the time-series of LAI and NEE before and after assimilation of MODIS LAI data from the central pixel of Loobos. The legend on the left corresponds to the top panel, which the right one is for the second panel.

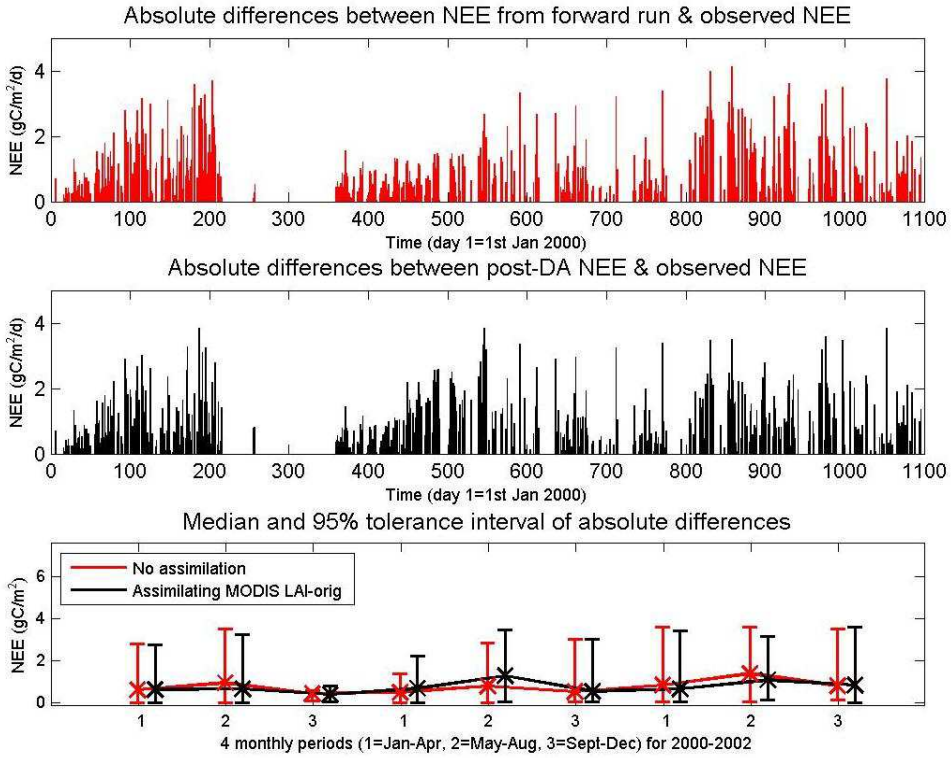


Figure 5.7 The top two panels show the absolute differences between the post-DA and observed NEE, and between the NEE from a forward run and the observed NEE. The third panel shows the same information as the first two except the absolute differences are summarised using the median (crosses) and 95% tolerance intervals (error bars) for each of the 4 monthly periods during 2000-02.

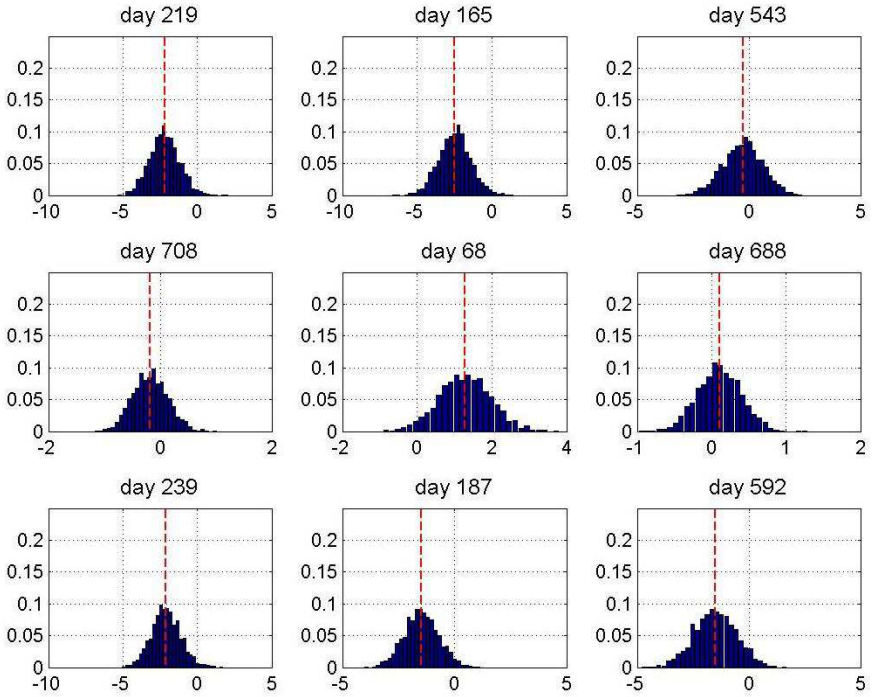


Figure 5.8 Histograms of the distribution of NEE from the 1500 ensemble members, for nine randomly chosen days, corresponding to the assimilation of the MODIS LAI data for the central pixel of Loobos. The red vertical dashed line indicates the mean.

shows these histograms of NEE after assimilation of MODIS LAI for the central pixel of Loobos. All plots are approximately Normal, and this was true for all elements of the state vector and for any randomly chosen set of nine days. Therefore the assumption of Normality is valid.

5.3.2 Assessing the impact of assimilating unprocessed MODIS LAI observations on estimates of NEE

In this subsection we show that for Oregon there is some evidence that assimilating the original MODIS LAI observations improves the fit of the modelled to the observed NEE, but for Loobos there is no consistent evidence that it makes a difference, and for Tharandt assimilation of MODIS LAI data actually results in a worse fit. The results were assessed by plotting the post-DA trajectories of the NEE and LAI alongside the trajectories of the forward runs and the corresponding observations (figures 5.9a and 5.9b) and by plotting the median and 95% tolerance interval (97.5^{th} – 2.5^{th} percentile) of the distribution of absolute differences between the post-DA and observed NEE and between the model-only (no DA) and observed NEE (figure 5.10). The absolute differ-

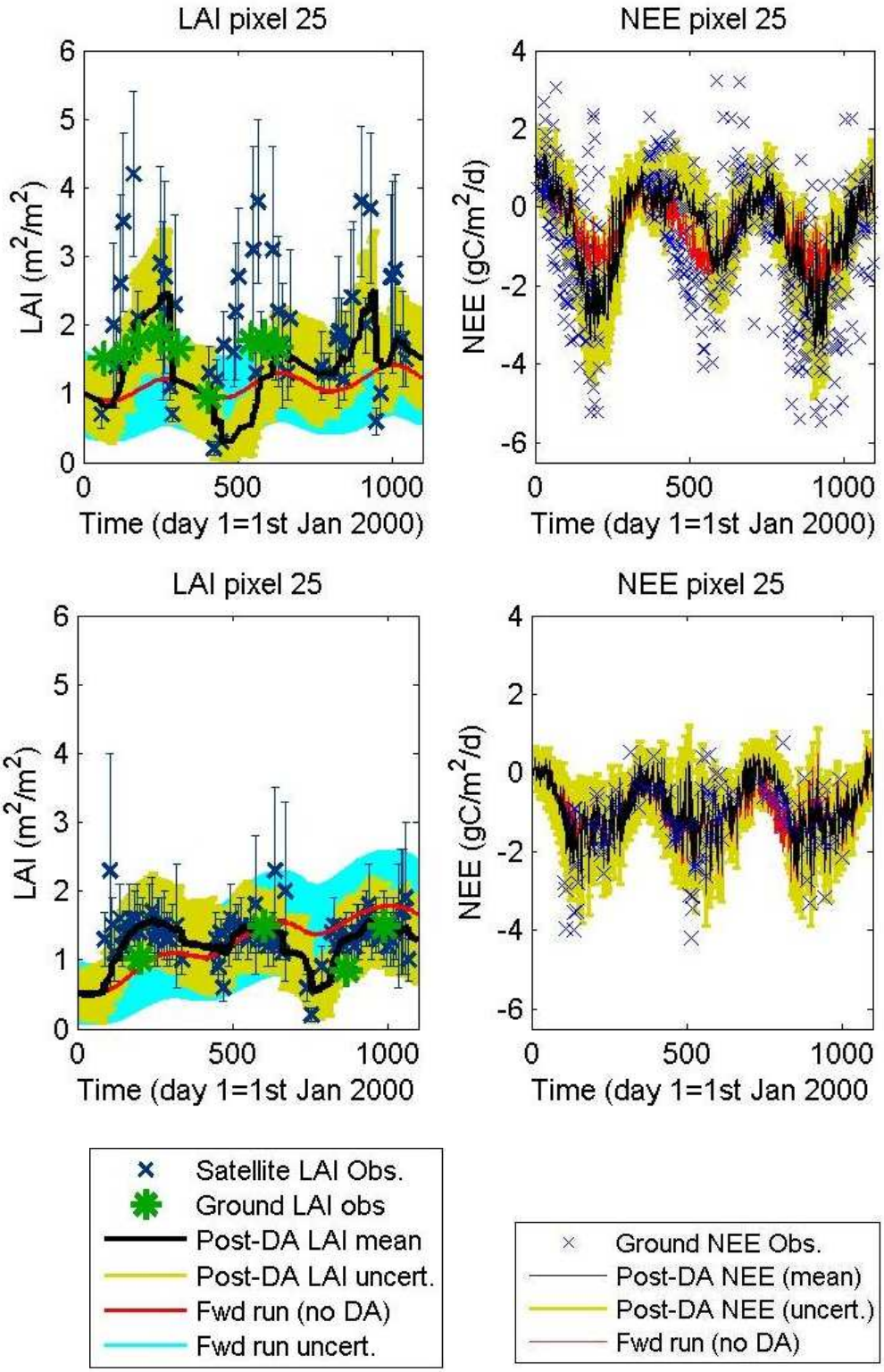


Figure 5.9a Plots of the post-assimilation LAI (left panel) and NEE (right panel) (black lines) for Loobos and Oregon, together with the MODIS LAI observations and ground NEE observations (blue crosses) and the forward runs. The uncertainty in the post-DA LAI/NEE and the forward run LAI is represented by +/- 1 standard deviation.

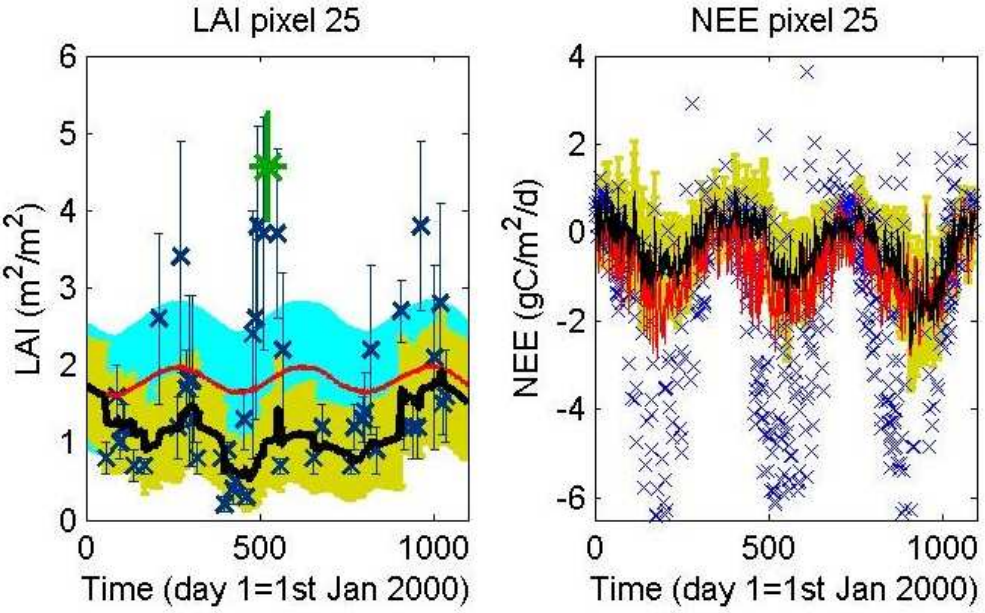


Figure 5.9b As figure 5.9a, except for Tharandt.

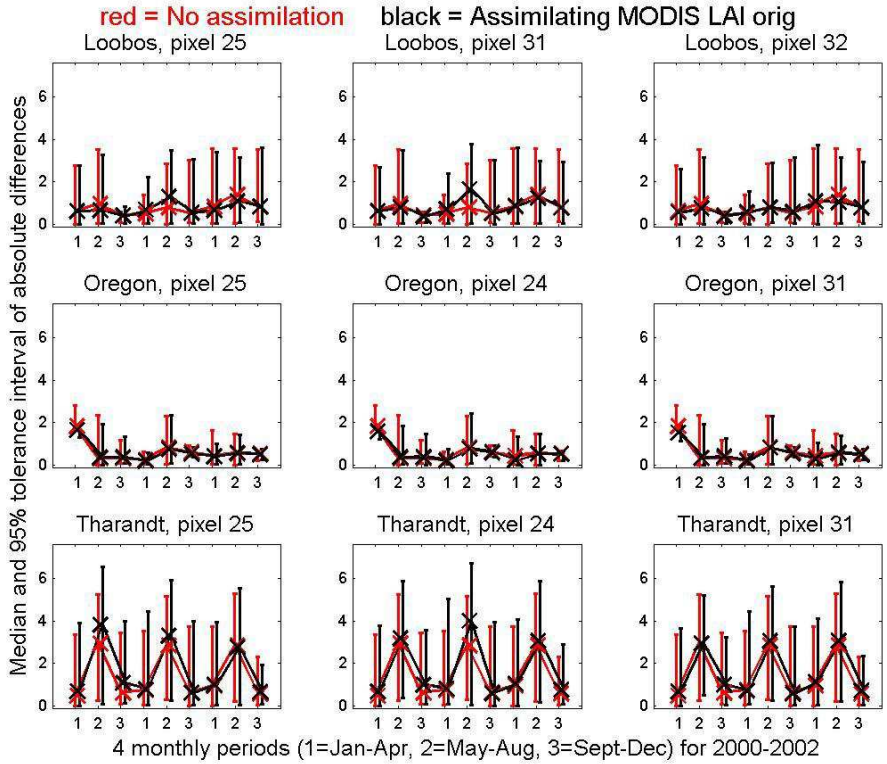


Figure 5.10 The median and 95th tolerance interval for: (red) the distribution of absolute differences between the modelled NEE (no DA) and the observed NEE, and (black) the distribution of absolute differences between the NEE after assimilating the MODIS LAI orig dataset and the observed NEE. The 1, 2 and 3 refer to {Jan-Apr}, {May-Aug} and {Sept-Dec}.

ences were subdivided into three periods of the year (Jan-Apr, May-Aug and Sept-Dec) for reasons given in subsection 5.2.7.

For Loobos, in the second year of pixels 25 and 31 (pixel 31 not shown), the small standard deviations of the small MODIS observations (top left panel of figure 5.9a) drag the post-DA LAI (black line) down to unrealistically low levels that are much lower than the LAI from the forward run (red line). During these periods, this results in a worse fit of the post-DA NEE to the observed NEE than for the forward run (top right panel of figure 5.9a and top row of figure 5.10). For the other periods, we mostly see an improved fit of the modelled to observed NEE due to assimilating the MODIS LAI observations (top row of figure 5.10); i.e. the median of absolute difference between the post-DA and observed NEE (black) is smaller than the median corresponding to the forward run (red), or the 95% tolerance intervals are smaller.

An interesting point to make here in relation to the top left panel of figure 5.9a, which is the same as the top panel of figure 5.6, is that the residuals between the MODIS LAI and the DALEC LAI are nearly always positive. This is interesting because in chapter 4 it was thought that this autocorrelation in the residuals was influenced to some degree by the fact that the data was synthetic there; but here the data are real.

For Oregon, the post-DA LAI trajectory is similar to the model-only LAI (figure 5.9a, bottom left panel), and we find that there is no difference in the fit of the modelled NEE to the observed NEE whether the MODIS data is assimilated or not, i.e. the medians and lengths of the tolerance intervals are approximately the same (second row of figure 5.10).

For Tharandt, assimilating the MODIS LAI observations tended to produce worse estimates of NEE than assimilating no data. Due to the small s.d. prescribed to the smaller MODIS LAI observations, the post-DA LAI trajectory is dragged to a lower than the model-only LAI trajectory, in some places by a large amount (e.g. around day 500 of the left panel of figure 5.9b). This is because the EnKF chooses the post-DA LAI to prefer the MODIS LAI observations with small s.d., since these are much smaller than the s.d. prescribed for the model. This causes the post-DA NEE to be much higher than the modelled NEE, particularly during the growing season (figure 5.9b). This is because the lower LAI mean than foliar carbon will be proportionally smaller, and due to the feedback in the DALEC model, GPP will also

become smaller, resulting in NEE getting larger, since NEE = total respiration – GPP. The median absolute difference between the post-DA and observed NEE (black crosses of the bottom row of figure 5.10) is equal to or greater than for the model-only NEE (red crosses). The 95% tolerance intervals for the modulus difference are also mostly always wider for the post-DA NEE estimates (black lines of bottom row of figure 5.10) than for the model-only NEE estimates (red lines), particularly for most of 2000-2002 for pixel 25 and for 2001-2002 for pixel 24. A final interesting observation is that the NEE observations at Tharandt (figure 5.9b, right panel) extend down to around -6.5gC/m² during the summer peak, whereas they extend only to around -4 g C/m² for Oregon and to around -5 g C/m² for Loobos (top and bottom right panels of figures 5.9a). As a result of this, the 95% tolerance intervals are significantly wider (mostly between 4 and 6 gC/m²; bottom row of figure 5.10) than for the other two sites (in general between 2 and 3 gC/m²; first two rows of figure 5.10).

5.3.3 Assessing the impact of processing the MODIS LAI data

After assimilating the MODIS LAI-smsd dataset, there is a noticeable improvement in the NEE estimates compared to assimilating the original MODIS LAI dataset, especially for Loobos and Tharandt. The smoothing increases the very small MODIS LAI values from the original dataset (first and third rows of figure 5.12). This means that small LAI observations after smoothing do not have as large an impact and so prevents the post-DA LAI trajectory from being dragged down as was the case when the original MODIS LAI was used. For Loobos, the post-DA LAI trajectory during the summer, labelled '2' on the x-axis, (e.g. around day 500) of pixels 25 (top left panel of figure 5.11a) and 31 (not shown) is now higher than after assimilating the original MODIS LAI (top left panel of figure 5.9a). As a result, the post-DA NEE is now extends to a lower level, particularly during the summer, and fits the observed NEE much better. This can be seen most clearly in figure 5.12, where the median absolute difference between the post-DA and observed NEE is less when the MODIS LAI-smsd dataset is assimilated (green crosses) than when the original LAI dataset is assimilated (black crosses), particularly during the summer. The corresponding tolerance intervals are also mostly shorter than those from assimilating the MODIS LAI-orig (black bars on figure 5.12). In contrast, assimilating the original LAI data,

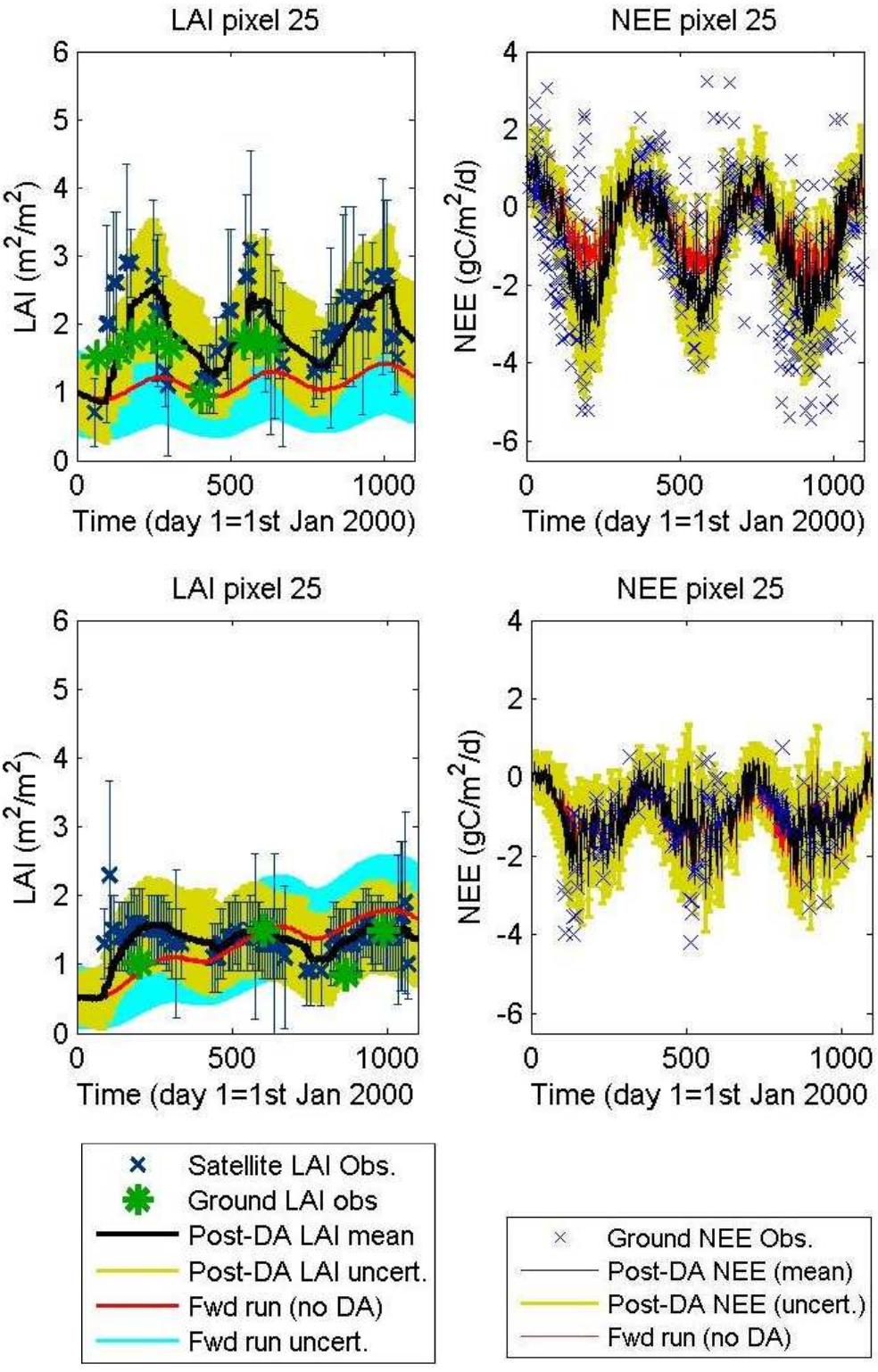


Figure 5.11a Plots of the post-assimilation LAI (left panel) and NEE (right panel) (black lines) for Loobos and Oregon, together with the MODIS LAI-smsd observations and ground NEE observations (blue crosses) and the forward runs. The uncertainty in the post-DA LAI/NEE is represented by +/- 1 standard deviation.

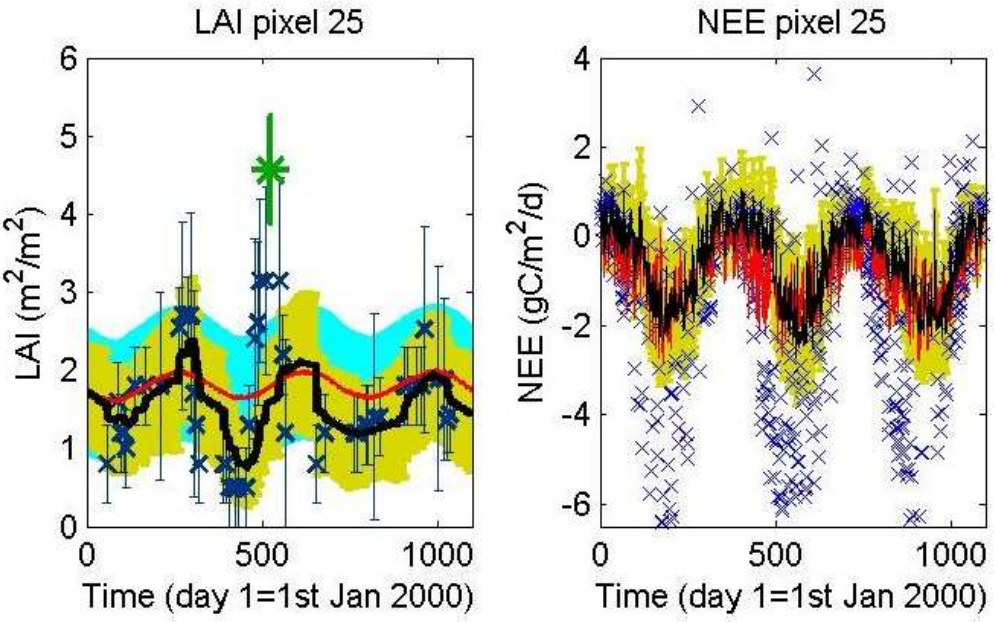


Figure 5.11b As figure 5.11a, except for Tharandt.

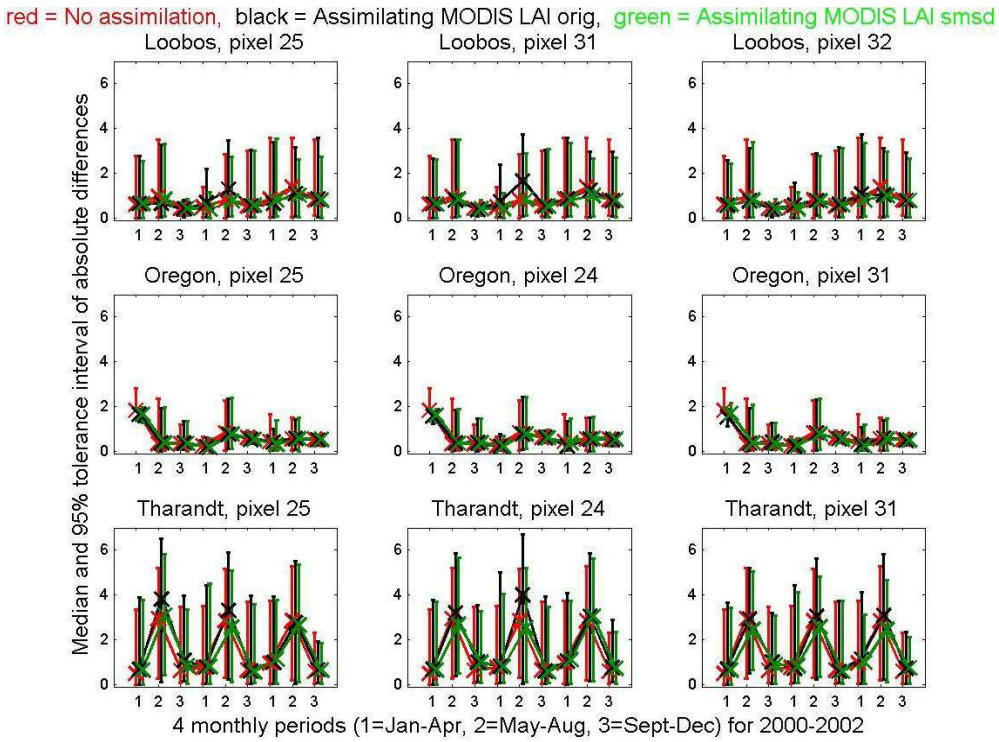


Figure 5.12 The median and 95th tolerance interval for: (red) the distribution of absolute differences between the modelled NEE (no DA) and the observed NEE; (black) the distribution of absolute differences between the NEE after assimilating the MODIS LAI-orig dataset and the observed NEE; and (green) differences between the NEE after assimilating the MODIS LAI-smsd dataset and the observed NEE. The 1, 2 and 3 refer to {Jan-Apr}, {May-Aug} and {Sept-Dec}.

we also find that assimilating LAI-smsd data does not result in a worse fit of the modelled to observed NEE for any of the 4-monthly periods, either in median difference or length of the tolerance interval differences.

Similar comments can be made for Tharandt (figure 5.11b), but the reductions in the median and the lengths of the tolerance intervals (bottom row of figure 5.12) are more pronounced and occur during the summer period for two of the three years for all three pixels considered. Although assimilating LAI-smsd data does on occasion result in a worse fit of the modelled to observed NEE, for other periods we see some improvement in the fit. This contrasts with assimilating the original LAI observations which results in a worse fit for all the 4-monthly periods during 2000-02.

For Oregon (second row of figure 5.11a), there is no noticeable improvement from assimilating the LAI smsd dataset, when compared to assimilating the original MODIS LAI dataset (central panel of figure 5.12). Therefore the improvements to the fit of the modelled to observed NEE observed when assimilating the MODIS LAI-orig data during May-August of the first year and January-April of the third year are also true when assimilating the MODIS LAI-smsd data.

5.3.4 Assimilating NEE and LAI observations

As expected, when in situ NEE observations are assimilated into DALEC in addition to the LAI-smsd dataset, there is a far better fit of the post-DA NEE to the observed NEE than in any of the previous experiments. This can be clearly seen from figure 5.14, where the median and spread of the absolute difference of the post-DA and observed NEE is significantly less when the LAI-smsd and NEE are assimilated (green) compared to when no data assimilation is carried out, or any of the MODIS LAI datasets are assimilated (not shown).

An interesting observation when assimilating NEE is its effect on the LAI, because we can learn what the MODIS LAI ideally would be in order to achieve a good fit of the post-DA NEE to observed NEE. In particular, for Tharandt (left panel of figure 5.13b) the post-DA LAI trajectory increases to a much higher maximum when assimilating just the LAI-smsd data (bottom left panel of figure 5.11b) although the troughs of both are at similar levels. This is caused by the NEE after assimilation being smaller. However, for the other two sites the LAI after assimilation roughly corresponds to the LAI obtained from just assimilating the MODIS LAI-smsd data, i.e.

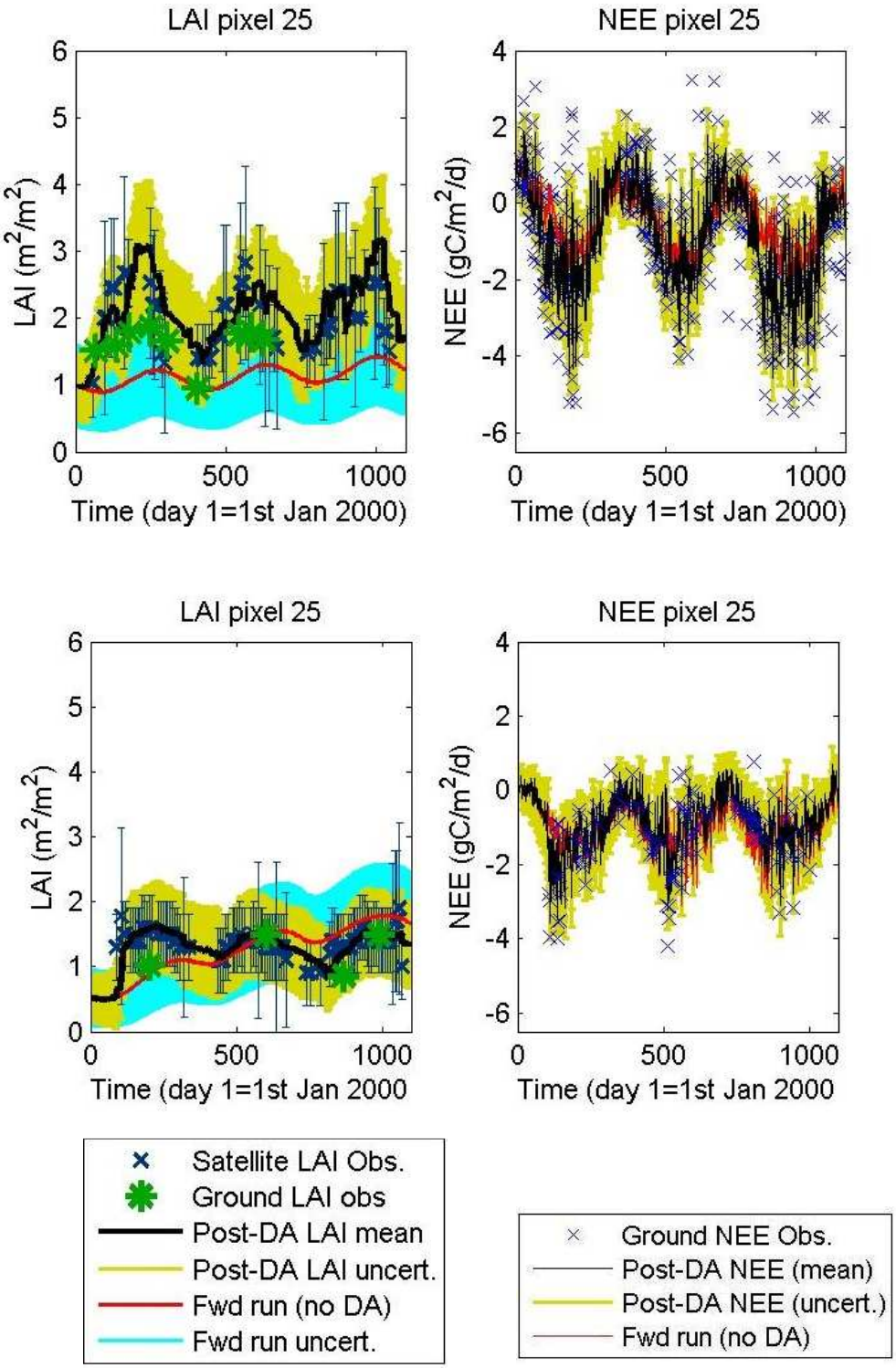


Figure 5.13a Plots of the post-assimilation LAI (rows 1 & 3) and NEE (rows 2 & 4) (black lines) together with the MODIS LAI-smsd observations and ground NEE observations (blue crosses) and the forward runs. The uncertainty in the post-DA LAI/NEE is represented by +/- 1 standard deviation. The first row corresponds to Loobos while the second one correspond to Oregon.

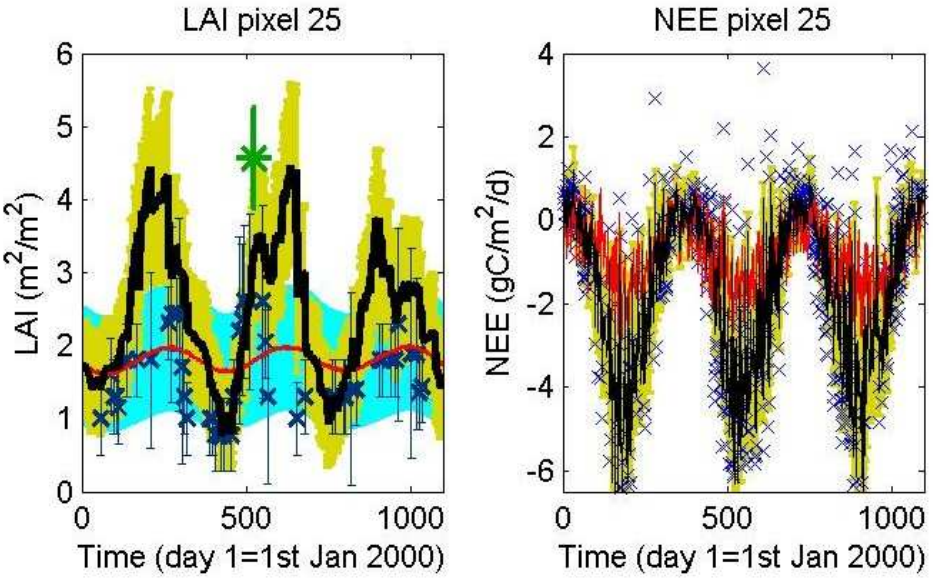


Figure 5.13b As figure 5.13a, except for Tharandt.

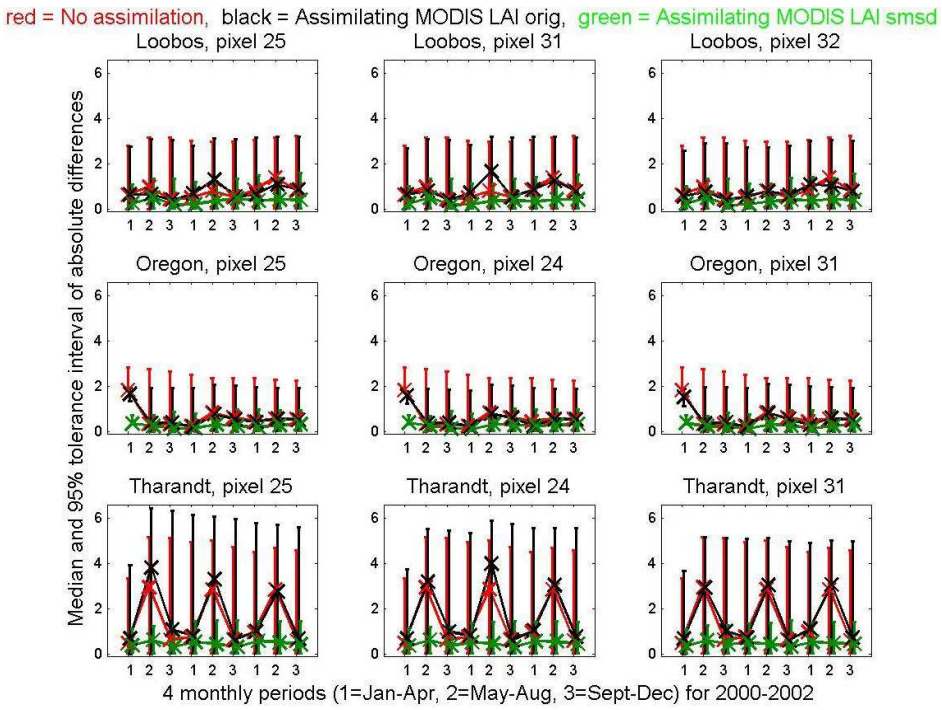


Figure 5.14 The median and 95th tolerance interval for: (red) the distribution of absolute differences between the modelled NEE (no DA) and the observed NEE; (black) the distribution of absolute differences between the NEE after assimilating the MODIS LAI-orig dataset and the observed NEE; and (green) differences between the NEE after assimilating the MODIS LAI-smsd and NEE datasets and the observed NEE. The 1, 2 and 3 refer to {Jan-Apr}, {May-Aug} and {Sept-Dec}.

i.e. each has a similar peak to peak amplitude with the peaks and troughs occurring at approximately the same time points (figures 5.13a and 5.13b versus figures 5.11a and 5.11b). The fit of the LAI after assimilation to the observed is also very similar, whether a processed LAI dataset is assimilated on its own or alongside the NEE ones.

5.3.5 Reduction in uncertainty

Figures 5.15a and 5.15b show how the uncertainty of the LAI from a forward run of the model (no DA) and after assimilating the unprocessed and processed MODIS LAI data. When assimilating the unprocessed data the uncertainty from the post-DA LAI (black lines of figure 5.15a) is less than the uncertainty of the modelled LAI from the forward run of the model for Oregon and Tharandt, but there is no significant difference for Loobos.

However when the processed MODIS LAI is assimilated we find that the reduction in uncertainty is not as great for Oregon and Tharandt, and for Loobos the uncertainty has in fact increased for the post-DA LAI. This increase in uncertainty of the post-DA when the processed MODIS LAI is used rather than the unprocessed can be explained by remembering that the unprocessed MODIS LAI had unrealistically low uncertainty at certain points in its dataset (particularly for low data values). However for the processed data (in the case of figure 5.15b the MODIS LAI-smsd dataset was used), the uncertainty for the very small MODIS LAI values were inflated. This resulted in an improved fit of the modelled to observed NEE, but at the cost of causing the uncertainty to slightly increase. Nonetheless, as stated above, for the two of the three sites (Oregon and Tharandt) it is still clear that the uncertainty in the modelled LAI is reduced for most or close to all of the time points when the processed MODIS LAI is assimilated compared to when no assimilation takes place.

5.4 Discussion

5.4.1 Advantages of using the processed MODIS LAI observations

We have shown that assimilating any one of the processed MODIS datasets compared

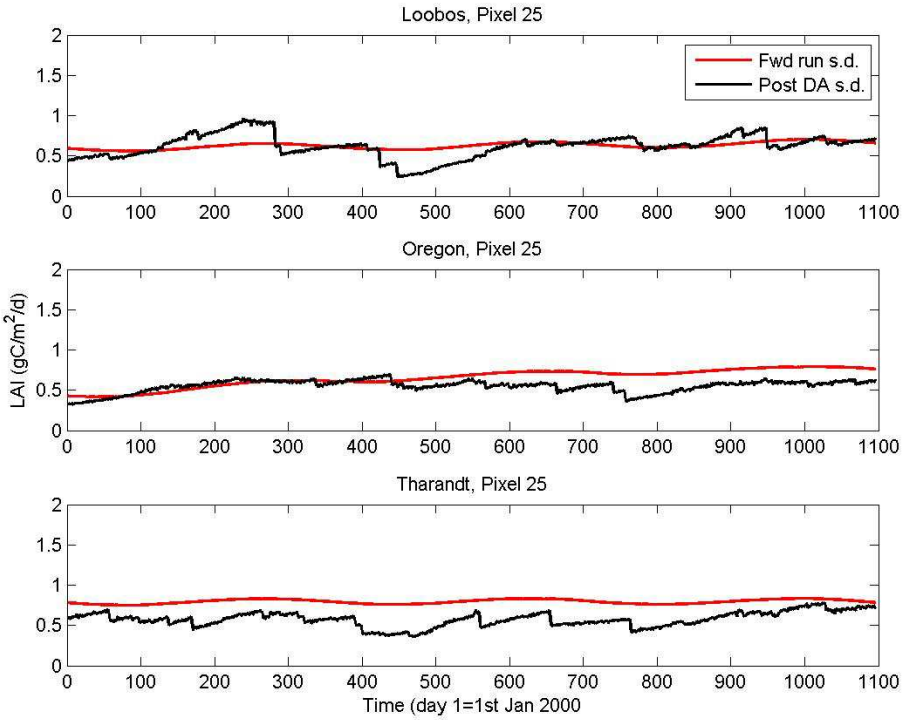


Figure 5.15a The uncertainty (expressed as a standard deviation) of the modelled LAI without DA (red) and with DA (black), corresponding to assimilating the original MODIS LAI dataset.

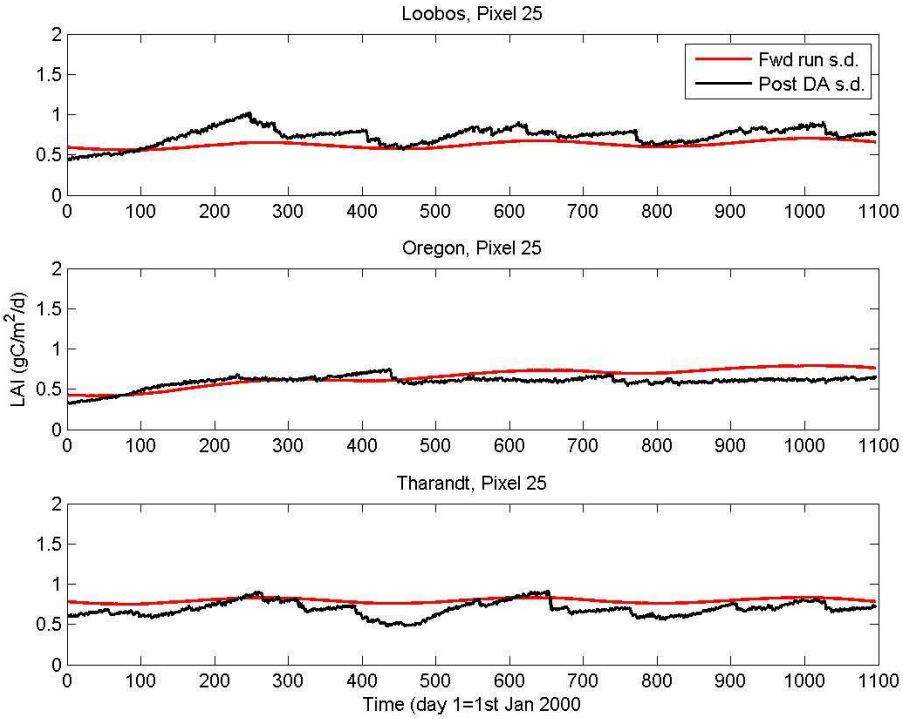


Figure 5.15b The same as figure 5.15a except these plots correspond to assimilating the MODIS-smsd LAI data.

to assimilating the original MODIS dataset noticeably improves the fit of the modelled to observed NEE. This was most evident during the summer period of 2002 for Oregon and Tharandt where the median absolute difference between the post-DA model NEE and observed NEE decreases by up to 50% and during most of the other summer periods for these two sites there was an improvement of at least 10%. Moreover we find that for Loobos and Tharandt, there is evidence of an improved fit when compared to the fit from the forward run. In Tharandt, while there remains some 4 monthly periods where fit of the modelled to the observed NEE is worse when the processed MODIS data are assimilated compared to the fit from the forward run, most of the periods report either a comparable or improved fit (i.e. green crosses and error bars versus red ones, for bottom row of figure 5.12) – this is in complete contrast to when the original MODIS data was assimilated where the fits were worse for all 4 monthly periods compared to the forward run (i.e. black crosses and error bars versus red ones, for bottom row of figure 5.12).

5.4.2 The oscillating NEE observations

One feature of the NEE observations is that they oscillate throughout this summer period from values greater than -1gC/m² to values below -3.5gC/m² in only a few days (figure 5.16). However, the trajectories of the NEE of the forward run and after assimilating solely MODIS LAI (red and black lines, respectively of figure 5.16) oscillate much less. Consequently the difference between the two NEE trajectories is not as large when the absolute differences between the modelled and observed NEE are calculated. For example, during the summer period of the first year of Loobos for pixel 25, the post-DA NEE being 1.3gC/m² lower on average (figure 5.16), but corresponds to a reduction of the average absolute difference of only 0.3gC/m² (i.e. black versus red crosses in top left hand panel of figure 5.10).

This inconsistency between the improvements we see in the post-DA LAI and NEE trajectories relative to the forward run, and the lack of improvement in the fit of the resulting modelled to observed NEE can also be seen when the processed MODIS LAI data are assimilated, for example the summer period of the second year of pixel 31 of Loobos (central panel of the first row of figure 5.10). This can also be explained by the large oscillation of NEE as described in the previous paragraph.

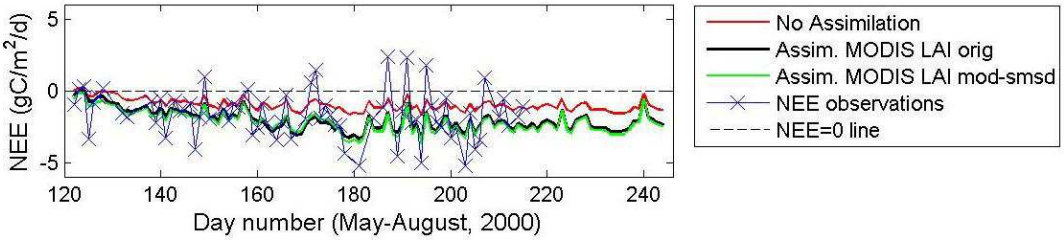


Figure 5.16 Time-series of the post-assimilation NEE states when assimilating the MODIS LAI orig dataset (black) and the LAI mod-smsd dataset (green) for the June-August period of 2000 for Loobos. The trajectory of the NEE from the forward run (red) and the NEE observations (blue) are plotted as well.

EXPLAINING THE OSSCILATORY BEHAVIOUR OF THE OBSERVED NEE

During the summer periods, such as the first year of Loobos (figure 5.16), while most of the NEE observations are significantly negative (e.g. less than -3 gC/m²) it is possible for NEE to increase and even become positive and to then drop down to values less than -3 gC/m² in the space of only a week. This variation is controlled predominantly by how much photosynthesis takes place, on a particular day or over a period of days, and the amount of precipitation. For periods of days where the solar radiation remains high and there is enough rainfall, the NEE at the start will remain in general less than -3 and -6 gC/m² depending on the weather, the intensity of the solar radiation and the available water. This can be seen by comparing the NEE, solar radiation and rainfall data side by side (table 5.3).

The NEE can change by larger amounts if there are larger changes in the solar radiation or rainfall change. If the solar radiation drops (or rises), we see a decrease (increase) in photosynthesis resulting in a decrease (increase) in GPP and thus an increase (decrease) in NEE (since NEE = Total Respiration – GPP). This can happen almost instantly, i.e. within one or two days. By comparing the NEE, solar radiation and rainfall data, we see that throughout the 3 year period, these sudden changes in the solar radiation and rainfall almost always correspond to a sudden change in NEE, resulting in this oscillatory behaviour. What is interesting is that, although the oscillatory behaviour in the observed NEE (e.g. during May-August of 2000) is not matched by the modelled NEE even after assimilating LAI (e.g. figure 5.16), the oscillations in GPP in situ measurements are in fact matched extremely well by the modelled GPP after assimilating LAI – see top panel of figure 5.17. We see similar behaviour for the second year of Loobos at pixel 25 where, unlike the first year, assimilating the processed LAI resulted in an improved fit of the modelled to observed NEE,

Day	NEE	Rad	Precip	Day	NEE	Rad	Precip
500	No data	15.2355	28.18	505	-2.731	20.8495	0
501	No data	19.5106	0.59	506	No data	22.6781	0
502	-0.544	13.6204	7.1	507	-2.379	25.1535	0
503	-2.41	17.7133	0.6	508	-2.283	28.8958	0
504	-0.027	9.8978	1.19	509	-1.188	28.9483	0

Table 5.3 Days 500-509 of 2000-02 at Loobos, showing in-situ measurements of NEE (gCm^{-2}), Solar Radiation or Rad (W/m^2) and Precipitation (mm). Days 505 to 509 are highlighted to show the effect of no rainfall on increasing NEE while solar radiation remains high.

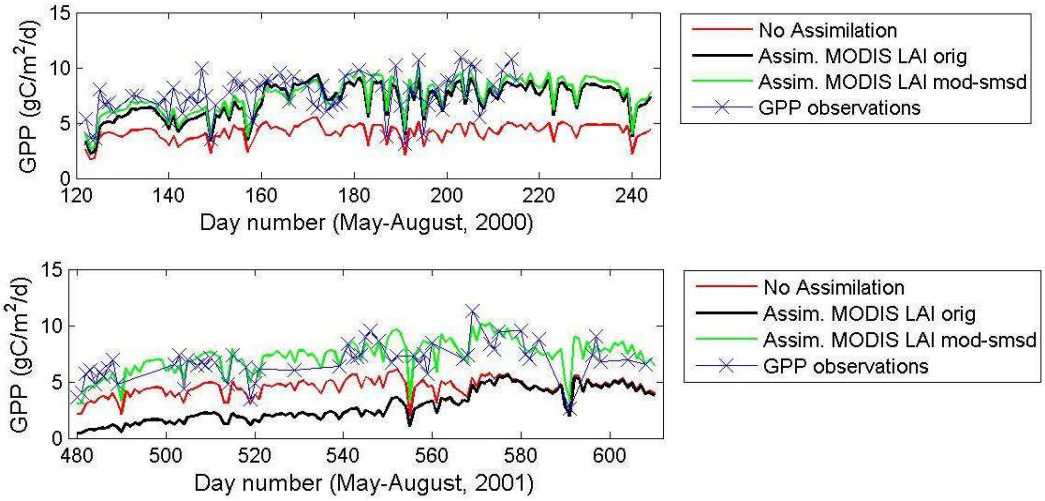


Figure 5.17 Time-series of the post-assimilation GPP states when assimilating the MODIS LAI orig dataset (black) and when assimilating the MODIS LAI mod-smsd dataset (green) for the May-August period of 2000 (top) and the May-August period of 2001 (bottom) for Loobos. The trajectory of the GPP from the forward run (red) and the in-situ GPP observations (blue) are plotted as well.

than assimilating unprocessed LAI or no assimilation. These findings suggest that in order for the model to be able to replicate the oscillatory behaviour of the observed NEE, respiration data needs to be assimilated as well as the MODIS LAI data, since NEE = respiration – GPP. Although satellite measurements of C respiration are not available, ground-based ones are.

If rainfall drops changes, this also causes a change in photosynthesis (because plants needs water to survive), but this is less sudden than changes due to changes in solar radiation. This is because when there is less solar radiation, the plant reacts immediately by closing its stomata (guard cells) in order to retain water. However, when rainfall reduces, the stomata remain open for a while as long as solar radiation is high enough for photosynthesis to take place until loss of water leads to water stress.

The plant will then begin to close its stomata to conserve water, leading to less photosynthesis and an increase in NEE.

5.4.3 Assimilating respiration and soil moisture data

Figure 5.16 shows that assimilating processed MODIS LAI data results in an extremely good fit of the modelled GPP to the GPP data. To further improve this fit, assimilating water based measurements (e.g. soil moisture) may be beneficial. However more importantly, to obtain a good fit of the modelled to observed NEE, the assimilation of respiration data (in addition to processed MODIS LAI data) is essential. Due to time constraints for this thesis it was not possible to assess whether assimilating these respiration and soil moisture data into the model would further improve our estimates of NEE. Instead we now describe how such measurements would be obtained and what changes to the model would be needed in order for them to be assimilated.

MEASUREMENTS OF RESPIRATION AND SOIL MOISTURE

Measurements of total soil and plant respiration are available, from the fluxnet database, by the use of isotopes. Isotopes enable the NEE, which is essentially the net vertical flux of CO_2, to be partitioned into its GPP and Re (respiration) components (Bowling et al., 2005). Measurements of the autotrophic and heterotrophic components of respiration are also available (Williams et al, 2005).

There are various types of water measurements available from ground and space and common ones are soil moisture content (Limer et al., 2007) and evapotranspiration, both of which are available at ground level from the GHG-Europe fluxnet database (http://www.europe-fluxdata.eu/newtcdc2/GHG-Europehome/login.aspx). Precipitation data are also often used to drive models which include water dynamics. From space, a common water based measurement is soil moisture. A number of satellite sensors have become available for measuring this (for example the ASCAT and AMSR-E satellites, see Draper et al., 2011), but two highly regarded ones dedicated specifically to soil moisture retrieval over land are the satellites from the SMOS (Soil Moisture and Ocean Salinity) mission, launched in November 2009, and the SMAP (Soil Moisture Active Passive) mission, due to be launched in 2014/15. SMOS uses a passive antenna to measure microwave L-band radiation offering global

coverage every 3 days at a resolution of 40km (Mecklenburg et al., 2008). The SMAP system is similar in its design to SMOS in many respects, but uses a simpler fixed passive microwave antenna allowing only one angle of the ground to be viewed and it has an additional active radar antenna which will enable much higher resolution measurements (around 1-3km) than are currently available (Gurney et al., 2011).

ASSIMILATING RESPIRATION AND SOIL MOISTURE INTO MODELS

To date, most data assimilation studies using soil moisture data have involved hydrological models. For example, Crow & Ry (2009) assimilated remotely sensed soil moisture observations into a hydrological model to improve runoff predictions. Moran et al. (2004) assimilated satellite-based soil moisture data into a soil-vegetation-atmosphere-transfer model and found improvements in the estimation of spatially distributed soil profiles, which have watershed applications such as drought and flood prediction and crop irrigation scheduling. Studies have also been also carried out to validate post-assimilation estimates of soil moisture. A good example is Draper et al., (2012), who assimilated near-surface soil moisture data from the ASCAT and AMSR-E satellite sensors into a catchment land surface model using an EnKF. The resulting modelled soil moisture improved the fit to ground-based measurements of soil moisture from 85 sites in the U.S. and Australia. Soil moisture has also been assimilated for agricultural applications. Nearing et al. (2012) assimilated observations of remotely sensed observations soil moisture and LAI into a dynamic crop development model using EnKF. The authors found that the resulting estimate of the end-of-season harvest yield agreed well with what was observed. There is also a growing number of studies where soil moisture satellite data has been assimilated into land surface models (LSMs), such as Kumar et al. (2008) and Tian et al. (2007). For example, Kumar et al. (2008) used the Ensemble Kalman filter to assimilate remotely sensed surface soil moisture into the Catchment and Noah LSMs, resulting in an improved to both models in the soil moisture estimates.

Despite the literature on assimilating soil moisture data, no studies have been carried out where satellite measurements of soil moisture have been used in terrestrial ecosystem models to improve estimates of Net Ecosystem Exchange. Therefore it is a matter of urgency that this is done, as this will make an enormous difference. However, the JULES model has recently been incorporated into a 4D-Var DA

scheme, and it is hoped as many global data streams as are available will be assimilated into the model (Luke, 2011). A major challenge for this will be the fact that JULES has a huge number of parameters which need to be estimated. The uncertainties on all the data will need to be correctly quantified, which will also be extremely difficult given the huge amount of data from across the globe which will be used.

5.4.4 Assimilating MODIS LAI-smsd and ground NEE

Further evidence for the need to assimilate measurements other than LAI, particularly respiration data, can be seen when we observe what the LAI trajectory looks like when MODIS LAI and NEE data are assimilated. Figure 5.18a shows the trajectories of LAI after assimilating MODIS LAI-smsd both on its own and with in situ NEE. For Oregon (second panel of figure 5.18a), the post-DA LAI trajectories match each other extremely well, which results from the corresponding post-DA NEE trajectories matching each other well (second panel of figure 5.18b). The slight differences observed in the median and spread of the absolute differences for some periods e.g. May-August 2001 (figures 5.12 and 5.14) is to do with the increased variability of the NEE observations around this period (e.g. days 500-600 in second panel of figure 5.18b); this is picked up when NEE is assimilated (green line of figure 5.18b) but not otherwise (black line of figure 5.16b).

For Loobos, the post-DA LAI trajectories matched each other fairly well (i.e. top panel of figure 5.18a) yet the medians and spreads of absolute differences (green symbols in the 1st rows of figures 5.12 and 5.14) are very different for many of the periods, particularly May-August. As for Oregon, this is due to the increased variability of the NEE observations during these periods (top panel of figure 5.18b). This can be picked up when assimilating the actual NEE observations, but not as well otherwise. This supports the case, as outlined at the end of subsection 5.4.3, that assimilating respiration data along with the MODIS LAI data would be extremely beneficial. This is because the respiration data may help in representing the observed variability of NEE better than just assimilating MODIS LAI.

For Tharandt, assimilating only MODIS LAI data results in an NEE that does not match the oscillatory behaviour of the observed NEE (bottom panel of figure 5.18b). Thus, as with Loobos, assimilating respiration data is also essential. Looking

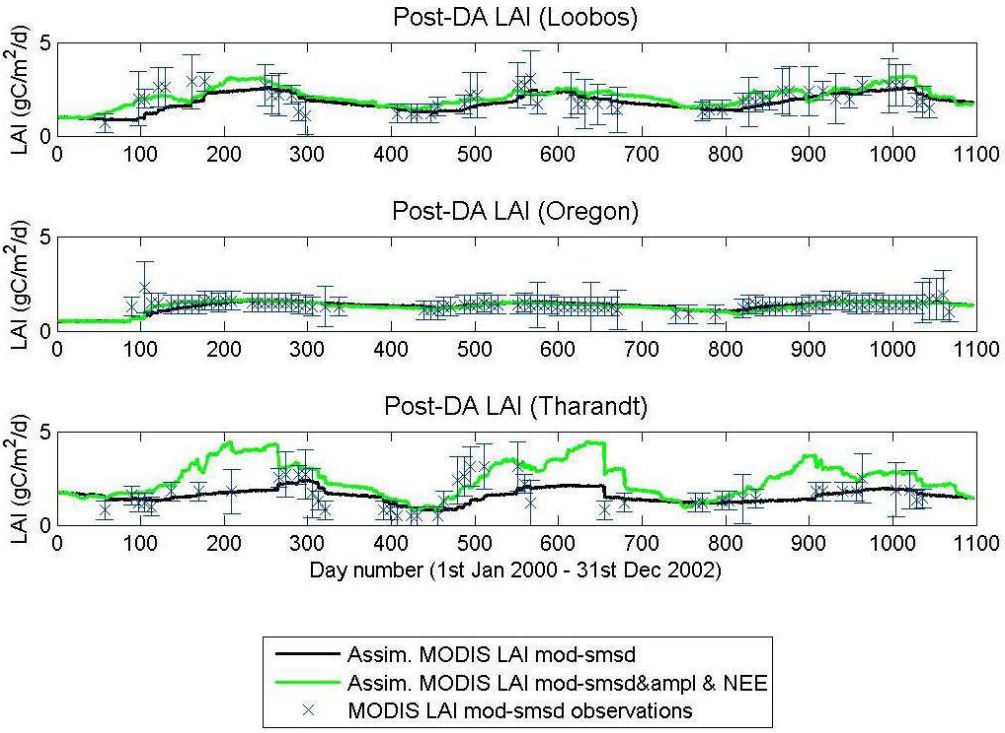


Figure 5.18a Time series of the post-DA LAI trajectory for pixel 25 when assimilating the MODIS LAI mod-smsd dataset on its own (black) and when assimilating the MODIS LAI-smsd and ground NEE datasets. The three panels correspond to the different sites: Loobos (top), Oregon (middle) and Tharandt (bottom).

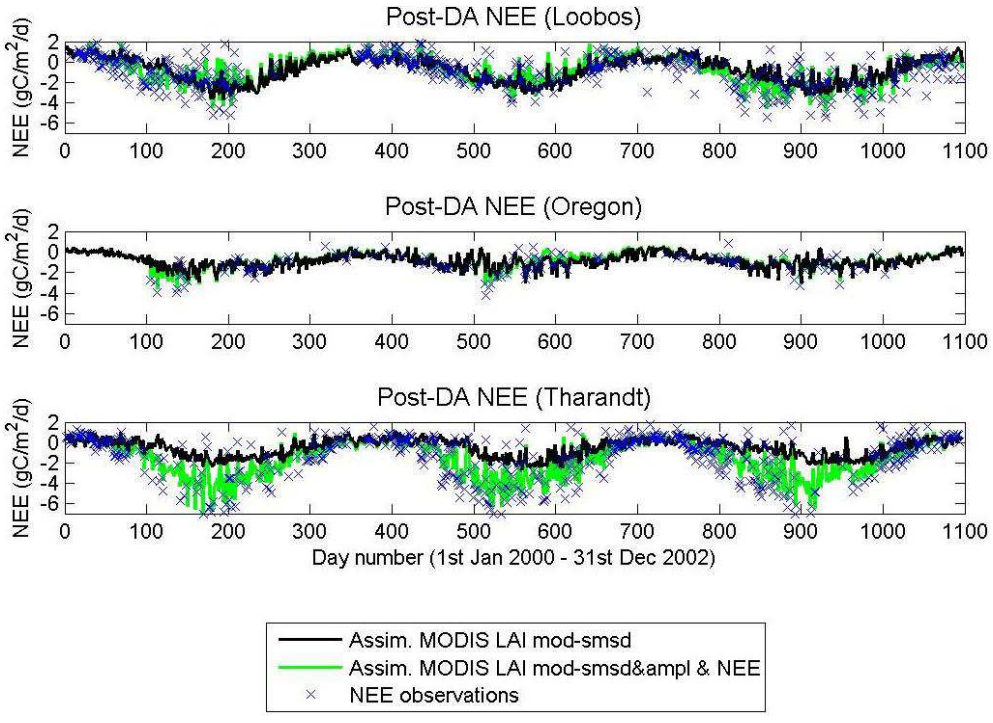


Figure 5.18b The same as figure 5.18a except the post-DA NEE is plotted instead.

at the bottom panel of figure 5.18a, we also can see that while the assimilated MODIS LAI observations need to be at a higher level in order for the post-DA NEE to better fit the observed NEE (bottom left panel of figure 5.13b), there are two other interesting points worth making:

(1) It is clear that the two post-DA LAI trajectories (i.e. black and green trajectories of bottom panel of figure 5.18a) have different summer maximums particularly in the first and third years (i.e. around 200-300 and 900-1000). This suggests that the MODIS LAI dataset may need to be further modified to account for this.

(2) Between days 475 and 550 (approximately), there are 5 MODIS LAI-smsd observations which appear to follow the green post-DA LAI trajectory (i.e. when the MODIS LAI-smsd and NEE are assimilated) in the bottom panel of figure 5.18a. However the black post-DA LAI trajectory (after assimilating the MODIS LAI-smsd data) is at a much lower level. This is probably due to the relatively larger s.d. values prescribed to these MODIS LAI-smsd observations resulting in the model having more influence on the post-DA trajectory (top left panel of figure 5.11b). This is unusual because in following the green line the actual values of these five MODIS LAI-smsd appear to be a more correct, yet the larger uncertainty prescribed to them has prevented them from influencing the post-DA trajectory as much as they should. This implies that either the observational uncertainty is too large or the model uncertainty is too low, or perhaps a combination of both. In any case this identifies how critically important it is to accurately estimate all uncertainties in a DA scheme.

5.4.5 Other issues

Although unrelated to the rest of the discussion but important to mention here, the issue of the parameters being treated as fixed should be briefly discussed. In a typical state estimation problem such as in Williams et al. (2005) and Quaife et al. (2008), the parameters are treated as fixed and then the model uncertainty is represented by adding stochastic noise to the elements of the state vector. In this way, we are indirectly saying that the parameter set used is uncertain by recognising that the resulting model output, i.e. the state vector, is uncertain. Nonetheless, given what has been learnt from the previous chapters, it would be beneficial to additionally treat the

parameters as directly uncertain, so that when the model is propagated forward for each ensemble member, either the parameters are augmented to the state vector or the parameter uncertainty is incorporated in the model execution, during periods of time in between data points, in some other way.

It is also necessary to mention here how these fixed parameter sets were estimated. For the Loobos and Tharandt sites, the parameters were optimized using ground based NEE from a site of similar size and characteristic to these two sites and which was reasonably close by (i.e. of similar meteorological conditions and climate etc…). However for the Oregon site, the parameter values used were adopted from Williams et al. (2005); this was because that study also used the EnKF with the same model over the same years, therefore it made sense to use the calibrated parameters used in that study. However, the parameters that were optimized in Williams et al. (2005) were done so using data, including NEE, which was then used again to estimate the states (i.e. using the EnKF for state estimation). In effect, the data was used twice. Since the same NEE data used in that study was used here, this NEE data has also been used twice. This is therefore a limitation of this study, but also of Williams et al. (2005) and also Quaife et al. (2007) who additionally used the NEE data twice in this way. This issue may help to explain why the effect of processing the MODIS LAI had little effect on improving the fit of the modelled to the observed NEE for Oregon.

Finally, it is important to comment how the post-assimilation LAI trajectories for the different experiments compare with the ground LAI data. Focusing on Loobos, where there is the most LAI data, it was mentioned in subsection 5.2.4 that the ground LAI data have a much lower peak to peak amplitude than the MODIS LAI, which is one of the reasons for modifying the MODIS LAI to reduce this amplitude. However in figure 5.5, it is interesting to note that all of the modified MODIS LAI datasets have a noticeably reduced peak to peak amplitude compared to the original dataset. This reduction in the peak to peak amplitude is also present in the post-DA LAI trajectory when the modified MODIS LAI data are assimilated (e.g. the black line of the top left panel of figure 5.11a). However comparing the ground LAI data with this post-DA LAI (i.e. comparing the green points with the black line of the top left panel of figure 5.11a) we see that there is a horizontal offset between the post-DA LAI and the ground LAI data. If we believed it is necessary to reduce this offset, this

could be done in numerous ways: first we could further modifying the MODIS LAI, but it would be hard to properly justify this; the DALEC parameters used to produce the non-DA modelled LAI (i.e. the red line in the top panel of figure 5.11a), could be calibrated against the ground LAI data; alternatively the ground LAI data could be assimilated along with the MODIS LAI, however it would then be more appropriate to use a DA scheme such as the Ensemble Kalman Smoother as using the EnKF would result in the ground LAI not informing the post-DA LAI in 2002 where there are no ground LAI data. For Tharandt, we can see in the left panel of figure 5.11b that the only ground LAI data point in 2001 is significantly higher than the post-DA LAI trajectory; this is perhaps slightly worrying. This is happening, despite the uncertainty intervals on this ground observation and the neighbouring MODIS LAI data do overlap, because of the large uncertainty intervals on the MODIS LAI data. Given the MODIS LAI and ground LAI tend to broadly agree with the higher level of LAI in 2001 compared to the other two years, it is concerning that we do not see this higher level of LAI with the post-DA LAI (i.e. the black line of the left panel of figure 5.11b). As was suggested above at the Loobos site, one way to take this into account is to assimilate the ground LAI alongside the MODIS LAI.

5.5 Conclusions

This chapter investigated whether processing MODIS LAI observations, to make them more realistic, before assimilating it into DALEC improved the fit of the modelled to the observed NEE. The processing involved smoothing the LAI signal and increasing the unrealistically small standard deviation values. While these changes made the MODIS LAI more realistic, it was unclear whether this processing made a difference when assimilating the LAI into the model. If we were to find out that it did not, then this would have suggested that DA is insensitive to the unrealistic features of the dataset being assimilated. We found the opposite, i.e. processing the MODIS LAI resulted in a significant improvement in the modelled NEE estimates after assimilation.

We found that, for Loobos and Tharandt, assimilating the original MODIS LAI resulted in a worse overall fit of the modelled to observed NEE. This was due to small uncertainties (expressed as standard deviations) prescribed to some of the

MODIS LAI observations, predominantly the small LAI data values. This resulted in the post-assimilation trajectory of LAI being dragged down to unrealistically low values. For the third site, there was no difference, but this was mainly because the MODIS LAI was very similar to the LAI from the forward run of the model. The assimilation of the processed MODIS LAI data resulted in vast improvements to the fit of the modelled to observed NEE, for Loobos and Tharandt, mainly because the small MODIS LAI standard deviations were now larger and not dragging the LAI, after assimilation, to unrealistically low values. For Oregon, there was no improvement in the fit of the modelled to the observed NEE when the processed MODIS LAI was used; this was mainly because the original MODIS LAI was far less noisy, had a smaller peak-to-peak amplitude and did not have contain as much data with small s.d. values, compared to the other two sides.

The results also revealed an interesting anomaly. They showed that where there was a large difference between the trajectories of the NEE from the forward run (no DA) and the NEE from assimilating the MODIS LAI data, this difference did not appear to be as large when comparing the fit of the each of the two NEE trajectories to the observed NEE. This was due to the oscillatory behaviour of the observed NEE which was only slightly evident in the two modelled NEE trajectories (figure 5.16). However, when we plotted GPP from the forward run and after assimilating the processed LAI, we found an incredibly improved fit of the modelled to observed GPP as a result of assimilating the processed MODIS LAI (figure 5.17). This suggests that the oscillating behaviour seen in the observed NEE could be replicated by the model if respiration data were assimilated with the processed MODIS LAI data. Further evidence for the need to also assimilate respiration data was realised when the NEE and MODIS LAI-smsd data were assimilated.

There were two main limitations with this chapter. First, there were not as many ground NEE observations in many of the nine 4-monthly periods. For Loobos the mean percentage of days with NEE observations present was 47%, with a minimum and maximum of 8% and 72% respectively. For Oregon the problem was worse: the mean was only 11% with a minimum and a maximum of 2% and 24%. For Tharandt the mean, minimum and maximum were 51%, 33% and 71% respectively. As a result of the lack of enough NEE observations, the plots of the medians and 90%

tolerance intervals for the absolute differences between the post-DA and observed NEE (figures 5.10, 5.12 and 5.14) may suffer from minor inaccuracies.

Secondly, the ground LAI observations may have been able to be utilized better in this study. A direct comparison with ground measured LAI and MODIS LAI is not generally advisable due to the different scales the two types of observations (10-20m and 1km scales, respectively) are measured at, and also the fact that they are measuring two different quantities from ground and space. For the scaling problem, if enough ground based LAI measurements are available for in a certain 1km×1km pixel then it may be possible to upscale the ground LAI measurements for a meaningful comparison with the MODIS LAI ones (De Kauwe, 2011). For this chapter, the few ground LAI observations available were made at different points in time, so this upscaling was not possible.

Chapter 6

Conclusion

6.1 Addressing the Thesis Aims

We first address the two aims of this PhD thesis, as stated at the end of chapter 2.

Aim 1: When using Data Assimilation (DA) to estimate parameters using ground observations:

(i) To understand the limitations of DA and the conditions under which it performs best.

(ii) To determine likely factors that cause parameter estimates to vary.

Aim 2: To assess the impact of assimilating satellite observations of leaf area index to improve the model states, and whether DA is robust against unrealistic features of the satellite data.

Aim 1 (i) To understand the limitations of DA and the conditions under which it performs best.

Correctly estimating the parameters is important if we wish to use them to forecast (for example) not only NEE but also other important quantities such as GPP, R_a (autotrophic respiration) and R_h (heterotrophic respiration). This is because although many different sets of parameters may result in a good fit of the modelled to observed NEE, if the model suffers from parameter equifinality, only the correct parameter set is likely to result in the other important quantities such as GPP, R_a and R_h being modelled well. In this thesis, we found many circumstances when DA wrongly estimated the parameters, or could only correctly estimate them under strict conditions. In other situations, DA could be robust and correctly estimate parameters with not much information. We summarise the main findings below:

- In chapter 3, we found that if only two years of daily NEE observations and a small number of LAI observations are assimilated into DALEC-D using the Metropolis algorithm, then the parameter estimates are biased if the initial value of the labile carbon ($C_{lab}(0)$) is not at its true value. However, if the initial value of the labile carbon pool is over-estimated by a reasonable amount ($110 gCm^{-2}$) compared to its true value ($70 gCm^{-2}$), then the modelled trajectories of the C pools and NEE are reasonably close to their true trajectories (figures 3.12-3.14, subsection 3.4.1), although the deviations away from the true trajectories increase with time for the C_w and C_{som} pools (figure 3.13). Nonetheless this deviation away from the true trajectory is significantly greater when $C_{lab}(0)$ is under-estimated ($35 gCm^{-2}$) or greatly over-estimated ($170 gCm^{-2}$). Moreover when $C_{lab}(0)$ is under-estimated, the C_r trajectory is not only completely out of phase with the true trajectory, unlike the other trajectories which remain in phase (figure 3.12b, subsection 3.4.2), but after 3 years C_r decreases to zero, meaning that the trees are essentially dead. This suggests that $C_{lab}(0)$ should be greater than 35 gCm^{-2} and less than 170 gCm^{-2} although based on these results it is difficult to give accurate estimates of exact lower and upper bounds $C_{lab}(0)$ should lie between in order for all of the C trajectories of the model to be close to the true trajectories.

  The results also indicated that even if $C_{lab}(0)$ is wrongly estimated, by the end of 3 years the trajectories of the fast C pools (using the different sets of estimated parameter sets) in fact converge towards each other (figure 3.12), so the effect of wrongly estimating $C_{lab}(0)$ may only be transient, and the issue could be resolved by assimilating a time series greater than 2 years. However, this 'transient' argument does not apply to the slowly changing C pools or if the $C_{lab}(0)$ is under-estimated ($35 gCm^{-2}$) where the trajectories are completely out of phase with the true trajectory.

- In chapter 4, we found that when assimilating a stream of daily NEE data, if the data period is 1 year this is likely to give biased estimates of parameters for EnKF and biased forecasts for EnKF and the Metropolis algorithm. We found that at least between 2 and 5 years of data was needed for unbiased estimates (figure 4.6a, subsection 4.3.3); however Metropolis algorithm runs for a dataset length of 10 years or greater were not included in the results because the runs

were not at the global minimum and produced strange behaviour, in particular the residuals between the model and observations in the cost function were autocorrelated, which is likely to be the reason for the corresponding parameter estimates being biased (see next bullet point for details). Nonetheless for the EnKF it is astonishing that, using only 2 years of noisy NEE data of which 60% is gap free, it can accurately estimate 15 of the 17 unknown parameters (figure 4.6a, subsection 4.3.3).

- In chapter 4, we found that when a stream of NEE data is assimilated using the Metropolis algorithm, the estimated parameters can be far from their true values whilst still resulting in a near optimal fit of the modelled to observed NEE. However, the forecasts produced using this wrong parameter set were found to not only be very accurate but robust against a severe drought imposed during the forecast period. While it is difficult to give concrete evidence from the results as to why this parameter equifinality is occurring, it is probable that the poor parameter estimates for 11 of the 15 runs is due to the residuals in the cost function being autocorrelated. These is for two reasons:

  (i) One of the assumptions of using the Likelihood function in DA schemes like the Metropolis algorithm is that these residuals are independent; a violation of this in probability terms means if we have two events A and B and wish to compute $P(A \cap B)$, using the expression $P(A) \times P(B)$ is only permissable if events A and B are independent. Therefore if the residuals are correlated but in the Likelihood function they are treated as being independent then the Likelihood function will be wrong, and it is therefore perfectly reasonable to expect the resulting parameter estimates to be biased.

  (ii) Of the 15 runs of the Metropolis algorithm, 11 had Likelihood functions whose residuals were autocorrelated and the same 11 runs exhibited parameter estimates far from the truth and uncertainty intervals which in general did not include the truth (figures 4.6b, 4.7b, 4.8b, 4.16, chapter 4). For the remaining 4 runs where the residuals were independent in time, the results contained the following features which were not present in the results to the other 11 runs: (I) the parameter estimates were close to the truth and the majority of the uncertainty intervals included the truth (figures 4.6c, 4.7b, 4.8b, 4.16, chapter 4); (II) the parameter estimates and uncertainties matched up fairly well with the EnKF

ones; (III) the modelled NEE using the parameter sets from the posterior distribution fitted the NEE data much better compared to the other 11 runs (figure 4.22); (IV) the true parameter set had a cost function value which was contained in the distribution of the cost function values corresponding to those parameter sets from the posterior distribution. The probability of all these features occurring by chance for these 4 runs is likely to be close to zero. Subsection 6.3 summarises ways to modify the Likelihood function to take account of autocorrelated residuals.

Aim 1 (ii): To determine likely factors that cause parameter estimates to vary.

It was found the following factors caused variations in the parameter estimates:

- The initial value of the labile C pool, and to a lesser extent the initial value of the litter C pool and the LAI weighting term in cost function, when using the Metropolis algorithm and when not treating the initial conditions as parameters.
- The length of dataset being assimilated for the EnKF and the Metropolis algorithm. The Metropolis algorithm runs also showed that there were variations in the parameter estimates when there were variations in the data density or observational error, but these runs were excluded from the results because the autocorrelated residuals in the cost function for most of these runs as described previously.

Aim 2: To assess the impact of assimilating satellite observations of leaf area index to improve the model states, and whether DA is robust against unrealistic features of the satellite data.

Here we were interested in understanding the utility and pitfalls of assimilating satellite observations into a model in order to improve the estimates of its states, using a fixed parameter set. We found that the LAI observations from the MODIS sensor of the TERRA satellite contained unrealistic features. In particular: (1) they contain excessive temporal variation, with the MODIS LAI changing by 2–4m^2/m^2 in only 1 or 2 weeks in some instances; (2) the prescribed uncertainty on these LAI measurements is excessively low at certain time points, especially when LAI is small; As a result the fit of the modelled to observed NEE after assimilating the LAI data into DALEC was much worse than not assimilating any data. However if the

MODIS LAI is processed prior to assimilation, to remove these unrealistic features, the fit of the modelled to observed NEE is vastly improved, with an even better fit for GPP.

6.2 Limitations

Here we outline three limitations of this thesis which are: (i) the use of synthetic data; (ii) an arguably poor assessment of the model uncertainty; and (iii) the lack in the model of a water dynamics representation.

(i) Synthetic data: While the model processes used to generate the 'true data' may approximate real-world processes, the complexity and multi-dimensionality of these processes may not be captured in DALEC, or even a more complex model. Nonetheless synthetic data are still of great value as they do offer a good approximation to real data, and can be used in ways real data cannot – for example obtaining real datasets which are identical except for the size of their observational errors would be impossible. Furthermore, the quantification of observational errors is an active area of research and as this work develops, it will enable synthetic data to be an even more accurate representation of real data.

(ii) New DALEC parameters: Despite the arguments for treating Leaf Mass per Area (LMA) and foliar Nitrogen as constants, it is recognised that they should really be treated as parameters; this is a limitation to this work but also to all other studies that have used the DALEC model. Chuter (2013) found that NEE as simulated by DALEC is most sensitive to one of the existing model parameters but also to LMA. Therefore, wrongly estimating it could result in biases to the resultant modelled NEE. Chapter 5 discussed the inclusion of a further parameter to improve process representation, namely a clumping parameter. DALEC assumes no clumping and so the modelled LAI are not attempting to simulate 'true LAI', unlike satellite and ground based inferred data of LAI where clumping is taken into account. This creates a theoretical problem when LAI data is being assimilated into the model or compared with modelled LAI. Including an additional parameter to take account of foliar clumping in the tree canopy would therefore help to overcome this problem.

(iii) Simulating water dynamics: Despite a growing number of DA studies have used DALEC, in particular Williams et al. (2005), Quaife et al. (2007), Fox et al., (2009), DALEC does not contain a representation of the water dynamics; in particular it does not simulate evapotranspiration or how water flows through the soil. This is important to include in a model, because a plant's ability to photosynethesise is dependent not only on the amount of solar radiation, but also on the amount of available water. DALEC was used because: (i) it is very fast to run (as a result of being simple), whilst containing an accurate representation of key processes of C dynamics in evergreen and deciduous vegetation (e.g. the GPP submodel used in DALEC is a simplified but accurate version of the more complex SPA model); (ii) it was also designed for DA. Furthermore, we were able to include a pseudo water representation in chapter 4, by artificially changing two of the fixed parameters in the GPP submodel, responsible for stomatal control, in order to simulate a drought.

6.3 Implications for terrestrial ecosystem knowledge

Although the findings in this thesis are informative and interesting, they need to relate to the larger scientific questions laid out in chapter 1. The motivation for this thesis was to improve estimates of the terrestrial carbon sink in terms of accuracy and reduction of uncertainty, as described at the end of section 1.1 (pages 13/14). This thesis assesses ways in which Data Assimilation (DA) obtains the optimal estimate of the terrestrial C sink, which on the daily to seasonal time scale is NEE, by combining the information from the two sources of knowledge, namely observations and models. A key advantage of DA is that the post-DA estimate of the terrestrial C sink, typically has a lower uncertainty than either the model or the observations. The purpose of this thesis was an exploratory analysis of how DA works in different situations, and so the knowledge gained from this thesis was intended to indirectly improve terrestrial C cycle representation by highlighting how DA should be optimally used, rather than providing actual estimates of the terrestrial C cycle on the global scale. It should be further added that the increase in knowledge gained from this thesis to help answer the

questions that arise at the bottom of page 13 / top of page 14 only relate to (i) and (iii). Question (ii), although not directly answered, is still addressed in the same way as (i) and (iii) are; as this thesis will help to further improve and constrain estimates of NEE globally, it will become clearer with this improved knowledge as to the response of the land-atmosphere flux of C to elevated CO_2.

In the following, we describe six ways in which this thesis has improved our understanding of the terrestrial C cycle through DA.

(1) Using synthetic datasets for DA can be inappropriate. The advantage of assimilating synthetic data is that we know the answer, i.e. the true parameter set, and this enables one to assess to some degree the accuracy of the parameter estimates and whether the uncertainty intervals are realistic. However as explained earlier, for 11 of the 15 Metropolis algorithm runs from chapter 4 – namely those corresponding to the datasets 10 years or more in length, all the ones where the observational error varied, and most of those where the data density varied – the residuals between the synthetic NEE data and the modelled NEE as used in the cost function were autocorrelated. An assumption of DA algorithms such as the Metropolis algorithm is that the residuals in the Likelihood are independent, and if they are not but treated as though they are this is likely to lead to biased parameter estimates as demonstrated in chapter 4 (figure 4.6). To date, no-one has properly addressed this issue but this has potentially large implications because of the fact that many terrestrial C modelling DA studies involve synthetic data such as Fox et al. (2009), McBean et al. (2011), and Trudinger et al. (2005). There are two main ways of eliminating or reducing to acceptable levels this autocorrelation:

(i) One is to use a modified cost function which can deal with synthetic data of this kind; a good example is McBean et al. (2011) who used the Metropolis algorithm to estimate model parameter and used the following Likelihood function:

$$P(D|f(\Theta)) = f(\Theta) - D = N(0,\sigma^2)$$

where $f(\Theta)$ is the model, Θ represents the parameters, D is the data and σ^2 represents the error or variance of the data. As can be seen in the above formula, McBean et al. (2011) ensured that the residuals in the Likelihood were independent.

(ii) The other is to reduce the density of points in the dataset; evidence that this results in independent residuals can be seen from run 8 of the Metropolis algorithm runs in chapter 4, which corresponded to a dataset with only 20% of data present and whose parameter estimates were unbiased and whose uncertainty intervals encompassed the truth.

The convergence to a local minima for the 11 'bad' runs from the Metropolis algorithm in chapter 4 could also be overcome by other means. First of all widening the domain of the starting locations of the chains, as opposed to them starting from the same location in the parameter space, may help; for example if the starting location all happen to be near a local minima this may be problematic. Secondly, using a continuous proposal distribution, rather than a discrete one, when stepping around the parameter space may also be beneficial.

(2) The novelty of chapter 3 was that it demonstrated for the first time the emulation of a DA scheme. This is significant because it offers a way of gaining insight into the performance of a DA scheme and in chapter 3 was used to assess how sensitive the parameter estimates were to changes in the initial conditions. Other uses of DA emulation include: (i) sensitivity analysis with other types of inputs or outputs, e.g. the length of the error bars on parameter estimates; (ii) uncertainty analysis, that is quantifying how uncertainty in an input (e.g. an initial condition or a prior estimate of a parameter) propagates forward to an output (e.g. a posterior parameter estimate). Both these sensitivity analysis and uncertainty analysis tools could help the implementation of DA algorithms, for example to help determine suitable uncertainty bounds on inputs or priors. The caveat of emulating a DA scheme is that an extra source of uncertainty needs to be included in the emulation process, and that is the fact that the DA algorithm is not deterministic, although in most cases approximates a deterministic function.

(3) Chapter 3 demonstrated that wrongly estimating the initial conditions of the model can have a large impact on correctly estimating model parameters. In particular if the initial value of the labile C pool is under-estimated by a reasonable amount or severely over-estimated, the resulting trajectories of the C pools tend to be much further from the truth than if this initial condition is over-estimated by a smaller amount (see section 6.1 for details). Treating these initial conditions as parameters overcomes these problems – in particular we found that although two of the three

initial conditions treated as parameters had modal values far from the truth and whose uncertainty intervals did not contain truth, this had no detrimental impact on the posterior estimates of the other model parameters. Moreover the trajectories of the C pools were not comprised unlike when these initial conditions were not treated as parameters (figures 3.10a and 3.10b). Whilst treating initial conditions as parameters is reasonably common, no known study had compared the effect on the posterior parameter distributions of treating the initial conditions as fixed versus treating them as parameters, using the same DA scheme.

(4) Chapter 4 showed that for the EnKF algorithm, increasing the length of the dataset improves parameter estimates and/or reduces the corresponding uncertainty whilst still containing the truth; this is also true for the estimates of the parameters and uncertainties from the Metropolis algorithm but there is less evidence to support this because the runs corresponding to dataset length of 10 years or more were discarded. For DALEC, between 2 and 5 years of daily NEE data was sufficient for good parameter estimates and NEE forecasts. For models of greater complexity than DALEC, the results do not indicate whether between 2 and 5 years of data at the daily time-step would be sufficient. However, what has been presented here is a first step in estimating the length of data needed for good parameter estimates and forecasts for more complex models.

(5) Chapter 4 also indicated that the posterior uncertainty of the model parameters remains unchanged with increases in the observational error. This is intriguing and potentially important because it suggests that assimilating very noisy data is just as good as assimilating non-noisy data. A practical application of this is the florescent data retrieved from the GOSAT satellite (Guanter et al., 2012). It is hoped that this data will be assimilated into a model, however there has been concern over this due to the excessively noisy nature of the signal. The results from this thesis suggest that despite this noisy data, assimilating it into a model will still result in near optimal solutions to parameters and model states.

(6) Chapter 5 demonstrated that unrealistic features of the MODIS LAI dataset can have negative impacts (such as being out of phase or far from the truth) on the post-assimilation estimation of important quantities, such as NEE. Processing the data, in order to reduce the impact of these unrealistic features such as smoothing, improved the accuracy of modelled NEE and GPP. The probable reason why the

s.d. values for the MODIS LAI are small at times is because when the radiative transfer model (RTM) is inverted to produce these estimates, all of the other ecological quantities within the RTM and within each 1km by 1km pixel are kept fixed. At the moment the s.d. in the RTM output (i.e. the LAI) is derived from the propagating the uncertainty from the inputs (the reflectance) through to the outputs. This is fine if the vegetation is homogeneous within the 1km by 1km pixel, but if not then this lack of representation of variation could cause a bias in the LAI uncertainty. It is therefore recommended that the RTM is modified to account for this extra source of uncertainty in the MODIS LAI s.d. calculation.

6.4 Unanswered questions / future work

The work in this thesis could be extended in many different ways.

(1) *Taking account of the stochastic nature of a DA scheme when emulating it.* In chapter 3 the Metropolis algorithm was assumed to be deterministic, because it is approximately so, but in reality it is stochastic. Therefore, when emulating a stochastic DA scheme such as the Metropolis algorithm for practical purposes, it is important that the additional variation in the parameter estimates arising from this stochastic feature is taken into account.

(2) *Assessing the sensitivity of parameter estimates to varying data lengths and data density using real data.* In chapter four, the main reason for using synthetic data was because flux sites typically have no more than 10 years of measurements and there would be no way of creating sets of real measurements which would be identical except for their level of observational error. However the study could be partially replicated using real data by comparing the estimates of parameters using 1, 2 and 5 year subsets of a 10 year dataset (as well as the complete dataset itself) and by creating duplicate datasets with different percentages of measurement-free days.

(3) *Repeating the MCMC runs from chapter using a more traditional variant of the Metropolis algorithm.* The primary reason for doing this is to assess whether we would get the same lack of convergence to the global minimum for the same 11 runs as was observed in the results in chapter four. The following two features are

aspects of a traditional variant of the Metropolis algorithm not adopted for this thesis:

a. The chains of the algorithm being initialised from random and different locations in the parameter space;
b. The proposal distribution being continuous, so as a movement around the parameters space are not multiples of the step-size but rather that step-size acts as a parameter in the distribution, e.g. a variance type term if the proposal is a Normal distribution.

(4) *Assessing the effect of assimilating a dataset to estimate parameters and uncertainties and prescribing the wrong observational uncertainty.* In chapter four we assessed the effect of varying the observational uncertainty on estimates of parameters and uncertainties from two DA schemes. It is relevant and important to determine the impact of wrongly prescribing the observational error in a similar type of parameter estimation scenario. This kind of experiment could apply to state estimation DA problems as well.

(5) *Testing a modified version of a radiative transfer model where the other variables are not kept fixed when the RTM is inverted to estimate LAI.* In light of the findings of chapter five, it is suggested that the radiative transfer model is modified in this way in order to provide improved estimates of the s.d. term. Furthermore, perhaps a more Bayesian framework could be incorporated? Ideally there would be two s.d. values for each LAI data point: one which describes the variation in LAI spatially within the 1km by 1km grid square (i.e how homogeneous the landscape is) and one which describes the uncertainty. At the very least, this would provide a clearer interpretation for the s.d. value than is currently provided.

(6) *Replicating the experiment in chapter five but estimate parameters as well as states.* A shortcoming of the work in chapter five is that the parameters were kept fixed. It would be relevant and interesting to understand the effects on the unrealistic features of the MODIS LAI data to DA estimates of parameter and uncertainties. In light of the results of chapter four, we wouldn't necessarily expect the parameter estimates and uncertainties to be affected too greatly by the temporal variability as long as the LAI dataset was long enough.

(7) *Addressing the implications of the observational error results from chapter four.* Chapter four found that the DA parameter estimates and uncertainties are insensitive to varying observational error. This has huge implications, in particular for the GOSAT fluorescent data (Guanter et al., 2012), this means that when assimilating it into a model, it does not matter if it is noisy. Other more sophisticated pieces of work would help to build on whether this 'insensitivity to observational error' claim is reliable:

   a. Repeating the observational error runs, but with synthetic data generated using another model, e.g. DALEC-water or SIPNET.
   b. Replicating observational error runs with real data if possible.
   c. Repeating the observational error runs but using less than 10 years, e.g. 2.5 years and 5 years. This would determine whether we see the same parameter results as the original done in chapter four which used 10 years. This would help determine what length of dataset one needs in order to get this insensitivity to large observational error.

6.5 Recommendations

As a consequence of the findings of this thesis, the following recommendations are made:

- Unknown initial conditions of a model should be treated as parameters. While it is reasonably common for this to be done, the results presented here show the consequences of treating these unknown as fixed and estimated by other means such as from site inventory data or from model spin up. In particular, treating the initial conditions as fixed imposes the strict condition that they must be estimated very precisely in order for the DA derived parameter estimates to be close to their true values, and within a specific range in order to the trajectories of the modelled C pools to be accurate. In contrast if the initial conditions are treated as parameters, wrongly estimating them has no noticeable negative impact on other parts of the model (chapter 3).
- There is strong evidence for the EnKF and some evidence for the Metropolis algorithm that these are good DA scheme to use for accurately estimating model parameters and predicting future states, from assimilating only one stream of

daily NEE data. However, it is better to assimilate longer datasets, particularly ones greater than one year (chapter 4).

- When using the Metropolis algorithm to estimate the parameters of a model using synthetic data, it is essential to check that the residuals between the data and the model in the cost function are independent in time. For the DALEC model we found that when the dataset was longer than 5 years or when more than 20% of the data were present in the dataset, the residuals were autocorrelated which resulted in biased parameter estimates with unrealistic uncertainties (chapter 4).
- It is strongly recommended that the MODIS LAI observations should be processed to account for unrealistic features, such as excessive variation and unrealistic errors, before being assimilated into a model to improve the estimates of its states (chapter 5). This processing, to at least provide more realistic s.d. values, should be done by the research group from Boston University (Knyazikhin et al., 1998, Myneni et al., 2002) responsible for the radiative transfer model in providing the MODIS LAI estimates for general use. This would require modifying the current version of the radiative transfer model so as the other variables are not kept fixed when the RTM is inverted to estimate LAI; further details can be found in point (5) of the previous section.

Appendix A

APPENDIX A1

I need to prove that:

$$\mathbf{P}_k^- = \mathbf{M}_{k-1}\mathbf{P}_{k-1}\mathbf{M}_{k-1}^T + \mathbf{Q}_{k-1} \quad \text{(A1)}$$

And $$\mathbf{P}_k = \mathbf{S}_k\mathbf{P}_k^-\mathbf{S}_k^T + \mathbf{K}_k\mathbf{R}_k\mathbf{K}_k^T \quad \text{(A2)}$$

We first derive the formula for $\mathbf{P}_k$.

$\mathbf{P}_k$ is defined as $\mathbf{P}_k = \mathbf{E}\left[(\hat{\mathbf{x}}_k - \mathbf{x}_k^t)(\hat{\mathbf{x}}_k - \mathbf{x}_k^t)^T\right]$.

Let, $\mathbf{e}_k = \hat{\mathbf{x}}_k - \mathbf{x}_k^t$ and $\mathbf{e}_k^- = \hat{\mathbf{x}}_k^- - \mathbf{x}_k^t$, then $\mathbf{P}_k = \mathbf{E}\left(\mathbf{e}_k\mathbf{e}_k^T\right)$

Now, $\mathbf{e}_k = \hat{\mathbf{x}}_k - \mathbf{x}_k^t$

$= \left[\hat{\mathbf{x}}_k^- + \mathbf{K}_k(\mathbf{y}_k - \mathbf{H}_k\hat{\mathbf{x}}_k^-)\right] - \mathbf{x}_k^t$ [from equation (2.9)]

$= (\hat{\mathbf{x}}_k^- - \mathbf{x}_k^t) + \mathbf{K}_k(\mathbf{H}_k\mathbf{x}_k^t + \mathbf{v}_k - \mathbf{H}_k\hat{\mathbf{x}}_k^-)$ [using the observation equation in subsection 2.4.1]

$= \mathbf{e}_k^- + \mathbf{K}_k(\mathbf{v}_k - \mathbf{H}_k\mathbf{e}_k^-)$

$= (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{e}_k^- + \mathbf{K}_k\mathbf{v}_k$

$= \mathbf{S}_k\mathbf{e}_k^- + \mathbf{K}_k\mathbf{v}_k$ where $\mathbf{S}_k = (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)$

Then, $\mathbf{P}_k = \mathbf{E}\left(\mathbf{e}_k\mathbf{e}_k^T\right)$

$= \mathbf{E}\left[\left(\mathbf{S}_k\mathbf{e}_k^- + \mathbf{K}_k\mathbf{v}_k\right)\left(\mathbf{S}_k\mathbf{e}_k^- + \mathbf{K}_k\mathbf{v}_k\right)^T\right]$

$= \mathbf{E}\left[\mathbf{S}_k\mathbf{e}_k^-(\mathbf{e}_k^-)^T\mathbf{S}_k^T + \mathbf{S}_k\mathbf{e}_k^-(\mathbf{v}_k)^T\mathbf{K}_k^T + \mathbf{K}_k\mathbf{v}_k(\mathbf{e}_k^-)^T\mathbf{S}_k^T + \mathbf{K}_k\mathbf{v}_k(\mathbf{v}_k)^T\mathbf{K}_k^T\right]$

$= \mathbf{S}_k\mathbf{E}\left(\mathbf{e}_k^-(\mathbf{e}_k^-)^T\right)\mathbf{S}_k^T + \mathbf{S}_k\mathbf{E}\left(\mathbf{e}_k^-(\mathbf{v}_k)^T\right)\mathbf{K}_k^T + \mathbf{K}_k\mathbf{E}\left(\mathbf{v}_k(\mathbf{e}_k^-)^T\right)\mathbf{S}_k^T + \mathbf{K}_k\mathbf{E}\left(\mathbf{v}_k(\mathbf{v}_k)^T\right)\mathbf{K}_k^T$

For reasons not given here, we assume that $\mathbf{E}\left(\mathbf{e}_k^-(\mathbf{v}_k)^T\right) = 0$ and $\mathbf{E}\left(\mathbf{v}_k(\mathbf{e}_k^-)^T\right) = 0$. This results in:

$$\mathbf{P}_k = \mathbf{S}_k\mathbf{E}\left(\mathbf{e}_k^-(\mathbf{e}_k^-)^T\right)\mathbf{S}_k^T + \mathbf{K}_k\mathbf{E}\left(\mathbf{v}_k(\mathbf{v}_k)^T\right)\mathbf{K}_k^T$$

Or, $\boxed{\mathbf{P}_k = \mathbf{S}_k\mathbf{P}_k^-\mathbf{S}_k^T + \mathbf{K}_k\mathbf{R}_k\mathbf{K}_k^T}$

This can be simplified to: $\mathbf{P}_k = \mathbf{S}_k\mathbf{P}_k^-$, which is best explained by looking at the scalar case of the KF. In equation (2.6), recall that $s_a^2 = \frac{s_f^2 s_o^2}{s_o^2 + s_f^2}$ - this is the same as $s_a^2 = (1-K)s_f^2$, where *K* is the Kalman gain given by $K = s_f^2/(s_o^2 + s_f^2)$.

To derive the formula for $\mathbf{P}_k^-$, we first note that:

$\mathbf{e}_k^- = \hat{\mathbf{x}}_k^- - \mathbf{x}_k^t$

$= \mathbf{M}_{k-1}\hat{\mathbf{x}}_{k-1} - (\mathbf{M}_{k-1}\mathbf{x}_{k-1}^t + \mathbf{w}_{k-1})$

[using the observation equation in subsection 2.2.1 and where $\hat{\mathbf{x}}_k = \mathbf{M}_{k-1}\hat{\mathbf{x}}_{k-1}$ is our prior estimate].

So, $\mathbf{e}_k^- = \mathbf{M}_{k-1}\mathbf{e}_{k-1} - \mathbf{w}_{k-1}$

Now, $\mathbf{P}_k^-$ is defined as $\mathbf{P}_k^- = \mathbf{E}\left(\mathbf{e}_k^-(\mathbf{e}_k^-)^T\right)$, and so:

$$\mathbf{E}\left(\mathbf{e}_k^-(\mathbf{e}_k^-)^T\right) = \mathbf{E}\left((\mathbf{M}_{k-1}\mathbf{e}_{k-1} - \mathbf{w}_{k-1})(\mathbf{M}_{k-1}\mathbf{e}_{k-1} - \mathbf{w}_{k-1})^T\right)$$

$$= \mathrm{M}_{k-1}\mathrm{E}\left(\mathrm{e}_{k-1}(\mathrm{e}_{k-1})^T\right)\mathrm{M}_{k-1}^T - \mathrm{M}_{k-1}\mathrm{E}\left(\mathrm{e}_{k-1}(\mathrm{w}_{k-1})^T\right) - \mathrm{E}\left(\mathrm{w}_{k-1}(\mathrm{e}_{k-1})^T\right)\mathrm{M}_{k-1}^T + \mathrm{E}\left(\mathrm{w}_{k-1}(\mathrm{w}_{k-1})^T\right)$$

For reasons not explained here, $\mathbf{E}\left(\mathbf{e}_{k-1}(\mathbf{w}_{k-1})^T\right) = 0$ and $\mathbf{E}\left(\mathbf{w}_{k-1}(\mathbf{e}_{k-1})^T\right) = 0$. So, we get:

$$\mathbf{P}_k^- = \mathbf{M}_{k-1}\mathbf{E}\left(\mathbf{e}_{k-1}(\mathbf{e}_{k-1})^T\right)\mathbf{M}_{k-1}^T + \mathbf{E}\left(\mathbf{w}_{k-1}(\mathbf{w}_{k-1})^T\right)$$

Or $\boxed{\mathbf{P}_k^- = \mathbf{M}_{k-1}\mathbf{P}_{k-1}\mathbf{M}_{k-1}^T + \mathbf{Q}_{k-1}}$

APPENDIX A2

Derivation that the Kalman gain matrix is: $\mathbf{K}_k = \mathbf{P}_k^-\mathbf{H}_k^T(\mathbf{H}_k\mathbf{P}_k^-\mathbf{H}_k^T + \mathbf{R}_k)^{-1}$

Recall that $\mathbf{P}_k = \mathbf{S}_k\mathbf{P}_k^-\mathbf{S}_k^T + \mathbf{K}_k\mathbf{R}_k\mathbf{K}_k^T$. We choose $\mathbf{K}_k$ such that the sum of the elements of the leading diagonal of $\mathbf{P}_k$ (i.e. the variances) are minimised.

Substituting $\mathbf{S}_k = (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)$ into $\mathbf{P}_k = \mathbf{S}_k\mathbf{P}_k^-\mathbf{S}_k^T + \mathbf{K}_k\mathbf{R}_k\mathbf{K}_k^T$ gives:

$$\begin{aligned}
\mathbf{P}_k &= (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{P}_k^-(\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)^T + \mathbf{K}_k\mathbf{R}_k\mathbf{K}_k^T \\
&= \mathbf{P}_k^- - \mathbf{P}_k^-\mathbf{H}_k^T\mathbf{K}_k^T - \mathbf{K}_k\mathbf{H}_k\mathbf{P}_k^- + \mathbf{K}_k\mathbf{H}_k\mathbf{P}_k^-\mathbf{H}_k^T\mathbf{K}_k^T + \mathbf{K}_k\mathbf{R}_k\mathbf{K}_k^T \\
&= \mathbf{P}_k^- - (\mathbf{P}_k^-)^T\mathbf{H}_k^T\mathbf{K}_k^T - \mathbf{K}_k\mathbf{H}_k\mathbf{P}_k^- + \mathbf{K}_k\mathbf{H}_k\mathbf{P}_k^-\mathbf{H}_k^T\mathbf{K}_k^T + \mathbf{K}_k\mathbf{R}_k\mathbf{K}_k^T \\
&\qquad\qquad \text{[since } \mathbf{P}_k^- \text{ is symmetric]} \\
&= \mathbf{P}_k^- - (\mathbf{K}_k\mathbf{H}_k\mathbf{P}_k^-)^T - \mathbf{K}_k\mathbf{H}_k\mathbf{P}_k^- + \mathbf{K}_k(\mathbf{H}_k\mathbf{P}_k^-\mathbf{H}_k^T + \mathbf{R}_k)\mathbf{K}_k^T
\end{aligned}$$

To minimise the trace of $\mathbf{P}_k$ with respect to $\mathbf{K}_k$, we need to minimise the trace of each of the four terms. We do this by differentiating each of the four terms with respect to $\mathbf{K}_k$. The rules for differentiating matrices are similar to traditional differentiation. So for the first term, $\mathbf{P}_k^-$, since it does not involve a $\mathbf{K}_k$, it will become zero after differentiation. For the second and third terms which involve just one $\mathbf{K}_k$ each, we use the following result:

$$\frac{d[tr(\mathbf{AC})]}{d\mathbf{A}} = \mathbf{C}^T$$

Since $tr(\mathbf{K}_k\mathbf{H}_k\mathbf{P}_k^-)^T = tr(\mathbf{K}_k\mathbf{H}_k\mathbf{P}_k^-)$, it follows that:

$$\frac{d[tr(\mathbf{K}_k\mathbf{H}_k\mathbf{P}_k^-)^T]}{d\mathbf{K}_k} = \frac{d[tr(\mathbf{K}_k\mathbf{H}_k\mathbf{P}_k^-)]}{d\mathbf{K}_k} = (\mathbf{H}_k\mathbf{P}_k^-)^T$$

For the last of the four terms of the $\mathbf{P}_k$, which involves two $\mathbf{K}_k$ matrices, we use the following result, where **C** is assumed to be a symmetric matrix:

$$\frac{d[tr(\mathbf{ACA}^T)]}{d\mathbf{A}} = 2\mathbf{AC}$$

Since $\mathbf{H}_k\mathbf{P}_k^-\mathbf{H}_k^T + \mathbf{R}_k$ is symmetric,

[ie since $(\mathbf{H}_k\mathbf{P}_k^-\mathbf{H}_k^T+\mathbf{R}_k)^T=(\mathbf{H}_k\mathbf{P}_k^-\mathbf{H}_k^T)^T+\mathbf{R}_k^T=(\mathbf{H}_k^T)^T(\mathbf{P}_k^-)^T(\mathbf{H}_k)^T+\mathbf{R}_k^T$
$=\mathbf{H}_k\mathbf{P}_k^-\mathbf{H}_k^T+\mathbf{R}_k$]
we get:

$$\frac{d\left[tr(\mathbf{K}_k(\mathbf{H}_k\mathbf{P}_k^-\mathbf{H}_k^T+\mathbf{R}_k)\mathbf{K}_k^T\right]}{d\mathbf{K}_k}=2\mathbf{K}_k(\mathbf{H}_k\mathbf{P}_k^-\mathbf{H}_k^T+\mathbf{R}_k)$$

Combining all these results, we get:

$$\frac{d\left[tr(\mathbf{P}_k\right]}{d\mathbf{K}_k}=\mathbf{0}-(\mathbf{H}_k\mathbf{P}_k^-)^T-(\mathbf{H}_k\mathbf{P}_k^-)^T+2\mathbf{K}_k(\mathbf{H}_k\mathbf{P}_k^-\mathbf{H}_k^T+\mathbf{R}_k)$$

Or
$$\frac{d\left[tr(\mathbf{P}_k)\right]}{d\mathbf{K}_k}=2\left[-(\mathbf{H}_k\mathbf{P}_k^-)^T+\mathbf{K}_k(\mathbf{H}_k\mathbf{P}_k^-\mathbf{H}_k^T+\mathbf{R}_k)\right] \qquad \text{(B1)}$$

In order to minimise the trace of $\mathbf{P}_k$, the left-hand side of (B1) needs to set to zero.
So, the equation becomes:

$$\mathbf{0}=2\left[-(\mathbf{H}_k\mathbf{P}_k^-)^T+\mathbf{K}_k(\mathbf{H}_k\mathbf{P}_k^-\mathbf{H}_k^T+\mathbf{R}_k)\right]$$

Or

$$-(\mathbf{H}_k\mathbf{P}_k^-)^T+\mathbf{K}_k(\mathbf{H}_k\mathbf{P}_k^-\mathbf{H}_k^T+\mathbf{R}_k)=\mathbf{0}$$

Solving for $\mathbf{K}_k$ gives:

$$\boxed{\mathbf{K}_k=\mathbf{P}_k^-\mathbf{H}_k^T(\mathbf{H}_k\mathbf{P}_k^-\mathbf{H}_k^T+\mathbf{R}_k)^{-1}}$$

APPENDIX A3

We need to show that $\mathbf{X}^a=\mathbf{X}^f+\mathbf{P}_e^f\mathbf{H}^T(\mathbf{H}\mathbf{P}_e^f\mathbf{H}^T+\mathbf{R}_e)^{-1}(\mathbf{D}-\mathbf{H}\mathbf{X}^f)$ (C1)
is equivalent to: $\mathbf{X}^a=\mathbf{X}^f+\mathbf{X}'\mathbf{X}'^T\mathbf{H}^T(\mathbf{H}\mathbf{X}'\mathbf{X}'^T\mathbf{H}^T+\mathbf{D}'\mathbf{D}'^T)^{-1}(\mathbf{D}-\mathbf{H}\mathbf{X})$ (C2)
We do this by finding expressions for $\mathbf{P}_e^f$ and that $\mathbf{R}_e$.
Recall from equation (2.15) in subsection 2.2.4, we had:

$$\tilde{\mathbf{P}}_k^f=\frac{1}{N-1}\left[\hat{\mathbf{X}}_k^f-\overline{\mathbf{X}_k^f}\right]\left[\hat{\mathbf{X}}_k^f-\overline{\mathbf{X}_k^f}\right]^T$$

Using the notation in (C1), this can be rewritten as $\mathbf{P}_e^f=\frac{1}{N-1}\left[\mathbf{X}-\overline{\mathbf{X}}\right]\left[\mathbf{X}-\overline{\mathbf{X}}\right]^T$

Or: $\boxed{\mathbf{P}_e^f=\frac{\mathbf{X}'(\mathbf{X}')^T}{N-1}}$ where $\mathbf{X}'=\mathbf{X}-\overline{\mathbf{X}}$

For $\mathbf{R}_e$, recall the expression in subsection 2.2.4 for a general *k*:

$$\mathbf{Y}_k=\left[\mathbf{y}_k\mid\mathbf{y}_k\mid\cdots\mathbf{y}_k\right]+\left[\delta_1\mid\delta_2\mid\cdots\delta_N\right],\qquad \delta_i\sim N_{n_x}(\mathbf{0},\mathbf{R}_k)\ \forall\ i$$

Again for consistency with notation of other authors, we use a **D** instead of an **Y** and drop the subscripts. We also write it in a slightly different way, ie:

$$\mathbf{D}=\left[\mathbf{d}_1\mid\mathbf{d}_2\mid\cdots\mathbf{d}_N\right]\ \text{ where }\ \mathbf{d}_i=\mathbf{d}+\delta_i\ \text{ and }\ \delta_i\sim N_n(\mathbf{0},\mathbf{R}_e)$$

In a similar way to $\mathbf{P}_e^f$, $\mathbf{R}_e$ is defined as:

$$\mathbf{R}_e = \frac{\mathbf{D}'(\mathbf{D}')^T}{N-1} \qquad \text{where } \mathbf{D}' = \mathbf{D} - \overline{\mathbf{D}}$$

Now, $\mathbf{D}' = \mathbf{D} - \overline{\mathbf{D}} = \left[\mathbf{d}_1 \mid \mathbf{d}_2 \mid \cdots \mid \mathbf{d}_N\right] - \left[\overline{\mathbf{d}} \mid \overline{\mathbf{d}} \mid \cdots \mid \overline{\mathbf{d}}\right]$

$= \left[\mathbf{d}_1 - \overline{\mathbf{d}} \mid \mathbf{d}_2 - \overline{\mathbf{d}} \mid \cdots \mid \mathbf{d}_N - \overline{\mathbf{d}}\right]$

$= \left[(\mathbf{d} + \delta_1) - (\mathbf{d} + \overline{\delta}) \mid (\mathbf{d} + \delta_2) - (\mathbf{d} + \overline{\delta}) \mid \cdots \mid (\mathbf{d} + \delta_N) - (\mathbf{d} + \overline{\delta})\right]$

$= \left[\delta_1 - \overline{\delta} \mid \delta_2 - \overline{\delta} \mid \cdots \mid \delta_N - \overline{\delta}\right]$

$= \left[\delta_1 \mid \delta_2 \mid \cdots \mid \delta_N\right]$ (since $\overline{\delta} = \mathbf{0}$)

So, $\boxed{\mathbf{R}_e = \frac{\mathbf{D}'(\mathbf{D}')^T}{N-1}}$ where $\mathbf{D}' = (\delta_1, \delta_2, ..., \delta_N)$

We now substitute $\mathbf{P}_e^f = \mathbf{X}'(\mathbf{X}')^T/(N-1)$ and $\mathbf{R}_e = \mathbf{D}'(\mathbf{D}')^T/(N-1)$ into (C1) gives:

$$\mathbf{X}^a = \mathbf{X}^f + \frac{\mathbf{X}'(\mathbf{X}')^T}{N-1}\mathbf{H}^T\left(\mathbf{H}\frac{\mathbf{X}'(\mathbf{X}')^T}{N-1}\mathbf{H}^T + \frac{\mathbf{D}'(\mathbf{D}')^T}{N-1}\right)^{-1}(\mathbf{D} - \mathbf{H}\mathbf{X}^f)$$

We can now take out the 1/(N-1) common factors out of the bracket:

$$\mathbf{X}^a = \mathbf{X}^f + \frac{\mathbf{X}'(\mathbf{X}')^T}{N-1}\mathbf{H}^T\left(\frac{1}{N-1}\right)^{-1}\left(\mathbf{H}\mathbf{X}'(\mathbf{X}')^T\mathbf{H}^T + \mathbf{D}'(\mathbf{D}')^T\right)^{-1}(\mathbf{D} - \mathbf{H}\mathbf{X}^f)$$

Or: $\boxed{\mathbf{X}^a = \mathbf{X}^f + \mathbf{X}'(\mathbf{X}')^T\mathbf{H}^T\left(\mathbf{H}\mathbf{X}'(\mathbf{X}')^T\mathbf{H}^T + \mathbf{D}'(\mathbf{D}')^T\right)^{-1}(\mathbf{D} - \mathbf{H}\mathbf{X}^f)}$

Appendix B

R script for implementing the MmLHD

```
z=seq(from=0, to=1-1/40, length=40)
u1=z+runif(40,0,1/40)
u2=z+runif(40,0,1/40)
u3=z+runif(40n,0,1/40)
u4=z+runif(40,0,1/40)
x1=qunif(u1,20,200)
x2=qunif(u2,20,200)
x3=qunif(u3,20,200)
x4=qunif(u4,0.0005,0.20)
x=cbind(x1,x2,x3,x4)
x.matrix=matrix(x,ncol=4)
order1=read.table("G:/Ch3_REFLEX_followup_study/3Results/1Inputs/LHD_order.t
xt",header=F)
#order1 is a text file which gives the Maximin Latin Hypercube design for n=40
design points #consisting of 4 inputs, obtained from http://www.spacefillingdesigns.nl
(Husslage et al., 2006)
order=as.matrix(order1) + 1
Croots=x.matrix[,1][order[,1]]
Clitter=x.matrix[,2][order[,2]]
Clabile=x.matrix[,3][order[,3]]
LAIobserr=x.matrix[,4][order[,4]]
inputs=cbind(Croots,Clitter,Clabile,LAIobserr)
```

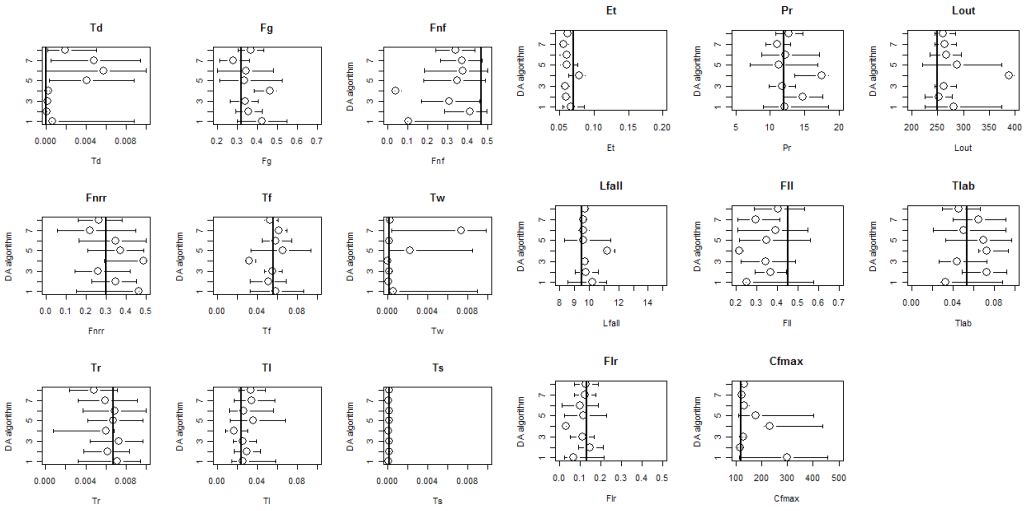


Figure B1 The estimates of the parameters and confidence intervals for the eight DA algorithms used in REFLEX. The DA schemes are the Genetic algorithm (1), the Ensemble Kalman filter (2), Metropolis algorithm (3), the SCEM-UA algorithm (4), Metropolis algorithm with spin-up (5), Metropolis algorithm with initial exploration (6), Combined Metropolis-Genetic algorithm (7), Metropolis algorithm with a beta prior (8).

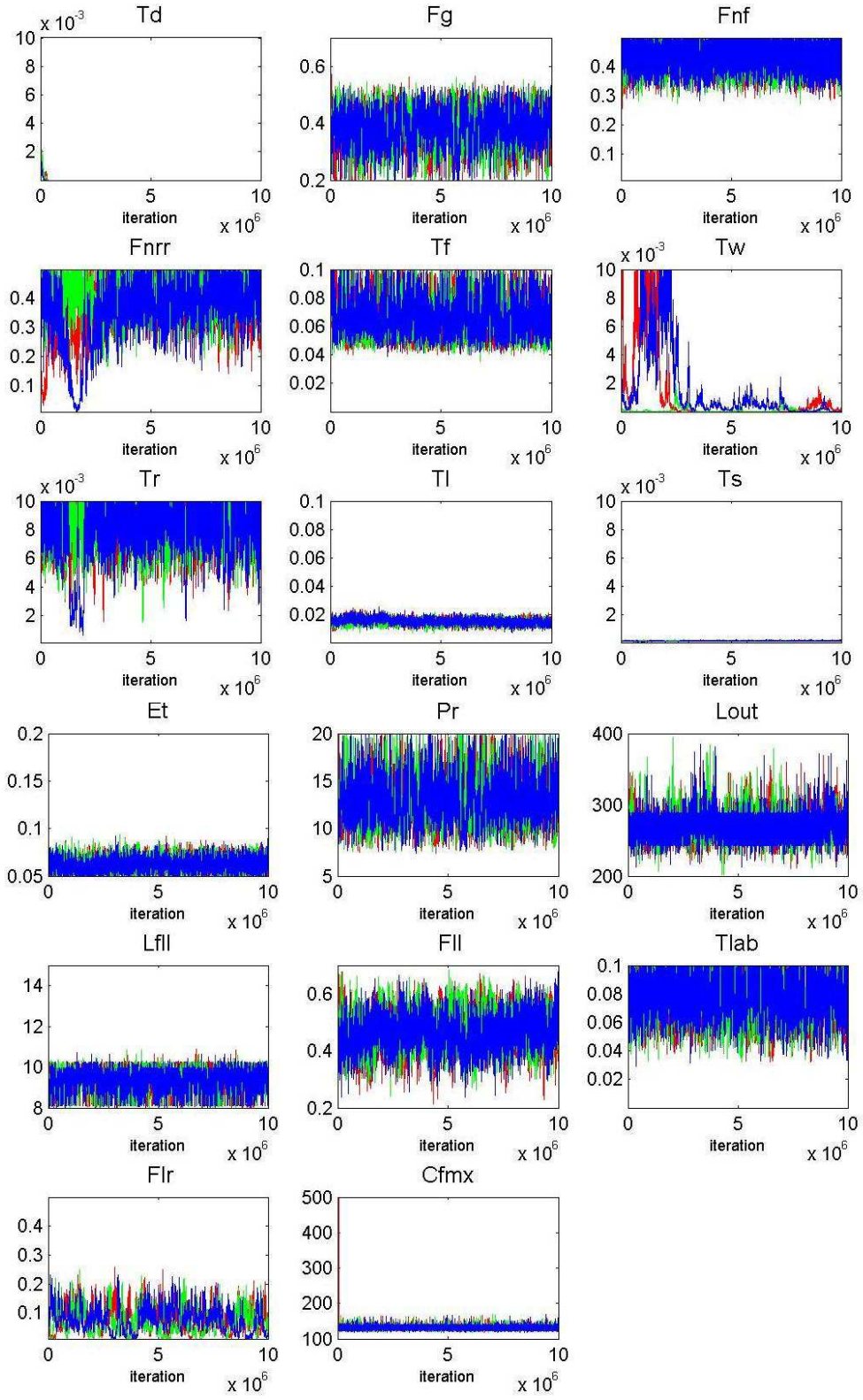


Figure B2 The trajectories of the three chains from the Metropolis algorithm for each of the 17 DALEC-D parameters.

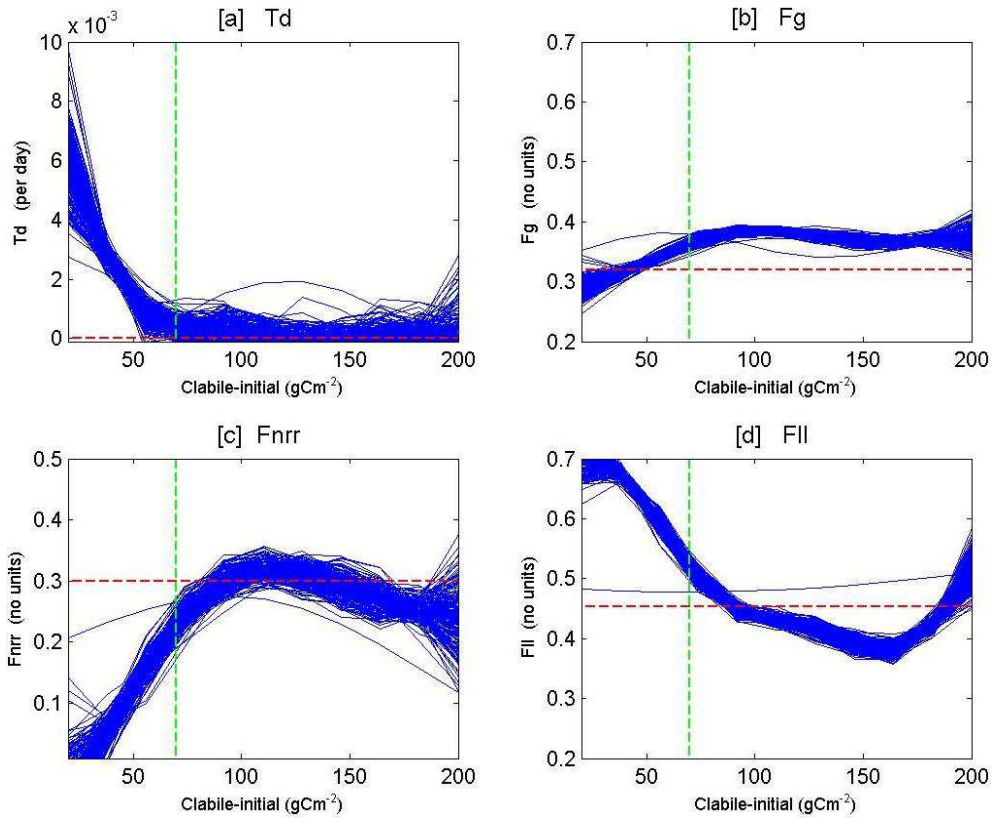


Figure B3 The expected value (blue lines) of T_d (a), F_g (b), F_{nrr} (c) and F_{ll} (d) for different values of the $C_{lab}(0)$ value. The red horizontal and green vertical dashed lines indicate the true values of the parameter and the true value of this input respectively.

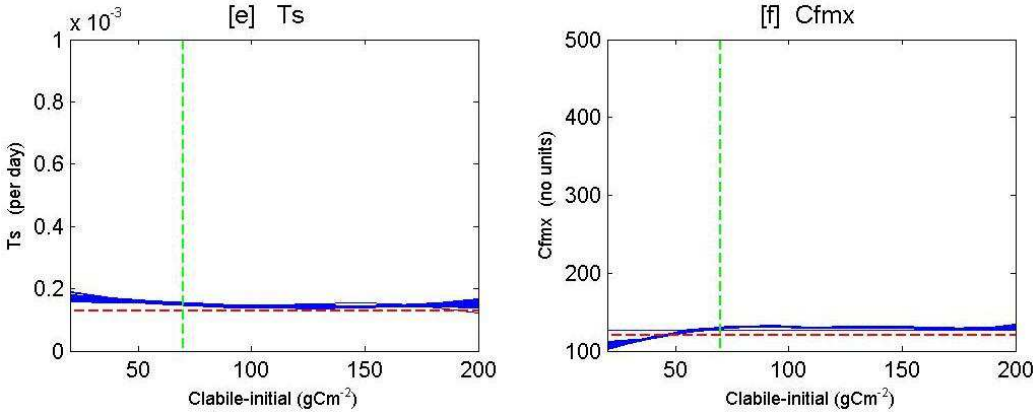


Figure B4 The expected value (blue lines) of T_s (e) and $C_{f,max}$ (f) for different values of the $C_{lab}(0)$ value. The red horizontal and green vertical dashed lines indicate the true values of the parameter and the true value of this input respectively.

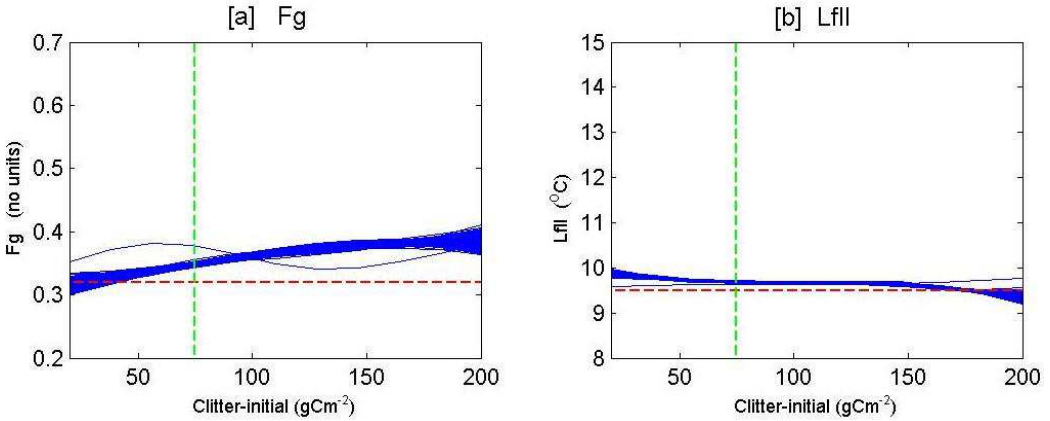


Figure B5 The expected value (blue lines) of F_g (a) and F_{ll} (b) for different values of the $C_{lit}(0)$ value. The red horizontal and green vertical dashed lines indicate the true values of the parameter and the true value of this input respectively.

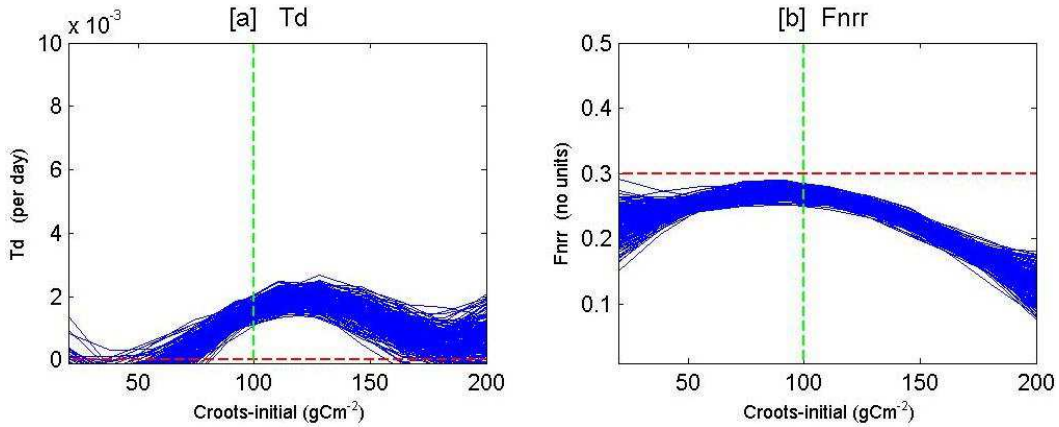


Figure B6 The expected value (blue lines) of T_d (a) and F_{nrr} (b) for different values of the $C_r(0)$ value. The red horizontal and green vertical dashed lines indicate the true values of the parameter and the true value of this input respectively.

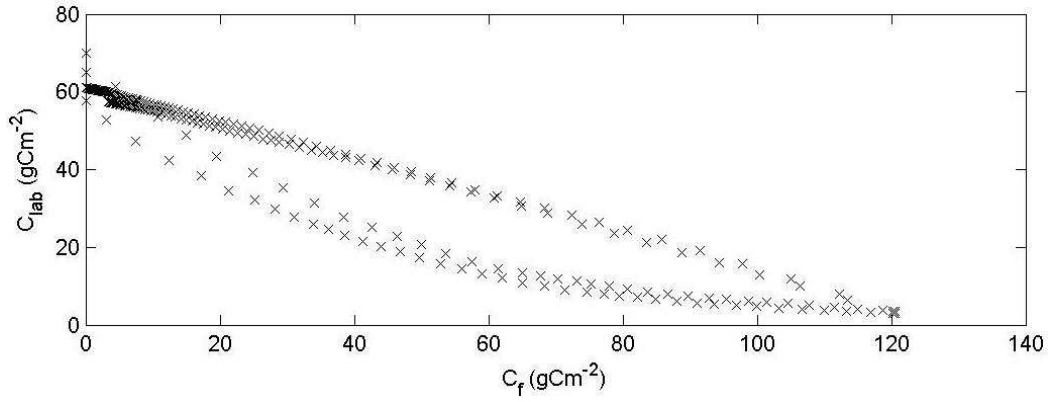


Figure B7 A scatter plot of daily values of Cf and Clab from the DALEC-D model using the true set of initial conditions and parameters as specified in table 3.1.

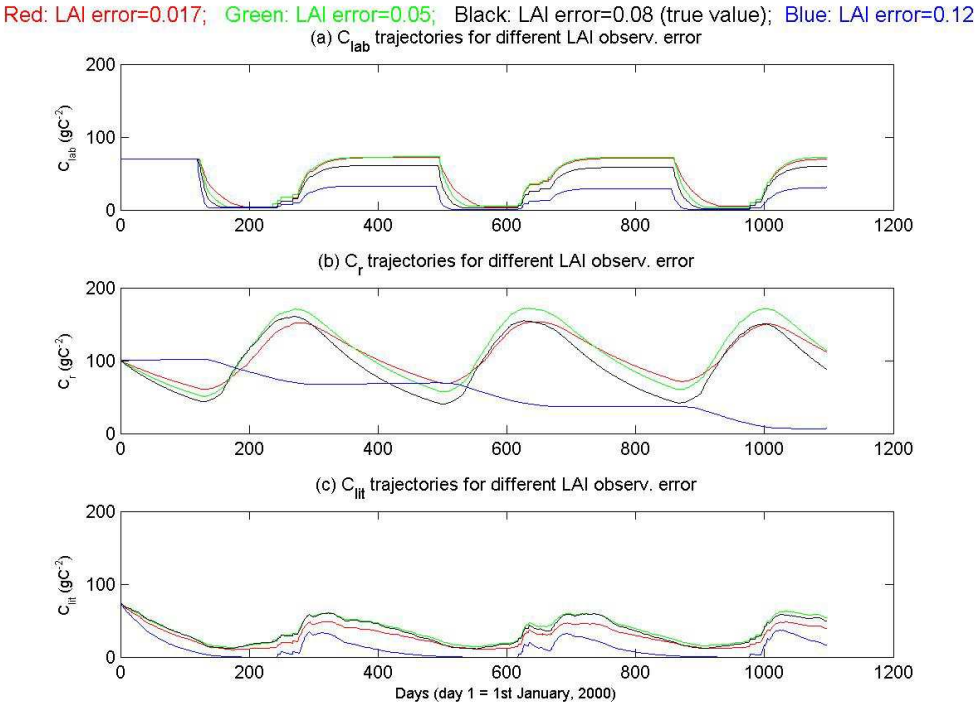


Figure B8 The trajectories of C_{lab}, C_r, and C_{lit} for different values of the LAI observational error using the emulator's estimates of the parameters corresponding to the values of the LAI observational error.

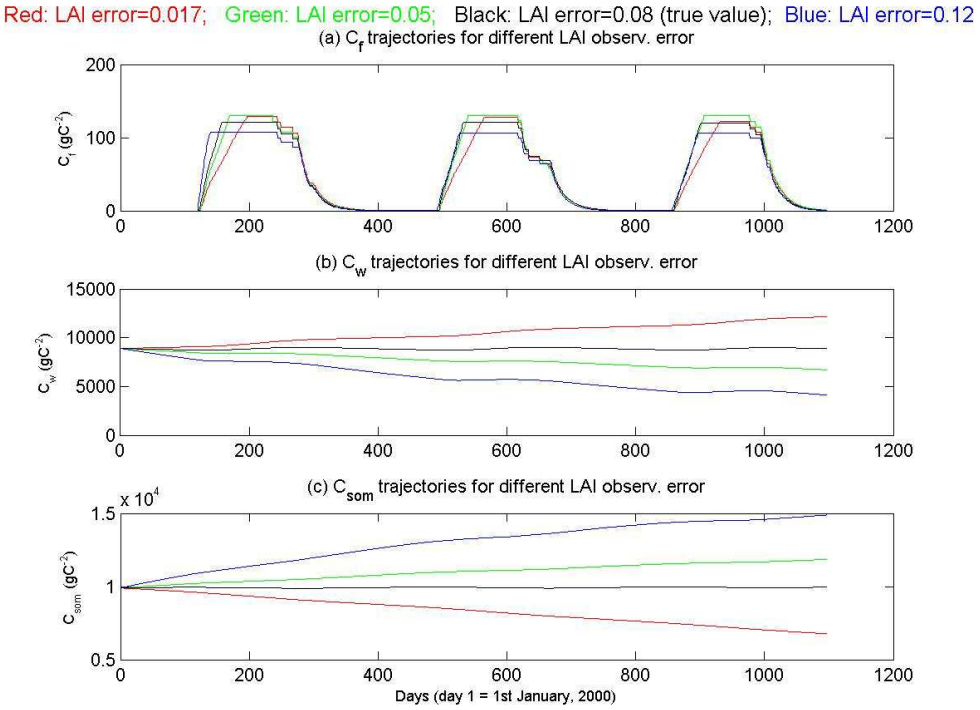


Figure B9 The trajectories of C_f, C_w, and C_{som} for different values of the LAI observational error, using the emulator's estimates of the parameters corresponding to the values of the LAI observational error.

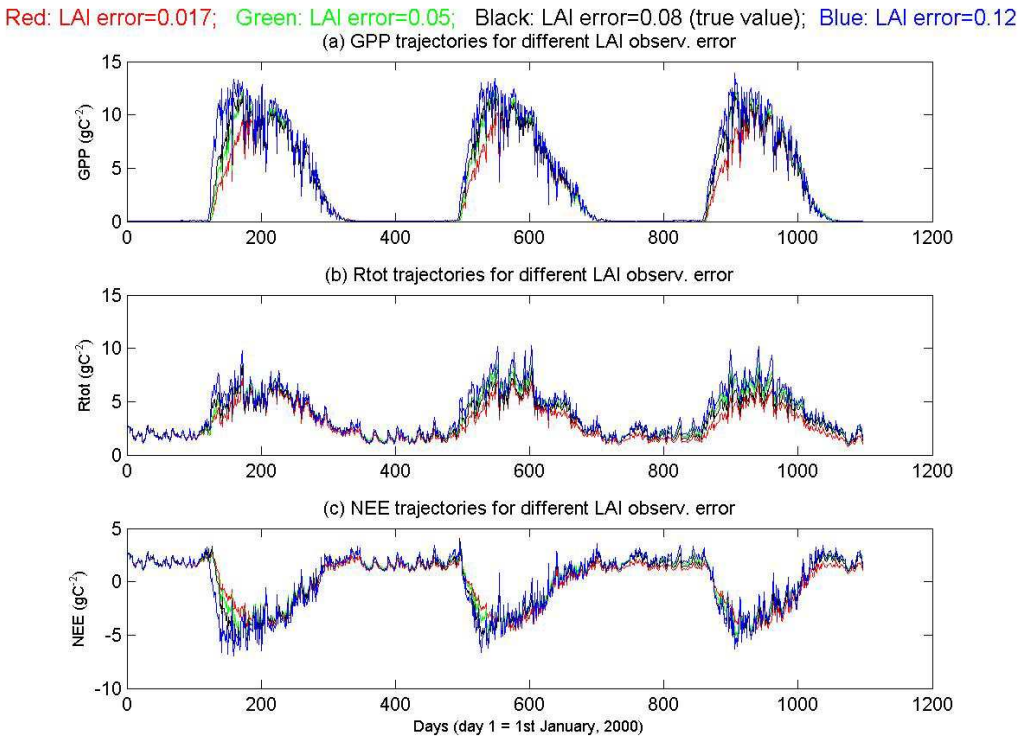
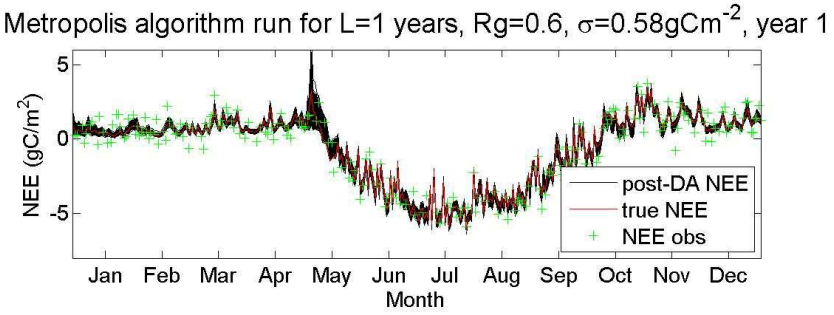


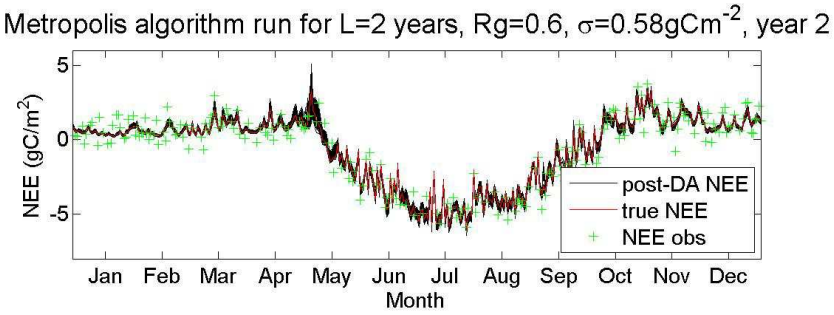
Figure B10 The trajectories of GPP, R_{tot}, and NEE for different values of the LAI observational error, using the emulator's estimates of the parameters corresponding to the values of the LAI observational value. Panel (d) shows the absolute difference between the NEE trajectories corresponding to LAI observational errors of 0.016, 0.05 and 0.11 m^2/m^2 and the true NEE trajectory.

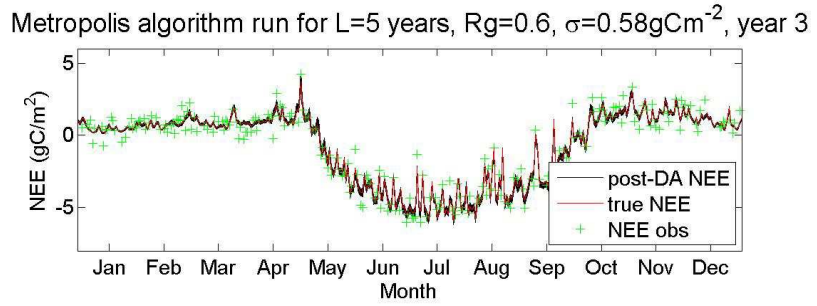
Appendix C

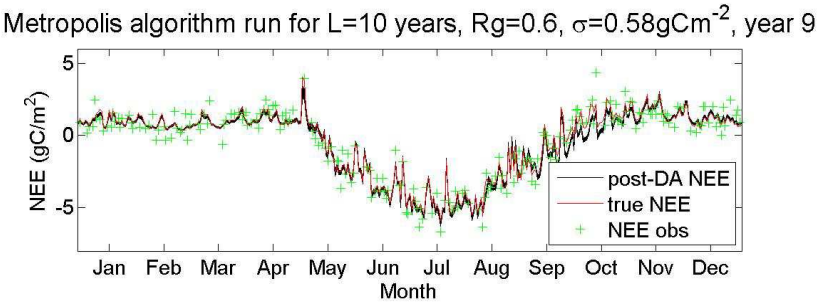
Plots of NEE trajectories for the Metropolis algorithm

Below are displayed the plots of NEE trajectories for all 15 of the runs for the Metropolis algorithm. For each run, DALEC-D is run forward using 500 randomly selected parameter sets from the posterior distribution. These are represented by the black lines. The red line shows the true NEE trajectory, and the green crosses are the observations. In each case, only one year of the trajectory is shown, for example for run 2, only the 2nd year is shown. This is because if the entire trajectory was displayed, much of the detail would be lost. For each run, the year that was chosen corresponds to a year representative of all the other years.

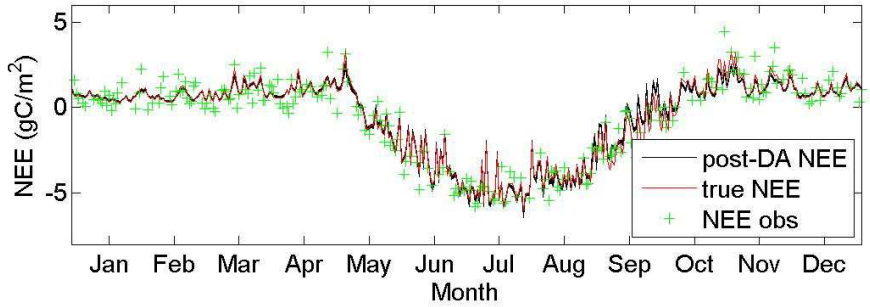




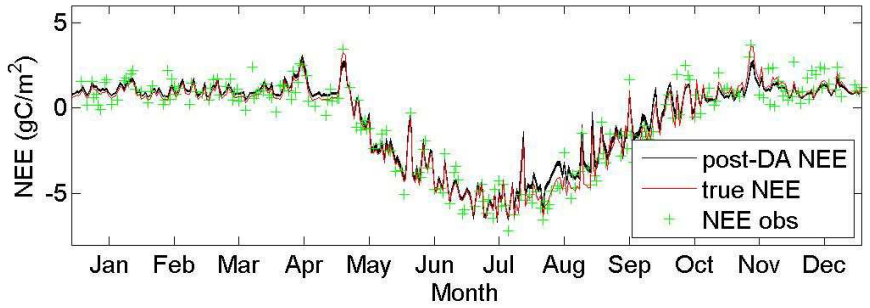




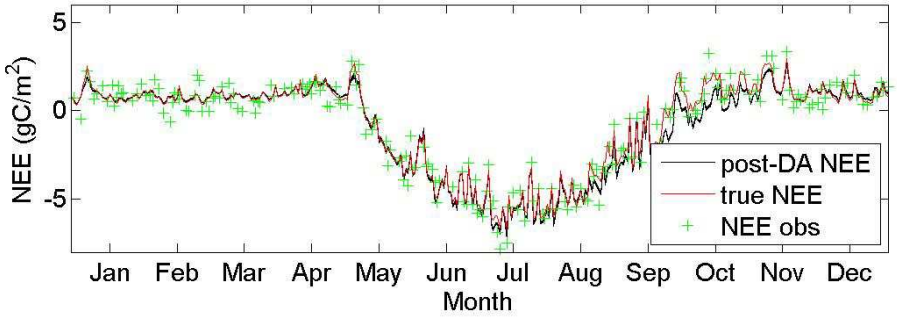
Metropolis algorithm run for L=20 years, Rg=0.6, σ=0.58gCm^{-2}, year 11



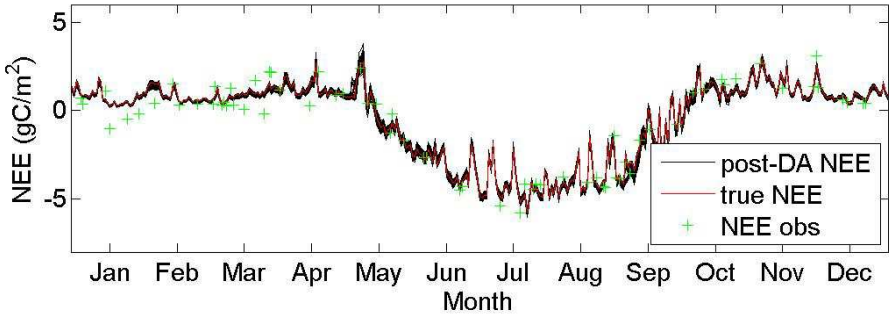
Metropolis algorithm run for L=30 years, Rg=0.6, σ=0.58gCm^{-2}, year 10

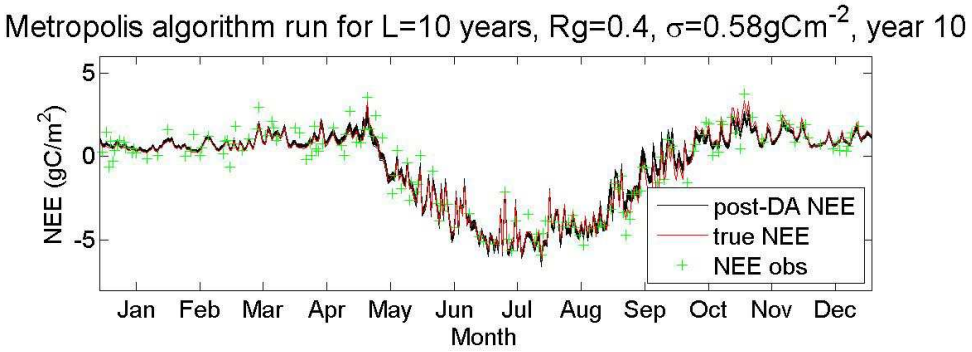


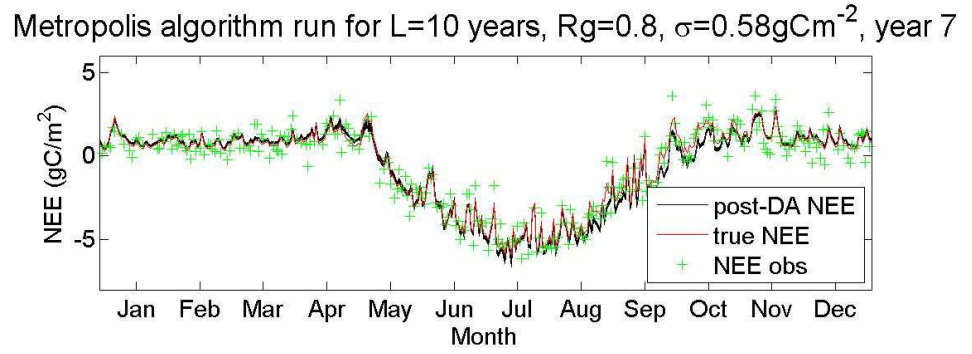
Metropolis algorithm run for L=50 years, Rg=0.6, σ=0.58gCm^{-2}, year 14

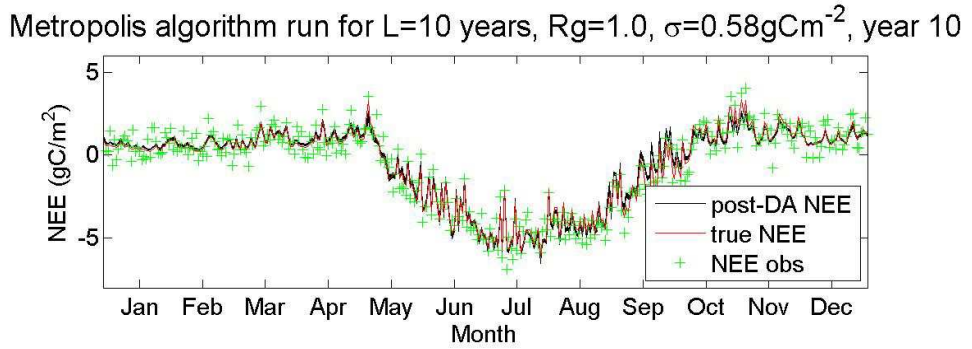


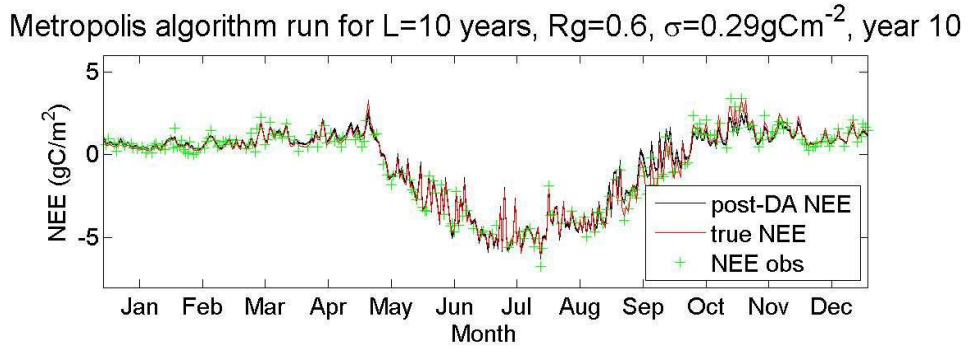
Metropolis algorithm run for L=10 years, Rg=0.2, σ=0.58gCm^{-2}, year 6

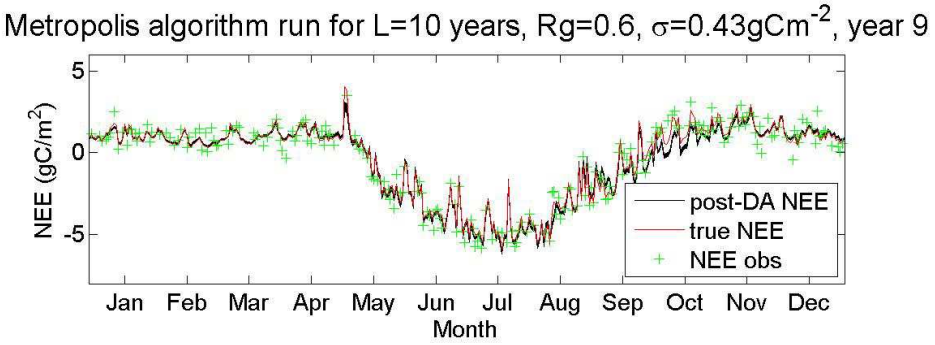


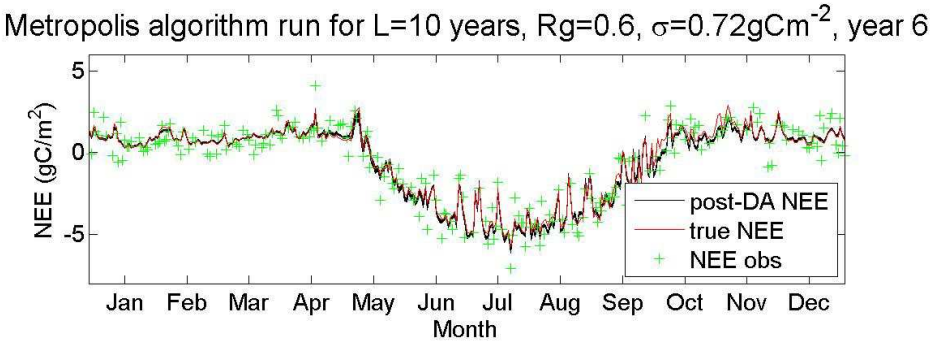

Metropolis algorithm run for L=10 years, Rg=0.4, σ=0.58gCm-2, year 10
NEE (gC/m2)
5
0
-5
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Month
post-DA NEE
true NEE
NEE obs

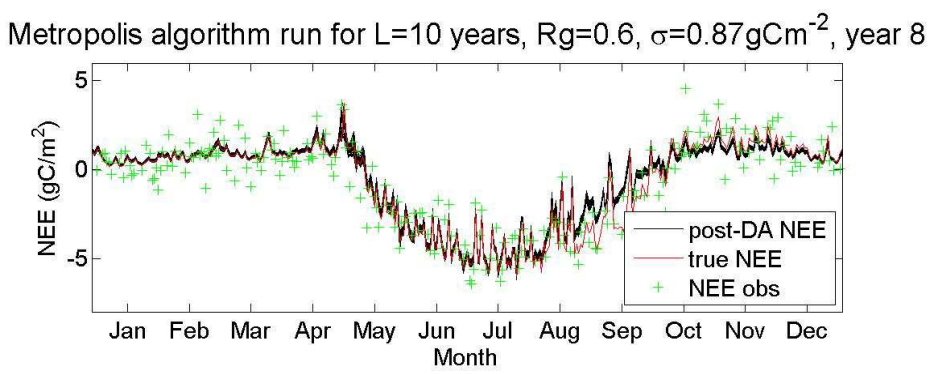

Metropolis algorithm run for L=10 years, Rg=0.8, σ=0.58gCm-2, year 7
NEE (gC/m2)
5
0
-5
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Month
post-DA NEE
true NEE
NEE obs


Metropolis algorithm run for L=10 years, Rg=1.0, σ=0.58gCm-2, year 10
NEE (gC/m2)
5
0
-5
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Month
post-DA NEE
true NEE
NEE obs


Metropolis algorithm run for L=10 years, Rg=0.6, σ=0.29gCm-2, year 10
NEE (gC/m2)
5
0
-5
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Month
post-DA NEE
true NEE
NEE obs

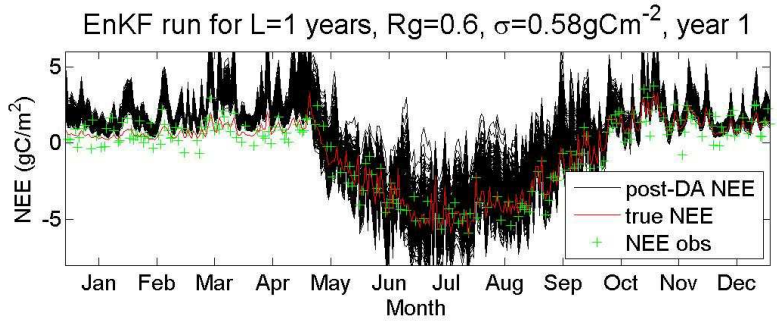

Metropolis algorithm run for L=10 years, Rg=0.6, σ=0.43gCm-2, year 9
NEE (gC/m2)
5
0
-5
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Month
post-DA NEE
true NEE
NEE obs

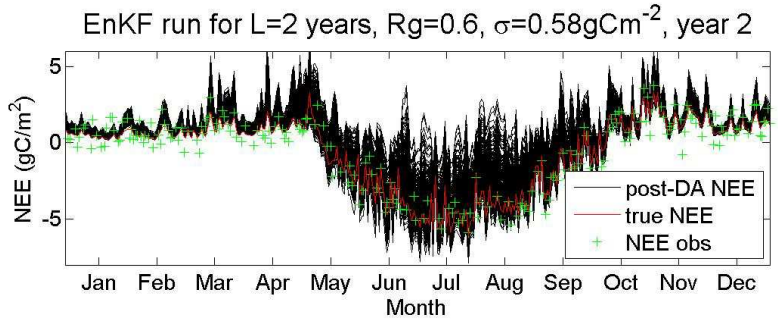

Metropolis algorithm run for L=10 years, Rg=0.6, σ=0.72gCm-2, year 6
NEE (gC/m2)
5
0
-5
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Month
post-DA NEE
true NEE
NEE obs

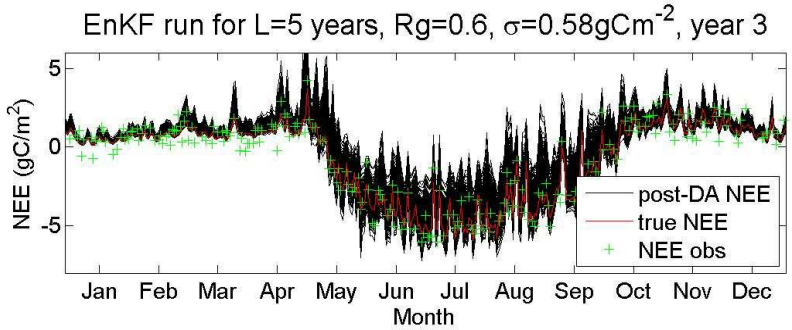

Metropolis algorithm run for L=10 years, Rg=0.6, σ=0.87gCm-2, year 8
NEE (gC/m2)
5
0
-5
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Month
post-DA NEE
true NEE
NEE obs

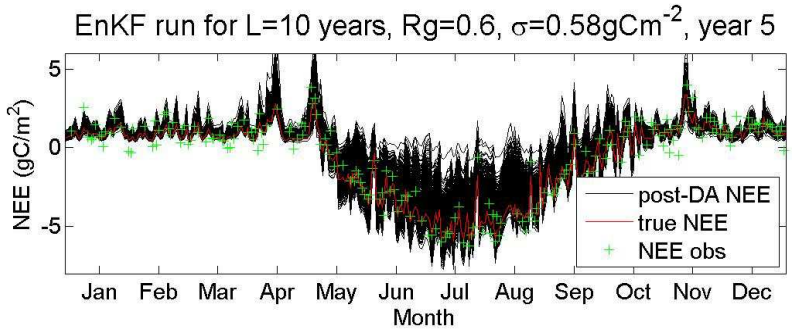
Plots of NEE trajectories for the EnKF

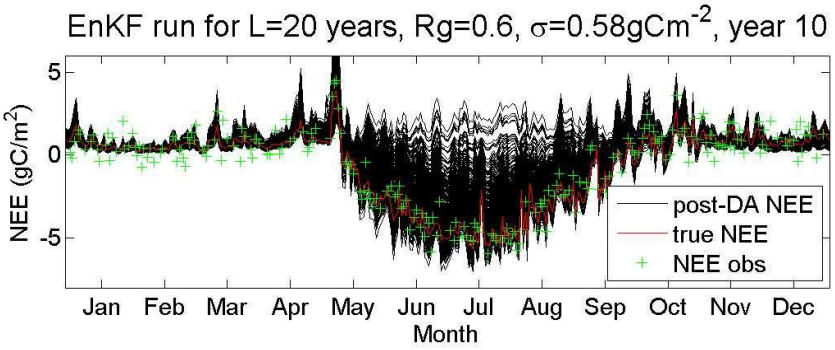
Below are displayed the plots of NEE trajectories for all 15 of the runs for the EnKF. For each run, DALEC-D is run forward using the parameter sets corresponding to the 500 ensembles. These are represented by the black lines. The red line shows the true NEE trajectory, and the green crosses are the observations. In each case, only one year of the trajectory is shown, for example for run 2, only the 2nd year is shown. This is because if the entire trajectory was displayed, much of the detail would be lost. For each run, the year that was chosen corresponds to a year representative of all the other years.

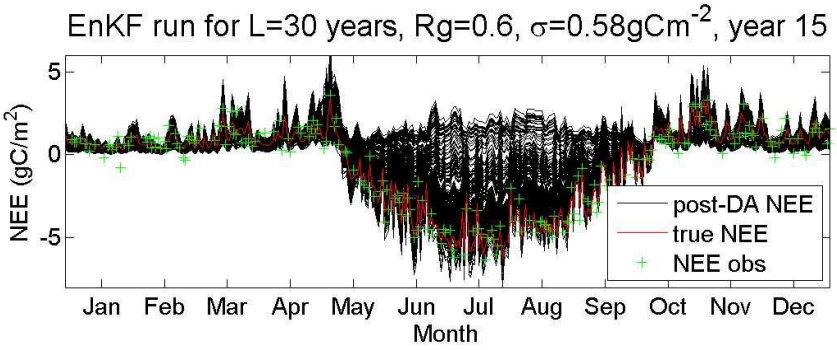


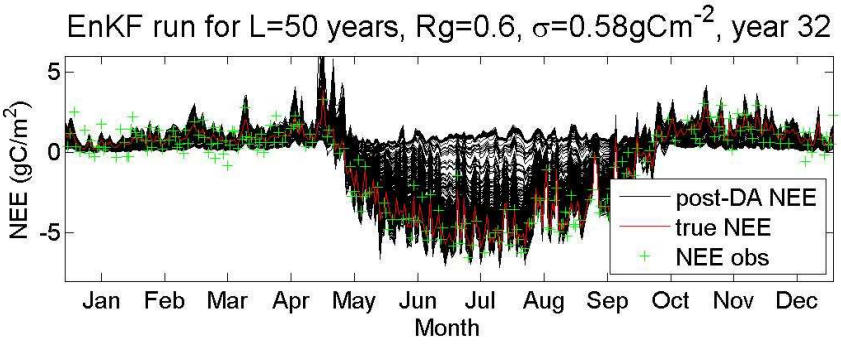


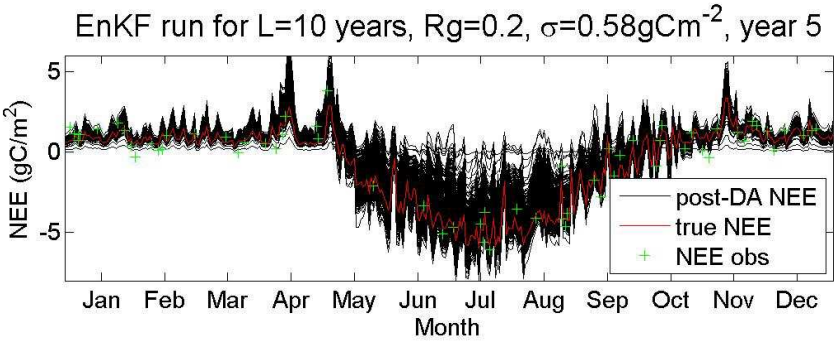


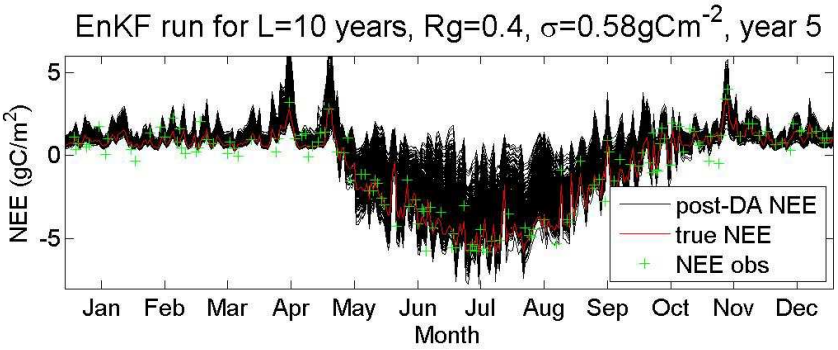


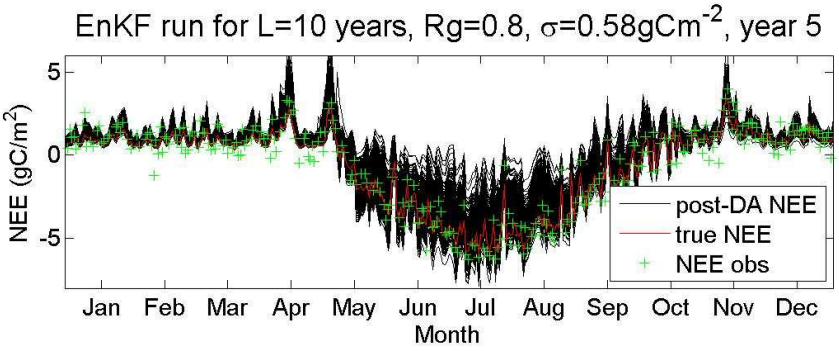

EnKF run for L=20 years, Rg=0.6, σ=0.58gCm-2, year 10
NEE (gC/m2)
5
0
-5
post-DA NEE
true NEE
NEE obs
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Month

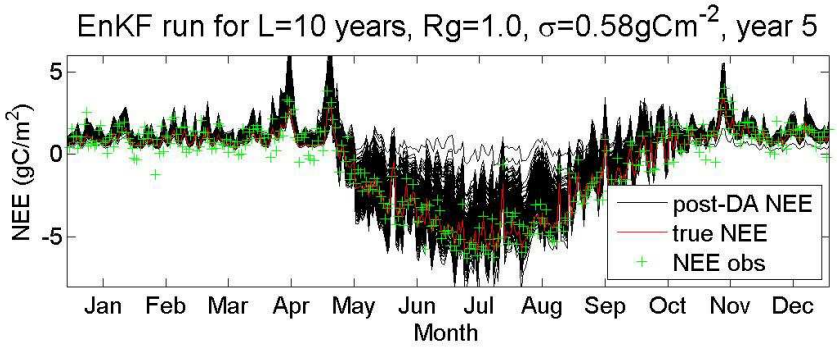

EnKF run for L=30 years, Rg=0.6, σ=0.58gCm-2, year 15
NEE (gC/m2)
5
0
-5
post-DA NEE
true NEE
NEE obs
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Month

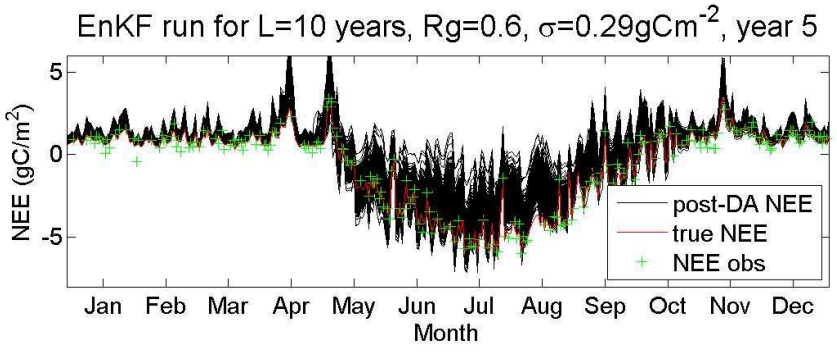

EnKF run for L=50 years, Rg=0.6, σ=0.58gCm-2, year 32
NEE (gC/m2)
5
0
-5
post-DA NEE
true NEE
NEE obs
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Month

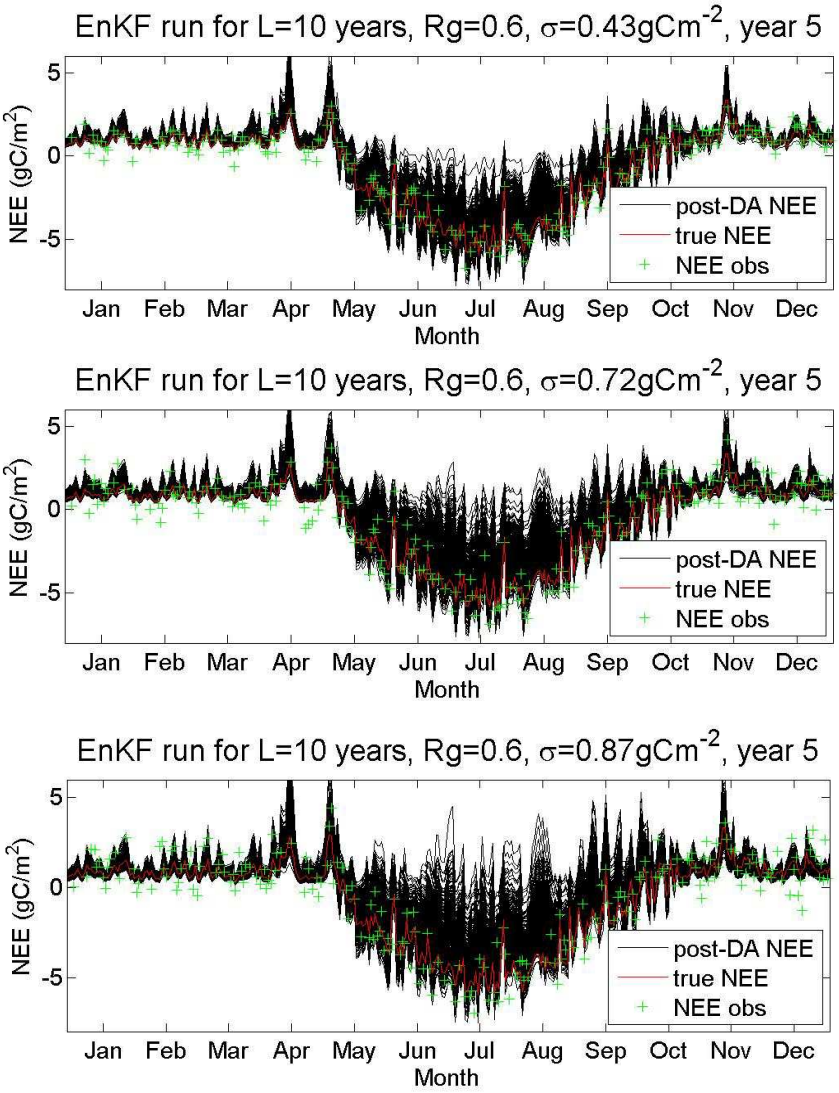

EnKF run for L=10 years, Rg=0.2, σ=0.58gCm-2, year 5
NEE (gC/m2)
5
0
-5
post-DA NEE
true NEE
NEE obs
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Month


EnKF run for L=10 years, Rg=0.4, σ=0.58gCm-2, year 5
NEE (gC/m2)
5
0
-5
post-DA NEE
true NEE
NEE obs
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Month


EnKF run for L=10 years, Rg=0.8, σ=0.58gCm-2, year 5
NEE (gC/m2)
5
0
-5
post-DA NEE
true NEE
NEE obs
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Month


EnKF run for L=10 years, Rg=1.0, σ=0.58gCm-2, year 5
NEE (gC/m2)
5
0
-5
post-DA NEE
true NEE
NEE obs
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Month


EnKF run for L=10 years, Rg=0.6, σ=0.29gCm-2, year 5
NEE (gC/m2)
5
0
-5
post-DA NEE
true NEE
NEE obs
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Month


EnKF run for L=10 years, Rg=0.6, σ=0.43gCm-2, year 5
NEE (gC/m2)
5
0
-5
post-DA NEE
true NEE
NEE obs
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Month
EnKF run for L=10 years, Rg=0.6, σ=0.72gCm-2, year 5
NEE (gC/m2)
5
0
-5
post-DA NEE
true NEE
NEE obs
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Month
EnKF run for L=10 years, Rg=0.6, σ=0.87gCm-2, year 5
NEE (gC/m2)
5
0
-5
post-DA NEE
true NEE
NEE obs
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Month

Appendix D

(A) Post-DA plots after assimilating LAI MODIS-orig dataset

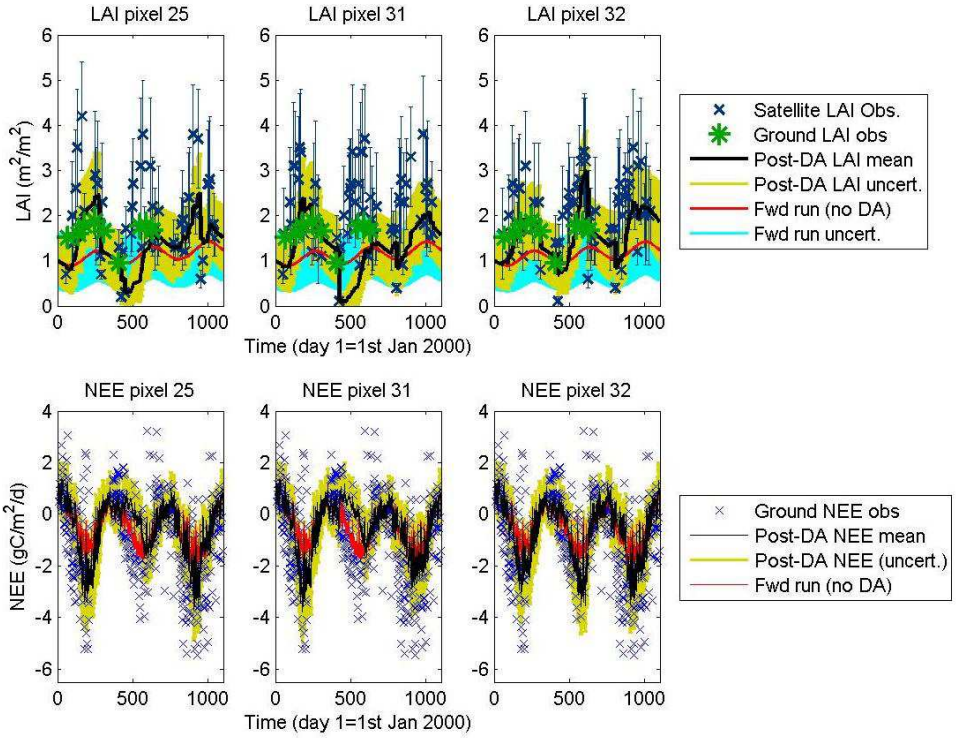


Figure D-A1 Plots (Loobos) of the post-assimilation LAI and NEE states (black lines) plotted alongside the MODIS LAI-orig observations / ground NEE observations (blue crosses) and the forward runs. The uncertainty in the post-DA LAI/NEE is represented by +/- 1 s.d.

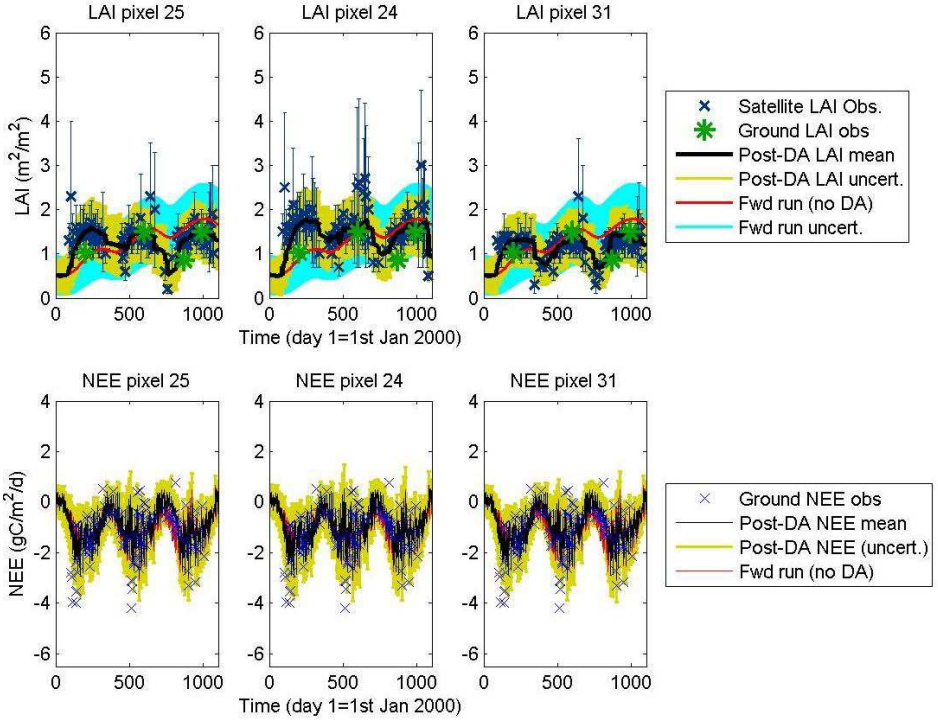


Figure D-A2 The same as figure D-A1 except the plots correspond to Oregon.

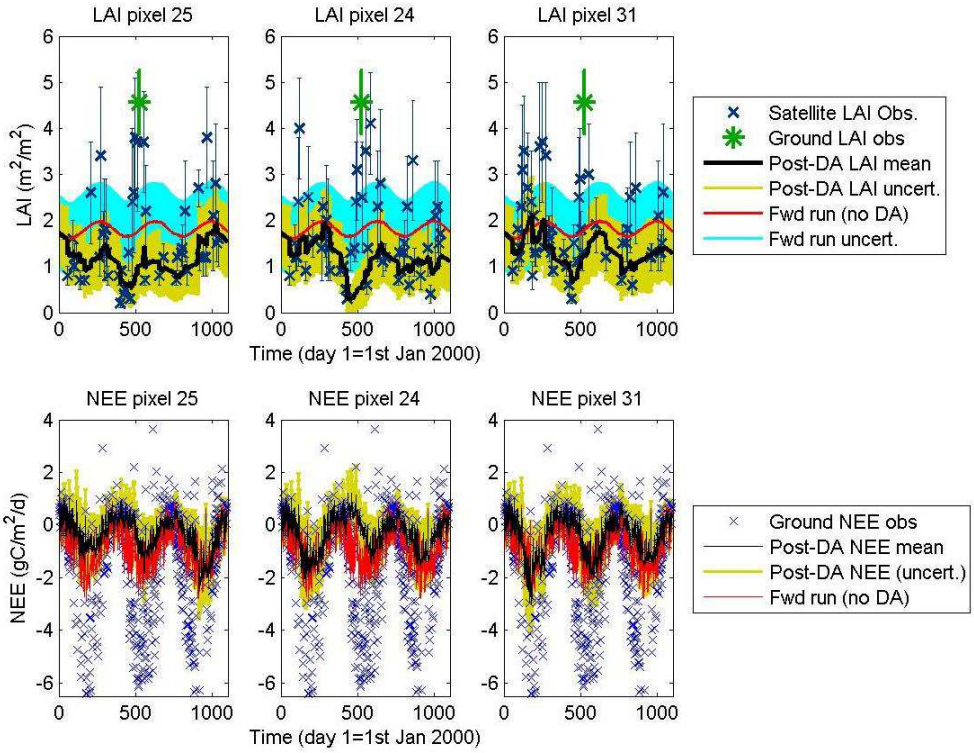


Figure D-A3 The same as figure D-A1 except the plots correspond to Tharandt.

(B) Post-DA plots after assimilating LAI MODIS-smsd dataset

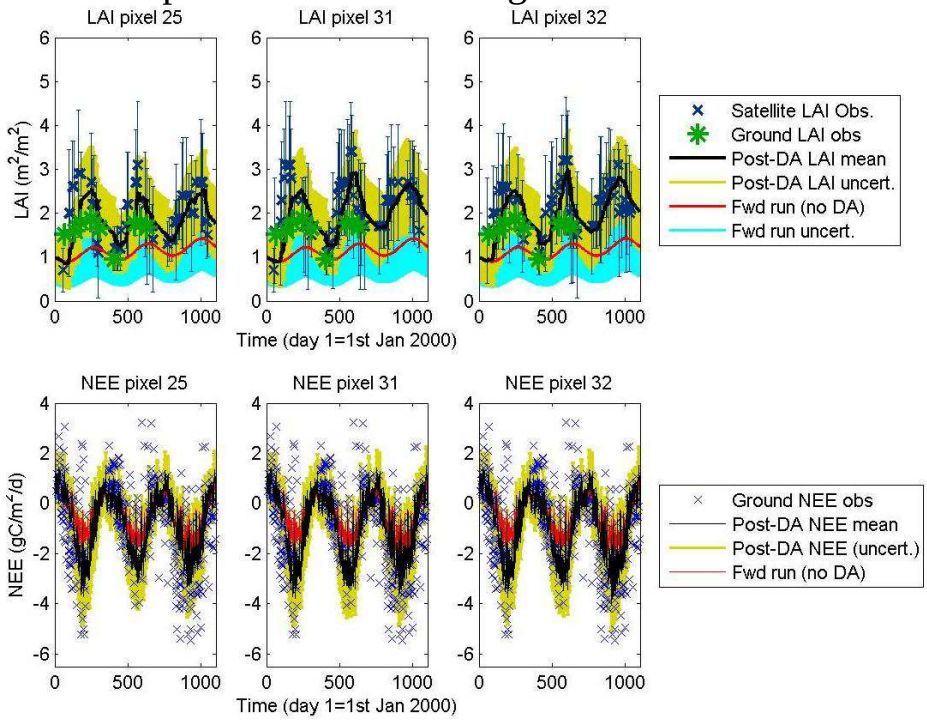


Figure D-B1 Plots (Loobos) of the post-assimilation LAI and NEE states (black lines) plotted alongside the MODIS LAI-smsd observations / ground NEE observations (blue crosses) and the forward runs. The uncertainty in the post-DA LAI/NEE is represented by +/- 1 s.d..

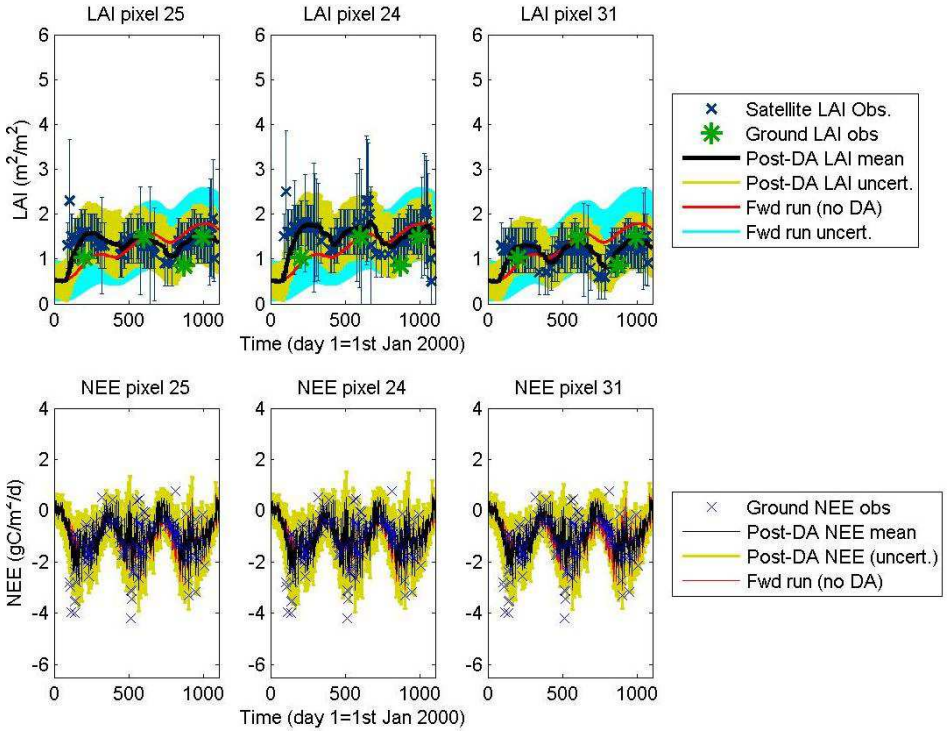


Figure D-B2 The same as figure D-B1 except the plots correspond to Oregon.

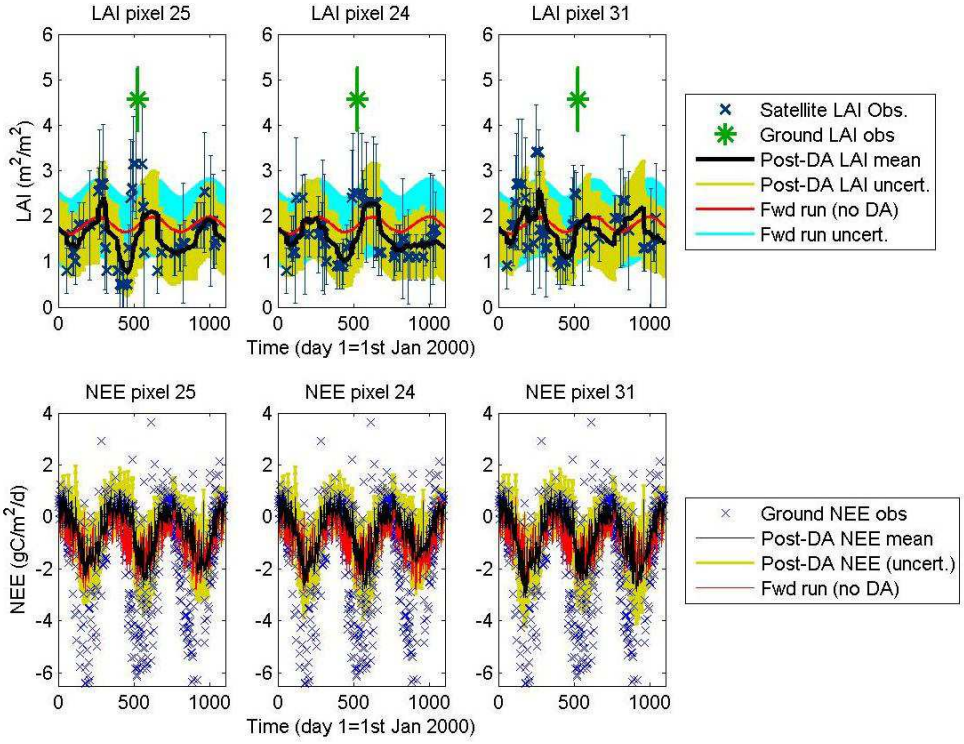


Figure D-B3 The same as figure D-B1 except the plots correspond to Tharandt.

(C) Post-DA plots after assimilating LAI MODIS-smsd dataset and the ground NEE dataset

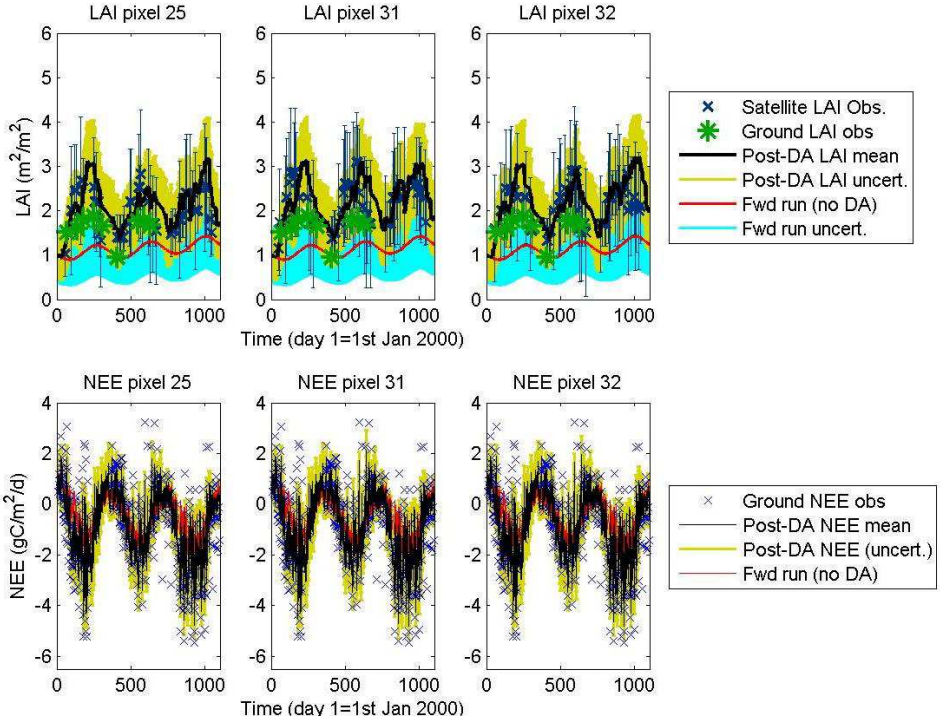


Figure D-C1 Plots (Loobos) of the post-assimilation LAI and NEE states (black lines) plotted alongside the MODIS LAI-smsd observations / ground NEE observations (blue crosses) and the forward runs. The uncertainty in the post-DA LAI/NEE is represented by +/- 1 s.d.

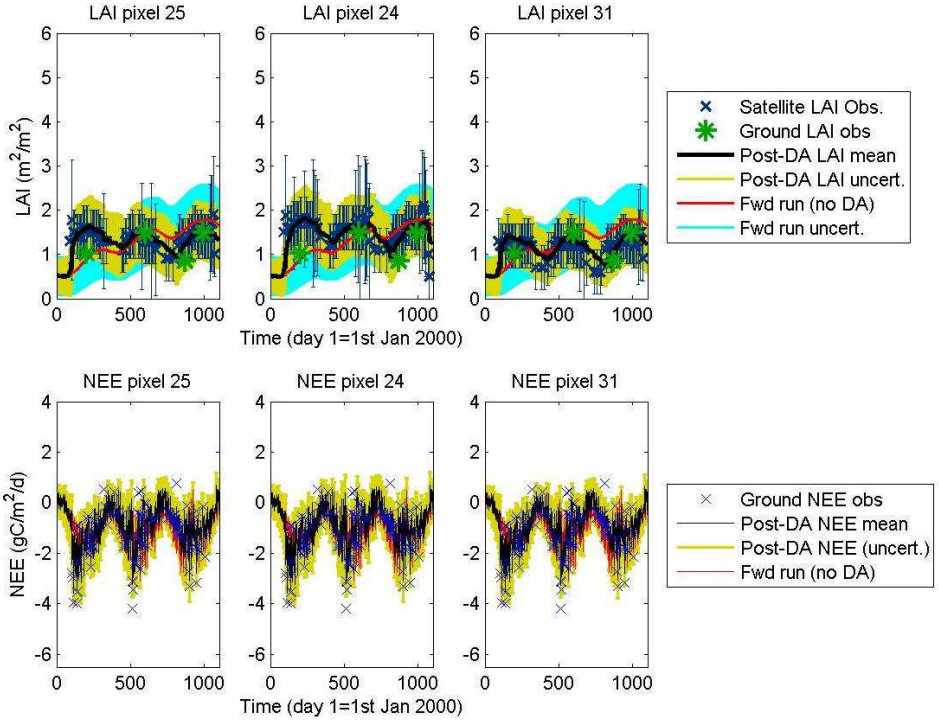


Figure D-C2 The same as figure D-C1 except the plots correspond to Oregon.

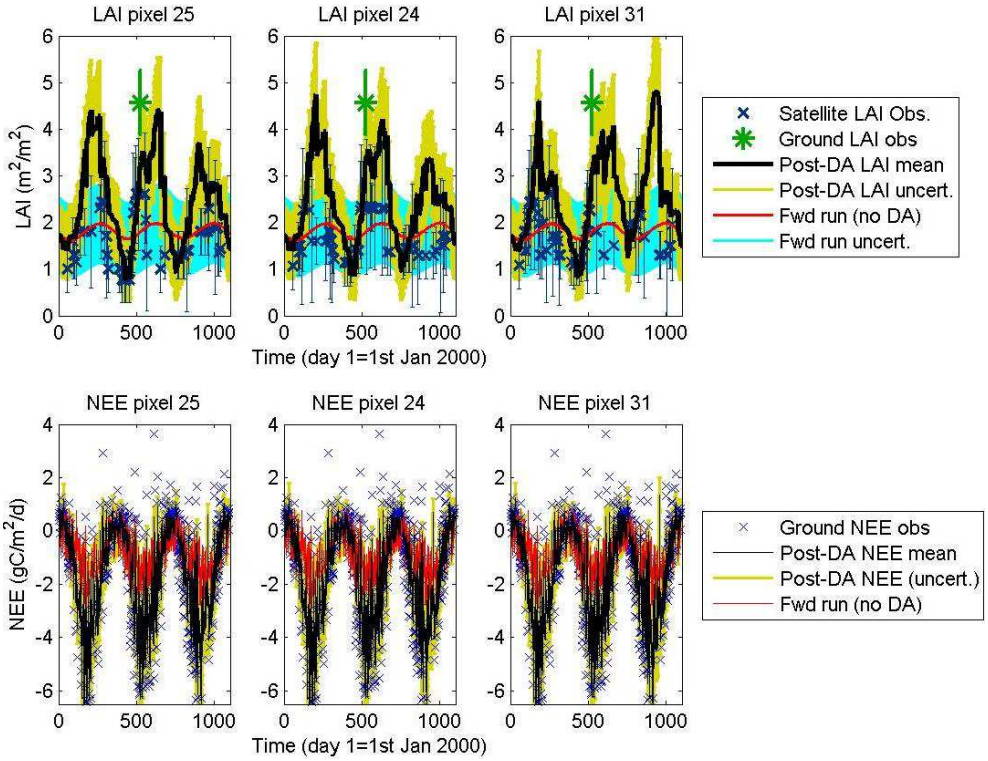


Figure D-C3 The same as figure D-C1 except the plots correspond to Tharandt.

References

Aber, J. & Federer, C. (1992) A generalized, lumped-parameter model of photosynthesis, evapotranspiration and net primary production in temperate and boreal forest ecosystems. *Oecologia*, **92,** 463-474

Abrahart, R. & See, L. (2007) Neural Network emulator of a rainfall runoff model. *Hydrology and Earth System Sciences Discussions*, **4**, 287–326

Abramowitz G. & Pitman A. (2007) Systematic bias in land surface models. *Journal of Hydrometeorology*, **8**, 989-1001.

Achard, F., Eva, H., Stibig, H., Mayaux, P., Gallego, J., Richards, T. & Malingreau, J. (2002). Determination of Deforestation Rates of the World's Humid Tropical Forests. *Science*, **297**: 999-1002

Ades, M. & van Leeuwen, P. (2012) An exploration of the equivalent weights particle filter. *Quarterly Journal of the Royal Meteorological Society*, **00**, 2–28.

Alton, P. (2010) Personal Communication at the 2010 conference of the National Centre of Earth Observation.

Amthor J. (1994) Scaling CO2-photosynthesis relationships from the leaf to the canopy. *Photosynthesis Research*, **39**, 321-350.

Atjay, G., Ketner, P. & Duvigneaud, P. (1979) Terrestrial primary production and phytomass. In: The Global Carbon Cycle, [Bolin, B., E.T. Degens, S. Kempe, and P. Ketner (eds.)]**,** John Wiley & Sons, Chichester, pp. 129-181.

Aubinet M., Berbigier P., Bernhofer C., Cescatti, A., Feigenwinter, C., Granier, A., Grunwald T., Havrankova, K., Heinesch, B., Longdoz, B., Marcolla, B., Montagnani, L. & Sedlak, P. (2005) Comparing CO_2 storage and advection conditions at night at different carboeuroflux sites. *Boundary-Layer Meteorology* **116**, 63-94.

Bachman, S., Heisler-White, J., Pendall, E., Williams, D., Morgan, J. & Newcomb, J. (2010) Elevated carbon dioxide alters impacts of precipitation pulses on ecosystem photosynthesis and respiration in a semi-arid grassland. *Oceanologia*, **162:** 791–802.

Bacour, C., Baret, F., Beal, D., Weiss, M. & Pavageau, K. (2006) Neural network estimation of LAI, fAPAR, fCover and LAIxCab, from top of canopy MERIS reflectance data: Principles and validation. *Remote Sensing of Environment*, **105(4)**, 313–325.

Baldocchi, D. (2012) Micrometeorological Flux Measurements, Eddy Covariance, Implementation. *ESPM 228, Advanced Topics in Biometeorology and Micrometeorology*. Berkeley University, Unpublished.

Baldocchi, D. (2008) Breathing of the terrestrial biosphere: lessons learned from a global network of carbon dioxide flux measurement systems. *Australian Journal of Botany*, **56**, 1-26.

Baldocchi, D. (2003) Assessing the eddy covariance technique for evaluating carbon dioxide exchange rates of ecosystems: past, present and future. *Global Change Biology*, **9**:479-492.

Baldocchi, D., Falge, E. & Gu, L. (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**, 2415-2434.

Balshi, M., McGuire, A., Duffy, P., Flannigan, M., Kicklighter, D. & Melillo, J. (2009) Vulnerability of carbon storage in North American boreal forests to wildfires during the 21st century. *Global Change Biology,* **15,** 1491-1510.

Bard, Y. (1974). *Nonlinear Parameter Estimation*. New York: Academic Press p.11.

Baret, F., Hagolle, O., Geiger, B., Bicheron, P., Miras, B., Huc, M., Berthelot, B., Iño, F., Weiss, M., Samain, O., Roujean, J., & Leroy, M. (2007), LAI, fAPAR and fCover CYCLOPES global products derived from VEGETATION: Part 1: Principles of the algorithm. *Remote Sensing of Environment*, **110**, 275– 286.

Baret, F. & Weiss, M. (2004) Can-Eye: processing digital photographs for canopy structure characterization. CAN EYE tutorial document, Avignon, France.

Barkley, M., Friess, U. & Monks, P. (2006), Measuring atmospheric CO2 from space using Full Spectral Initiation (FSI) WFM-DOAS. *Atmospheric Chemistry and Physics Discussions*, **6**, 2765–2807.

Bastos, L. & O'Hagan, A. (2008) Diagnostics for Gaussian process emulators. *Technometrics* **51**, 425-438

Bayes, T. (1763) An Essay Towards Solving a Problem in the Doctrine of Chances. *Philosophical Transactions*, **53**, 370-418.

Beer, C., Reichstein, M., Tomelleri, E., Ciais, P., Jung, M., Carvalhais, N., Rödenbeck, C., Arain, M., Baldocchi, D., Bonan, G., Bondeau, A., Cescatti, A., Lasslop, G., Lindroth, A., Lomas, M., Luyssaert, S., Margolis, H., Oleson, K., Roupsard, O., Veenendaal, E., Viovy, N., Williams, C., Woodward, I. & Papale, D. (2010). Terrestrial gross carbon dioxide uptake: global distribution and covariation with climate. *Science,* **329**, 834-838.

Behrenfeld, M., Siegel, D., O'Malley, R., & Maritorena, S. (2009) Global ocean phytoplankton, *American Meteorological Society*, **90**, S68–S73.

Berrut, J.P. & Trefethen, L. (2004) Barycentric Lagrange Interpolation. *SIAM REVIEW*. **46**, 3, 501–517

Blankenship, R. (2002) Molecular mechanism of photosynthesis. Blackwell Science Ltd. London.

Blyth, E., Gash, J., Lloyd, A., Pryor, M., Weedon, G. & Shuttleworth, J. (2010) Evaluating the JULES Land Surface Model Energy Fluxes Using FLUXNE Data. *Journal of Hydrometeorology.* **11 (2),** 509 – 519.

Boden, T., Marland, G. & Andres, R. (2010). Global, Regional, and National Fossil-Fuel CO_2 Emissions. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tenn., U.S.A.

Bond-Lamberty, B., C. Wang, & S. T. Gower (2004), Net primary production and net ecosystem production of a boreal black spruce fire chronosequence. *Global Change Biology*, **10**, 473–487.

Boose, E. (2001) Fisher Meteorological Station. *Harvard Forest Data Archive,* **HF001**.

Bouttier, F. & Courtier, P. (1999) Data assimilation concepts and methods. *Meteorological Training Course Lecture Series*. ECMWF. Unpublished.

Bowling, D., Burns, S., Conway, T., Monson, R. & White, J. (2005) Extensive observations of CO_2 carbon isotope content in and above a high-elevation subalpine forest. *Global Biogeochemical Cycles,* **19(3)**, GB3023.

Bowman, D., Balch, J., Artaxo, P., Bond, W., Carlson, J., Cochrane, M., D'Antonio, C., DeFries, R., Doyle, J., Harrison, S., Johnston, F., Keeley, J., Krawchuk, M., Kull, C., Marston, B., Moritz, M., Prentice, C., Roos, C., Scott, A., Swetnam, T., van der Werf, G. & Pyne, S. (2009). Fire in the Earth System. *Science,* **324**, 481–484.

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

Breda, N. (2003) Ground-based measurements of leaf area index: a review of methods, instruments and controversies. *Journal of Experimental Botany*. **54 (392)**, 2403–2417.

Breon, F., Ciais, P. (2010), Spaceborne remote sensing of greenhouse gas concentrations. *Geoscience*, **342**, 412–424.

Brooks S.P. & Gelman A. (1998) General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, **7(4):** 434-455.

Brovkin, V., Raddatz, T., Reick, C., Claussen, M. & Gayler, V. (2009) Global biogeophysical interactions between forest and climate. *Geophysical Research Letters.* **36**, L07405.

Burba G., Mcdermitt D., Grelle A., Anderson D. & Xu L. (2008) Addressing the influence of instrument surface heat exchange on the measurements of CO_2 flux from open-path gas analyzers. *Global Change Biology*, **14**, 1854-1876.

Burgess, E. (1837) General Remarks on the Temperature of the Terrestrial Globe and the Planetary Spaces; by Baron Fourier. *American Journal of Science*, Vol 32, pp. 1-20. Translation from the French, of Fourier, J. B. J., 1824, "Remarques Générales Sur Les Températures Du Globe Terrestre Et Des Espaces Planétaires.", *Annales de Chimie et de Physique*, Vol. 27, pp. 136–167.

Butler, N. (2001) Optimal and orthogonal Latin hypercube designs for computer experiments *Biometrika* **3,** 847-857.

Cable, J., Ogle, K., Lucas, R., Huxman, M., Smith, S., Tissue, D., Ewers, B., Pendall, E., Welker, J., Charlet, T., Cleary, M., Griffin, A., Nowak, R., Rogers, M., Steltzer, H., Sullivan, P. & van Gestel, N. (2009) The temperature responses of soil respiration in deserts: a seven desert synthesis. *Biogeochemistry*, **103:** 71-90

Canadell, J., Le Quere, C., Raupach, M., Field, C., Buitenhuis, E., Ciais, P., Conway, T., Gillet, N., Houghton, R. & Marlan, G. (2007) Contributions to accelerating atmospheric CO_2 growth from economic activity, carbon intensity, and efficiency of natural sinks. *PNAS*, **104,** 18866-18870.

Carter, A. & Scholes, R. (2000) Spatial Global Database of Soil Properties. IGBP Global Soil Data Task CD-ROM. International Geosphere-Biosphere Programme (IGBP) Data Information Systems. Toulouse, France.

Carvalhais, Nm, Reichstein, M. & Seixas J. (2008) Implications of the carbon cycle steady state assumption for biogeochemical modeling performance and inverse parameter retrieval. *Global Biogeochemical Cycles*, **22**, Gb2007.

Černý, V. (1985). Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of Optimization Theory and Applications* **45,** 41–51.

Chen, M., Liu, S., Tieszen, L. & Hollinger, D. (2008) An improved state-parameter analysis of ecosystem models using data assimilation. *Ecological Modelling*, **219,** 317–326.

Chokkalingam, U. (2011) Global warming and the dual role of forests. *Science Review. Forest Asia Brief No. 1.* Available at http://www.forestcarbonasia.org/wp-content/uploads/2011/04/ Forest-Carbon-Asia-Brief1-Science.pdf [Accessed 12th April 2012].

Chuter, A. (2013) A Qualitative Analysis of the Data Assimilation Linked Ecosystem Carbon Model, DALEC. *Ph.D. Thesis*. University of Surrey: U.K.

Clark, D., Mercado, L., Sitch, S., Jones, C., Gedney, N., Best, M. , Pryor, M., Rooney, G., Essery, R., Blyth, E., Boucher, O., Harding, R., Huntingford, C. & Cox, P. (2011) The Joint UK Land Environment Simulator (JULES), model description – Part 2: Carbon fluxes and vegetation dynamics. *Geoscientific Model Development*, **4,** 701-722.

Clark, D., Olivas, P., Oberbauer, S., Clark, D. & Ryan, M. (2009) First direct landscape-scale measurements of tropical rain 1 Leaf Area Index, a key driver of global primary production. **11(2),** 163-172.

Conti, S. & O'Hagan, A. (2010). Bayesian Emulation of Complex Multi-Output and Dynamic Computer Models. *Journal of Statistical Planning and Inference*, **140,** 640-651.

Couutier, P. & Thepaut, A. (1994) A strategy for operational implementation of 4D-Var, using an incremental approach. *Quarterly Journal of the Royal Meteorological Society*, **120 (519),** 1367–1387.

Cox, P. (2001) Description of the "TRIFFID" Dynamic Global Vegetation Model. Hadley Centre. UK.

Cressie, N. (1990) The Origins of Kriging. *Mathematical Geology*. **22(3)**, 239-252

Crow, W., & Ryu, D. "A new data assimilation approach for improving runoff prediction using remotely-sensed soil moisture retrievals," *Hydrologic and Earth System Sciences*, **13** ,1-16, 2009.

De Fries**,** R., Field, C., Fung, I., Collatz, G. & Bounoua, L. (1999) Combining satellite data and biogeochemical models to estimate global effects of human-induced land cover change on carbon emissions and primary productivity. *Global Biogeochemical Cycles*, **13,** 803-815.

De Kauwe, M., Disney, M. Quaife, T. Lewis, P. & Williams, M. (2011). An assessment of the MODIS collection 5 leaf area index product for a region of mixed coniferous forest. *Remote Sensing of Environment*, **115**, 767-780.

De Kauwe, M., Quaife, T., Lewis, P., Disney, M. & Williams, M. (2008) Estimating the Spatial Exchange of Carbon through the Assimilation of Earth Observation Derived Products using an Ensemble Kalman Filter. *Geoscience and Remote Sensing Symposium, 2008. IGARSS 2008. IEEE International* , vol.3, no., pp.III-1044-III-1047, 7-11 July 2008

Dekker, S., Bouten, W. & Schaap, M. (2001) Analysing forest transpiration model errors with artificial neural networks. *Journal of Hydrology*, **246**, 197-208.

Demarty, J., Chevallier, F., Friend, A., Viovy, N., Piao, S. & Ciais, P. (2007) Assimilation of global MODIS leaf area index retrievals within a terrestrial biosphere model. *Geophysical Research Letters*, **34**, L15402.

Denman, K., Brasseur, G., Chidthaisong, A., Ciais, P., Cox, P., Dickinson, R., Hauglustaine, D., Heinze, C., Holland, E. Jacob, D., Lohmann, U., Ramachandran, S., da Silva Dias, P., Wofsy, S. & Zhang, X. (2007) Couplings Between Changes in the Climate System and Biogeochemistry. In: *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* [Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M.Tignor and H.L. Miller (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Dente, L., Satalino, G., Mattia, F., & Rinaldi, M. (2008). Assimilation of leaf area index derived from ASAR and MERIS data into CERES-Wheat model to map wheat yield. *Remote Sensing of Environment*, **112**, 1395-1407.

Disney, M., & Quegan, S. (2007) EO data assimilation in land process models, PowerPoint presentation, ESA GlobModel Workshop, 12-13th September 2007 [Available at: http://www.globmodel.info/]

Draper, C., Reichle, R., De Lannoy, G. & Liu, Q. (2011) Assimilation of passive and active microwave soil moisture retrievals" *Geophysical Research Letters*, **39**, L04401.

Entekhabi, D., Njoku, E. G., O'Neill, P. E., Kellogg, K. H., Crow, W. T., Edelstein, W. & Van Zyl, J. (2010). The soil moisture active passive (SMAP) mission. *Proceedings of the IEEE*, **98(5)**, 704-716.

Etheridge, L., Steele, P., Langenfelds, R., Francey, R., Barnola, J. & Morgan, V. (1998) Historical CO_2 records from the Law Dome DE08, DE08-2, and DSS ice cores. In Trends: A Compendium of Data on Global Change. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tenn., U.S.A.

Evans, G. (2003) Defining misfit between biogeochemical models and data sets. *Journal of Marine Systems*, **40**, 49-54.

Evensen, G (2009) The Ensemble Kalman Filter for Combined State and Parameter Estimation. Monte Carlo Techniques for Data Assimilation in Large Systems. *Ieee Control Systems Magazine*, **29**, 83-104.

Evensen, G. (2003) The Ensemble Kalman Filter: theoretical formulation and practical implementation. *Ocean Dynamics*, **53**, 343-367.

Evensen, G. & van Leeuwen, P. (2000) An Ensemble Kalman Smoother for Nonlinear Dynamics. *Monthly Weather Review*. **128**, 1852-1867,

Evensen, G., Dee, D., Schroter, J. (1998) Parameter estimation in dynamical models. In: Chassignet EP, Verron J (eds.) Ocean modeling and parameterizations Kluwer Academic, The Nederlands, pp 373–398

Evensen, G. (1994) Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research*, **99**, 10143–10162.

Evensen, G. (1992) Using the Extended Kalman Filter with a Multilayer Quasi-Geostrophic Ocean Model. *Journal of Geophysical Research*, **97,** 17,905-17,924.

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

Fang, C. & Moncrieff, J. (1999) A model for soil CO_2 production and transport: 1. Model development. *Agricultural and Forest Meteorology*, **95**, 225–236.

Fang, W., Wei, S. & Liang, S. (2012) Validation of MODIS and CYCLOPES LAI products using global field measurement data. *Remote Sensing of Environment*, **119,** 43–54.

Farquhar G . & Von Caemmerer S . (1982) Modelling of photosynthetic response to the environment. In: Physiological Plant Ecology II. Encyclopedia of Plant Physiology, New Series, Vol . 12B (eds O . L . Lange, P . S . Nobel, C . B . Osmond & H . Ziegler), pp . 549-587 . Springer-Verlag, Berlin.

Feller, W. (1966) An Introduction to Probability Theory and its Applications, 2nd ed. New York: John Wiley & Sons, Inc.

Feyen, L., Vrugt, J., Breanndan, N., der Knijff, J. & De Roo, A. (2007). Parameter Optimisation and uncertainty assessment for large-scale streamflow simulation with the LISFLOOD model. *Journal of Hydrology*, **332,** 276-289.

Field, C., Berry, J. & Mooney, H. (1982) A portable system for measuring carbon dioxide and water vapour exchange of leaves. *Plant, Cell and Environment*, 5, 179-186.

Fox, A, Williams, M., Richardson, A., Cameron, D., Gove, J., Quaife, T., Ricciuto, D., Reichstein, M., Tomelleri, E., Trudinger, C., van Wijk, M. (2009) The REFLEX project: comparing different algorithms and implementations for the inversion of a terrestrial ecosystem model against eddy covariance data. *Agricultural and Forest Meteorology* **149**, 1597–1615.

Frankenberg, C., Fisher, J., Worden, J., Badgley, G., Satchi, S., Lee, J., Toon, G., Butz, A., Jung, M., Kuze, A., Yokota, T. (2011*),* New global observations of the terrestrial carbon cycle from GOSAT: Patterns of plant fluorescence with gross primary productivity, *Geophysical Research Letters*, **38,** L17706

Fricker, T., Oakley, J. & Urban, N. (2010) Multivariate Emulators with Nonseparable Covariance Structures. MUCM Technical Report. Sheffield, U.K..

Friend, A., Arneth, A., Kiang, N., Lomas, M., Ogee, J., Rodenbeck., C., Running, S., Santaren, J., Sitch, S., Viovy, N., Woodward, F. I. & Zaele, S. (2007) FLUXNET and modelling the global carbon cycle. *Global Change Biology*, **13**, 610-633.

Gallagher, K. & Sambridge, M. (1994) Genetic Algorithms: A powerful tool for large-scale non-linear optimization problems. *Computers and Geoscience*. **20**, 1229-1236.

Gao, C., Wang, H., Weng, E., Lakshmivarahan, S., Zhang, Y. & Luo, Y. (2011) Assimilation of multiple data sets with the ensemble Kalman filter to improve forecasts of forest carbon dynamics. *Ecological Applications.* **21(5),** 1461 – 1473.

Gao, F., Morisette, J., Wolfe, R., Ederer, G., Pedelty, Masuoka, E., Myneni, R., Tan, B. & Nightingale, J. (2008) An Algorithm to produce temporally and spatially continuous MODIS-LAI time series. *IEEE Geoscience and Remote Sensing Letters* **5(1)**, 60-64.

Garrigues, S., R. Lacaze, F. Baret, J. T. Morisette, M. Weiss, J. E. Nickeson, R. Fernandes, S. Plummer, N. V. Shabanov, R. B. Myneni, Y. Knyazikhin, & W. Yang (2008) Validation and intercomparison of global Leaf Area Index products derived from remote sensing data. *Journal of Geophysical Research,* **113**, G02028

Gelman, A. (1995), *Bayesian data analysis*, xix, 526p. pp., Chapman & Hall, London.
Gemen, S. & Gemen, D. (1984) Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-6(6)**, 721-741.
Ghent, D., Kaduk, J., Remedios, J., Ardo, J. & Balzter, H. (2010) Assimilation of land-surface temperature into the land surface model JULES with an Ensemble Kalman Filter. *Journal of Geophysical Research*, **115**, D19112
Gilks, W. & Best, N. (1995) Adaptive Rejection Metropolis Sampling within Gibbs sasmpling. *Applications of Statistics*, **44(4)**, 455-472.
Global Carbon Project (2011) Carbon budget and trends 2010. http://www.globalcarbonproject.org/carbonbudget.
Gomez, J. (2011) *Inference on flux measurements using an ecosystem model*. London: UCL. Unpublished
Gosling, J. (2006) Differences between estimates of expected computer code output with GEM-SA. MUCM Technical Report 07/03. Department of Probability and Statistics, The University of Sheffield.
Goulden, M., Munger, J., Fan, S., Daube, B. & Wofsy, S. (1996) Measurements of carbon sequestration by long-term eddy covariance: Methods and a critical evaluation of accuracy. *Global Change Biology*, **2**, 169-182.
Gower, S.T., Kucharik, C. & Norman, J. (1999) Direct and Indirect Estimation of Leaf Area Index, *f*APAR, and Net Primary Production of Terrestrial Ecosystems. *Remote Sensing of Environment*, **70**, 29–51.
Grace, J. (2005) Estimating Biogenic Carbon Fluxes from Flux tower measurements and Earth Observation data' in 'UK Emissions by Sources and Removals by Sinks due to Land Use, Land Use Change and Forestry Activities, Report, June 2005. *Centre for Ecology & Hydrology, Aberdeen University, Rothamsted Research, Forest Research Alice Holt, Centre for Terrestrial Carbon Dynamics (University of Edinburgh, University College London, University of Sheffield).*
Gruber, N, Gloor, M., Mikaloff, S., Doney, S., Dutkiewicz, S., Follows, M., Gerber, M., Jacobson, A., Joos, F., Lindsay, K., Menemenlis, D., Mouchet, A., Muller, S., Sarmiento, J. & Takahashi, T. (2009) Oceanic sources, sinks and transport of atmospheric CO_2. *Global Biogeochemical cycles*, **23**, GB1005.
Gu, Y., Belair, S., Mahfouf, J. & Deblonde, G. (2006) Optimal interpolation analysis of leaf area index using MODIS data. *Remote sensing of Environment*. **104,** 283-296.
Guanter, L., Frankenberg, C., Dudhia, A., Lewis, P.E., Gomez-Dans, J., Kuze, A. , Suto , H., Grainger, R. (2012) Retrieval and global assessment of terrestrial chlorophyll fluorescence from GOSAT space measurements, *Remote Sensing of Environment*, **121**, 236-251
Hankin, R. (2012) Introducing multivator: A Multivariate Emulator. *Journal of Statistical Software*, **46(8)**, 1-20.
Harmon, M., Bond-Lamberty, B., Tang, J. & Vargas, R. (2011) Heterotrophic respiration in disturbed forests: A review with examples from North America. *Journal of Geophysical research*. **116**, G00K04.

Hastings, G., K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their applications. *Biometrika*, **57(1)**, 97-109.
Haupt, S. & Haupt, R. (2004) Genetic Algorithms and their Applications in Environmental Sciences, Proceedings of the Third Conference on Artificial Intelligence Applications to Environmental Sciences, American Meteorological Society, Long Beach, CA, 1.1A.
Haylock, R. & O'Hagan, A. (1996). On inference for outputs of computationally expensive algorithms with uncertainty on the inputs. In: Bayesian Statistics 5, J. M. Bernardo et al (eds.). Oxford University Press, 629.637.
Henrici, P. (1982) Essentials of Numerical Analysis. Wiley, New York.
Hill, T., E. Ryan & M. Williams (2012) The use of CO2 flux time series for parameter and carbon stock estimation in carbon cycle research. Global Change Biology **18**, 179-193.
Hill, T., Quaife, T. & Williams, M. (2011) A data assimilation method for using low-resolution Earth Observation data in heterogeneous ecosystems. *Journal of Geophysical Research* **116**, D08117.
Hill, T., Williams, M., Woodward, F., & Moncrieff, J. (2011) Constraining ecosystem processes from tower fluxes and atmospheric profiles. *Ecological Applications,* **21,** 1474-1489.
Hill, T. & Williams, M. (2009). A simple inverse modelling approach applied to a coupled terrestrial ecosystem / planetary boundary layer model. *Chapter two of 'A modelling approach to carbon, water and energy feedbacks and interactions across the land-atmosphere interface', by Timothy Hill, PhD thesis, supervised by Mathew Williams*. Edinburgh University: U.K. Available at: http://www.geos.ed.ac.uk/homes/thill/PhD_thesis_MASTER_002.006.pdf
Holben, B. (1986). Characteristics of maximum-value composite images from temporal AVHRR data. *International Journal of Remote Sensing*, **7**, 1417−1434.
Hollinger, D. & Richardson, A. (2005) Uncertainty in eddy covariance measurements and its application to physiological models. *Tree Physiology*, **25**, 873-885.
Houtekamer, P. & Mitchell, H. (2005) Ensemble Kalman filtering. *Quarterly Journal of Review of the Meteorological Society*, **131**, 3269–3289.
Husslage, B., Rennen, G., Van Dam, E. & Den Hertog, D. (2006). Space-Filling Latin Hypercube Designs for Computer Experiments. Available at SSRN: http://ssrn.com/abstract=895464 or http://dx.doi.org/10.2139/ssrn.895464
Hyer, E. & Goetz, S. (2004) Comparison and sensitivity analysis of instruments and radiometric methods for LAI estimation: assessments from a boreal forest site. *Agricultural and Forest Meteorology*, **122**, 157–174
IPCC (2007) Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change [Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M.Tignor and H.L. Miller (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.
Jarlan, L., Balsamo, G., Lafont, S., Beljaars, A., Calvet, J. & Mougin, E. (2008) Analysis of Leaf Area Index in the ECMWF land surface scheme and impact on latent heat and carbon fluxes: Applications to West Africa, *Journal of Geophysical Research*, **113**, D24117.

Jones, H. (1992). Plants and microclimate. Cambridge University Press, Cambridge, England.

Jonsson, P. & Eklundh, L. (2004) TIMESAT – a program for analyzing time-series of satellite sensor data. *Computers & Geosciences*, **30(8)**, 833-845

Jonsson, P. & Eklundh, L. (2004). TIMESAT - a program for analyzing time-series of satellite sensor data. *Computers & Geosciences*, **30**, 833-845.

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

Kaminski, T., Knorr, W., Scholze, M., Gobron, N., Pinty, B., Giering, R. & Mathieu, P. (2012) Consistent assimilation of MERIS FAPAR and atmospheric CO_2 into a terrestrial vegetation model and interactive mission benefit analysis, *Biogeosciences Discussions*, **8**, 10761-10795.

Kattge, J., Diaz, S., Lavorel, S., Prentice, C., Leadley, P., Bonisch, G., Garnier, E., Westoby, M., Reich, P., Wrights, I. & Cornelissen, H. (2011) TRY – a global database of plant traits. *Global Change Biology*, **17(9)**, 2905-2935.

Keeling, C. (1960) The Concentration and Isotopic Abundances of Carbon Dioxide in the Atmosphere. *Tellus* **12**: 200–20.

Keeling, R., Piper, S. Bollenbacher, A. & Walker, J. (2008). Atmospheric CO_2 records from sites in the SIO air sampling network. In Trends: A Compendium of Data on Global Change. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tenn., U.S.A

Keeling, R. (2005) Comment on "The Ocean Sink for Anthropogenic CO_2". *Science*. **308**, 1743.

Kennedy, M., Anderson, C., O'Hagan, A., Lomas, M., Woodward, I, Gosling, J & Heinemeyer, A. (2008) Quantifying Uncertainty in the BiosphericCarbon Flux for England and Wales. *Journal of the Royal Statistical Society Annals*, **171,** 109–135.

Kennedy, M. C., Anderson, C. W., Conti, S. & O'Hagan, A. (2006). Case studies in Gaussian process modelling of computer codes. *Reliability Engineering and System Safety*, **91**, 1301-1309.

Knyazikhin, Y., Martonchik, J., Myeni, R., Diner, D. & Running, S. (1998) Synergistic algorithm for estimating vegetation canopy leaf area index and fraction of absorbed photosynthetically active radiation from MODIS and MISR data. *Journal of Geophysical Research*, **103**, 32257–32276.

Knorr, W., Kaminski, T. Scholze, M. Gobron, N. Pinty, B. Giering, R. & Mathieu, P. (2010), Carbon cycle data assimilation with a generic phenology model, *Journal of Geophysical Research.* **115**, G04017.

Knorr, W. & Kattge, J. (2005) Inversion of terrestrial ecosystem model parameter values against eddy covariance measurements by Monte Carlo sampling. *Global Change Biology*, **11**, 1333-1351.

Knorr, W. (2000) Annual and interannual CO2 exchanges of the terrestrial biosphere: Process-based simulations and uncertainties, *Global Ecology and Biogeography*. **9**, 225-252.

Kucharik, C. & Twine, T. (2007) Residue, respiration, and residuals: Evaluation of a dynamic agroecosystem model using eddy flux measurements and biometric data. *Agricultural Forest Meteorology*, **146**, 134–158.

Kumar, S., Reichle, R., Peters-Lidard, C., Koster, R., Zhan, X., Crow, W., Eylander, J. & Houser, P. (2008) A land surface data assimilation framework using the land information system: Description and applications. *Advances in Water Resources*, **31(11)**, 1419-1432.

Kurz, W., Dymond, C., Stinson, G., Rampley, G., Neilson, E. Carroll, A., Ebata, T. & Safranyik, L. (2008) Mountain pine beetle and forest carbon feedback to climate change, *Nature*, **452**, 987–990.

Kutsch, W., Kolle, O., Rebmann, C., Knohl, A., Ziegler, W. & Schulze, E. (2008) Advection and resulting CO2 exchange uncertainty in a tall forest in central Germany. *Ecological Applications*, **18**, 1391-1405.

Larcher, W. (2003) Physiological Plant Ecology. Springer-Verlay, Stuttgart.

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

Law, B. & Waring, R. (1994) Remote sensing of Leaf Area Index and Radiation intercepted by understory vegetation. *Ecological Applications*, **4**, 272-279.

Lawless, A. (2001) Development of Linear Models for Data Assimilation in Numerical Weather Prediction. *Unpublished Thesis (PhD)*. Met Office: U.K.

Le Quéré, C., Raupach, M., Canadell J. & Marland, G. (2009). Trends in the sources and sinks of carbon dioxide. *Nature Geosciences*, **2**, 831-836.

Le Quéré, C., Andres, R. J., Boden, T., Conway, T., Houghton, R. A., House, J. I., Marland, G., Peters, G. P., van der Werf, G., Ahlström, A., Andrew, R. M., Bopp, L., Canadell, J. G., Ciais, P., Doney, S. C., Enright, C., Friedlingstein, P., Huntingford, C., Jain, A. K., Jourdain, C., Kato, E., Keeling, R. F., Klein Goldewijk, K., Levis, S., Levy, P., Lomas, M., Poulter, B., Raupach, M. R., Schwinger, J., Sitch, S., Stocker, B. D., Viovy, N., Zaehle, S., and Zeng, N. (2012) The global carbon budget 1959–2011. *Earth System Science Data Discussion*, **5,** 1107-1157

Leuning, R., Zhang, Y., Rajaud, A., Cleugh, H. & Tu, K. (2008) A simple surface conductance model to estimate regional evaporation using MODIS leaf area index and the Penman-Monteith equation. *Water resources research*. **44:** W10419.

Lewis, P. E., Gomez-Dans, J., Kaminski, T., Settle, J., Quaife, T., Gobron, N., Styles, J. & Berger, M. (2012). An Earth Observation Land Data Assimilation System (EO-LDAS). *Remote Sensing of Environment*, **120**:219-235.

Lewis, P. E. (1999) Three-dimensional plant modelling for remote sensing simulation studies using the Botanical Plant Modelling System. *Agronomie*, **19,** 185–210.

Leakey, A., Ainsworth, E., Bernacchi, C., Rogers, A., Long, S. & Ort, D. (2009) Elevated CO_2 effects on plant carbon, nitrogen and water relations: six important lessons from FACE. *Journal of Experimental Botany*. **60**, 2859-2876.

Li, H., Zhang, Y., Chiew, F. & Xu, S. (2009) Predicting runoff in ungauged catchments by using Xinanjiang model with MODIS leaf area index. *Journal of Hydrology*. **370(1-4):** 155-162.

Limer, L., Anderson, C., O'Hagan, A. & Quegan, S. (2007) Uncertainty in the modelling of soil moisture in SDGVM. *Unpublished Thesis (PhD).* Sheffield, U.K..

Lloyd, J. & Taylor, J. (1994) On the temperature dependence of soil respiration. *Functional Ecology*. **8(3),** 315 – 323.

Lorenz, E. (1963) Deterministic nonperiodic flow. *Journal of Atmospheric Science*, **20**, 130–141.

Luke, C. (2011) Modelling aspects of land-atmosphere interaction: thermal instability in peatland soils and land parameter estimation through data assimilation. *Unpublished Thesis (PhD).* University of Exeter, U.K..

Luo, Y., Ogle, K., Tucker, C., Fei, S., Gao, C., LaDeau, S., Clark, J. & Schimel, D. (2011) Ecological forecasting and data assimilation in a data-rich era. *Ecological Applications,* **21(5)**, 1429-1442.

MacBean, N. Disney, M. I.**,** Lewis, P. & Ineson, P. (2010) Using satellite observations to improve model estimates of CO2 and CH4 flux: a Metropolis Hastings Markov Chain Monte Carlo approach, EGU meeting, Vienna, Austria, 3-7 May 2010.

Macfarlane, C., Arndt, S., Livesley, S., Edgar, A., White, D., Adams, M. & Eamus, D. (2007) Estimation of leaf area index in eucalypt forest with vertical foliage, using cover and fullframe fisheye photography. *Forest Ecology Management*, **242**, 756-763.

Mackay, D., Ewers, B., Loranty, M., Kruger, E. & Samata, S. (2012) Bayesian analysis of canopy transpiration models: a test of posterior parameter means against measurements. *Journal of Hydrology*. **432–433**, 75–83.

Mackenzie, D. (2003) Ensemble Kalman Filters bring Weather Models up to date. *Siam news,* **36(8)**, 1-4.

Mahalanobis, P. (1930) On tests and measures of group divergence, *Journal of the Asiatic Society of Bengal*, **26,** 541–588.

Mandel, J. (2007) – A brief tutorial of the Ensemble Kalman Filter [report]. *University of Colorado. U.S.A.*.

McKay, M., Conover, W. & Beckman, R. (1979). A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. *Technometrics*, **21(2)**, 239-245.

Mecklenburg S., Kerr, Y. Font, J. & Hahne, A. (2008) The Soil Moisture and Ocean Salinity (SMOS) Mission – An overview. *Geophysical Research Abstracts, Vol. 10.*

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

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. & Teller, E. (1953) Equation of state calculations by fast computing machines, *Journal of Chemical Physics*, **1(6)**, 1087–1092.

Meziane, D. & Shipley, B. (2001) Direct and Indirect Relationships between Specific Leaf Area, Leaf Nitrogen and Leaf Gas Exchange; Effects of Irradiance and Nutrient Supply. *Annals of Botany*, **88,** 915 – 927.

Mo, X., Chen, J., Ju, W. & Black, A. (2008) Optimization of ecosystem model parameters through assimilating eddy covariance flux data with an ensemble Kalman filter *Ecological Modelling*, **217**, 157-173

Moncrieff, J., Malhi, Y. & Leuning, R. (1996) The propagation of errors in long-term measurements of land-atmosphere fluxes of carbon and water. *Global Change Biology*, **2**, 231-240.

Mood, A., Graybill, F. & Boes, D. (1974). *Introduction to the theory of statistics*. 3rd ed. London: McGraw-Hill.

Mooney, H., J. Roy & B. Saugier (2001) *Terrestrial Global Productivity: Past, Present and Future*, Academic Press, San Diego.

Moran, M., Peters-Lidard, C., Watts, J. & McElroy, S. (2004) Estimating soil moisture at the watershed scale with satellite-based radar and land surface models. *Canadian Journal of Remote Sensing.* **30**, 1-22.

Morales, P., Sykes, M., Prentice, C., Smith, P., Smith, B., Bugmann, H., Zierl, B., Friedlingstein, P., Vivovy, N., Sabate, S., Sanchez, A., Pla, E., Grcia, C., Sitch, S., Arneth, A. & Ogee, J. (2005) Comparing and evaluating process-based ecosystem model predictions of carbon and water fluxes in major European forest biomes. *Global Change Biology*. **11**, 2211–2233.

Morgan, J., LeCain, D. Pendall, E. Blumenthal, D. Kimball, B. Carrillo, Y. Williams, D. Heisler-White, J. Dijkstra, F. & West., M. (2011) C4 grasses prosper as carbon dioxide eliminates desiccation in warmed semi-arid grassland. *Nature.* **476**, 202-206.

Mosegaard, K. & Tarantola, A. (1995) Monte-Carlo Sampling of Solutions to Inverse Problems. *Journal of Geophysical Research-Solid Earth*, **100**, 12431-12447.

Mrawira, D., Welch, W., Schonlau, M. & Haas, R. (1999) Sensitivity analysis of computer models: World Bank HDM-III model. *Journal of Transporting Engineering*, **125**, 421–428.

Muchow, R., Sinclair, T. & Bennett, J. (1990). Temperature and solar radiation effects on potential maize yield across locations. *Agronomy Journal*, **82**, 338– 343.

Munger, W., & Wofsy, S. (1999) EMS - Canopy-Atmosphere Exchange of Carbon, Water and Energy. *Harvard Forest Data Archive,* **HF004**.

Myneni, R., Hoffman, S., Knyazikhin, Y., Privette, J., Glassy, J., Tian, Y., Wang, X., Song, Y., Zhang, G., Smith, A., Lotsch, M., Friedl, J., Morisette, P., Votava, R., Nemani, S. & Running, S. (2002) Global products of vegetation leaf area and fraction absorbed PAR from year one of MODIS data. *Remote Sensing of Environment*, **83**, 214–231.

Myeni, R., Ramakrishna, R., Nemani, R. & Running, S. (1997) Estimation of global leaf area index and absorbed par using radiative transfer models. *Geoscience and Remote sensing*. **35(6):** 1380-1393.

Nearing, G., Crow, W., Thorp, K., Moran, M., Reichle, R. & Gupta, H. (2012) Assimilating Remote Sensing Observations of Leaf Area Index and Soil Moisture for Wheat Yield Estimates: An Observing System Simulation Experiment. *Water Resources Research*, **48(5)**, W05525.

Neftel, A., H. Friedli, E. Moor, H. Lötscher, H. Oeschger, U. Siegenthaler, & B. Stauffer. (1994) Historical CO_2 record from the Siple Station ice core. In Trends: A Compendium of Data on Global Change. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tenn., U.S.A.

Niinemets, U., Cescatti, A., Rodeghiero, M. & Tosens, T. (2005) Leaf internal diffusion conductance limits photosynthesis more strongly in older leaves of Mediterranean evergreen broad-leaved species. *Plant, Cell & Environment*, **28(12),** 1552-1566.

Oakley, J. & O'Hagan, A. (2004). Probabilistic sensitivity analysis of complex models: a Bayesian approach. *Journal of the Royal Statistical Society, Series B* **66**, 751-769.

Oakley, J. & O'Hagan, A. (2002) Bayesian inference for the uncertainty distribution of computer model outputs. *Biometrika*, **89**, 769–784.

Oakley, J. (1999) Bayesian uncertainty analysis for complex computer models. *SAMO (2004)*, 90-94.

Oak Ridge National Laboratory (2001). *MODIS Land Product Subsets* [online] Available at: <http://daac.ornl.gov/MODIS/MODIS-menu/background.html] [accessed 26th April, 2012]

Ogle, K. & Pacala, S. (2009). A modeling framework for inferring tree growth and allocation from physiological, morphological and allometric traits (2009). *Tree Physiology*, **29,** 587–605.

O'Hagan, A. & Oakley, J. E. (2004). Probability is perfect, but we can't elicit it perfectly. *Reliability Engineering and System Safety.* **85**, 239-248.

O'Hagan, A. & Kennedy, M. (2001) Bayesian calibration of computer models (with discussion). *Journal of the Royal Statistical Society, Series B.* **63**, 425-464.

O'Hagan, A. (1994). Kendall's Advanced Theory of Statistics: Bayesian Inference. Wiley.

Pan, M. & Wood, E. (2006) Data assimilation for estimating the terrestrial water budget using a constrained ensemble Kalman filter. *Journal of Hydrometeorology*, **7**, 534–547.

Parton, W., Morgan, J., Wang, G. & Del Grosso, S. (2007) Projected ecosystem impact of the Prairie Heating and CO_2 Enrichment experiment. *New Phytologist*, **174**, 823-834.

Parton, W., Holland, E., Del Grosso, S., Hartman, M., Martin, R., Mosier, A., Ojima, D. & Schimel, D. (2001) Generalized model for NOx and N20 emissions from soils. *Journal of Geophysical Research.* **106,** 17,403 – 17419.

Parton, W., Scurlock, J., Ojima, D., Gilmanov, R., Scholes, D., Schimel, T., Kirchner, J., Menaut, C., Seastedt, T., Garcia Moya, E., Apinan Kamnalrut, E. & Kinyamario, J. (1993) Observations and modeling of biomass and soil organic matter dynamics for the grassland biome worldwide. *Global Biogeochemical Cycles,* **7**, 785-809.

Parker, C. (2009) The Little REDD+ Book. An updated guide to governmental and non-governmental proposals for reducing emissions from Deforestation and degradation. London: Global Canopy Foundation.

Pearson, D. (2008) *A Primer on Data Assimilation and Parameter Estimation with Examples.* Exeter: The Hadley Centre.

Pendall, E., Bridgham, S., Hanson, P., Hungate, B., Kicklighter, D., Johnson, D., Law, B., Luo, Y., Megonigal, J., Olsrud, J., Ryan, M., Thornton, P. & Wan, S. (2004) Belowground Process Responses to Elevated CO2 and Temperature: A

Discussion of Observations, Measurement Methods, and Models. *New Phytologist.* **162**, 311–322 .
Petit, J., Jouzel, J., Raynaud, D., Barkov, N., Barnola, J. Basile, I., Bender, M., Chappellaz, J., Davisk, M., Delaygue, G. Delmotte, M., Kotlyakov, V., Legrand, M., Lipenkov, V., Lorius, C. Pe´ pin, L., Ritz, C., Saltzmank, E. & Stievenard, M. (1999). Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. *Nature*, **399**, 429-436.
Pólya, G. **(1920)** *Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momenten-problem, Mathematische Zeitschrift*, **8,** 171–181.
Poorter, H., Niinemets, Ü., Poorter, L., Wright, I. J. & Villar, R. (2009), Causes and consequences of variation in leaf mass per area (LMA): a meta-analysis. *New Phytologist*, **182,** 565–588.
Post, W., Peng, T., Emanuel, W., King, A., Dale, V. & DeAnglis, D. (1990) The Global Carbon Cycle. *American Scientist.* **78**, 310–326.
Potter, C., Randerson, J., Field, C., Matson, P., Vitousek, P., Mooney, H., Klooster, S. (1993) Terrestrial Ecosystem Production: A process model based on global satellite and surface data. *Global Biogeochemical Cycles*, **7,** 811 – 841.
Press, W., Teukolsky, S., Vetterling, W., Flannery, B. (1992). Numerical Recipes in FORTRAN. The Art of Scientific Computing, 2nd edition, Cambridge University Press, Cambridge, pp963.
Privette, J., Myneni, R., Knyazikhin, Y., Mukufute, M., Roberts, G., Tian, Y., Wang, Y., Leblanc, S. (2002). Early spatial and temporal validation of MODIS LAI product in Africa. *Remote Sensing of Environment*, **83**, 232−243.
Quaife, T., t.l.quaife@exeter.ac.uk (2011) *Particle Filter* [e-mail]. Message to E. Ryan (edmund.ryan@shef.ac.uk) sent 21st November, 2011.
Quaife, T., Lewis, P. (2010) Temporal Constraints on Linear BRDF Model parameters. *IEEE Transactions on Geosciences and Remote sensing*. **48(5),** 2445-2450.
Quaife, T., Lewis, P., De Kauwe, M. (2008) Assimilating Earth observation data into land surface models. UCL, U.K.. Unpublished.
Quaife, T., Lewis, P., De Kauwe, M., Williams, M., Law, B., Disney, M. & Bowyer, P. (2007) Assimilating canopy reflectance data into an ecosystem model with an Ensemble Kalman Filter. *Remote Sensing of Environment*, **112**, 1347-1364.
Quegan, S., LeToan, T., Chave, J., Dall, J., Pappathanassiou, K., Rocca, F., Ulander, L., & Williams, M. (2012) BIOMASS: report for mission selection. *European Space Agency*, Norwiijk, Netherlands.
Raich, J., Rastetter, E., Melillo, J., Kicklighter, D., Steudler, P., Peterson, B., Grace, A., Moore, B. & Vorosmarty, C. (1991) Potential Net Primary Productivity in South America: Application of a global model. *Ecological Applications*, **1(4),** 399 – 429.
Raupach, M., Rayner, P. & Barrett, D. (2005) Model-data synthesis in terrestrial carbon observation: methods, data requirements and data uncertainty specifications. *Global Change Biology*, **11**, 378-397.
Rayner, P. (2010). The current state of carbon-cycle data assimilation. *Current Opinion in Environmental Sustainability*, **2**, 289-296.

Rayner, P., Scholze, M., Knorr, W., Kaminski, T., Giering, R. & Widmann, H. (2005) Two decades of terrestrial carbon fluxes from a carbon cycle data assimilation system (CCDAS). *Global Biogeochemical Cycles*. **19**, GB2026,

Reichle, R. (2001) Extended versus Ensemble Kalman Filtering for Land Data Assimilation. Journal of Hydrometeorology, **3**, 728-740.

Reichstein, M., Falge, E., Baldocchi, D., Papale, D., Aubinet, M., Berbigier, P., Bernhofer, C., Buchmann, N., Gilmanov, T., Granier, A., Grunwalk, T., Havrankova, K., Ilvesniemi, H., Janous, D., Knohl, A., Laurila, T., Lohila, A. & Loustau, D. (2005) On the separation of net ecosystem exchange into assimilation and ecosystem respiration: review and improved algorithm. *Global Change Biology*. **9**, 1424 – 1439.

Richardson, A., Williams, M., Hollinger, D., Moore, D., Dail, D., Davidson, E., Scott, N. Evans, R., Hughes, H., Lee, J., Rodrigues, C. & Savage, K. (2010). Estimating parameters of a forest ecosystem C model with measurements of stocks and fluxes as joint constraints. *Oecologia*, **164**, 25-40.

Richardson, A., Mahecha, M., Falge, E., Kattge, J., Moffat, A., Papale, D., Reichstein, M., Stauch, V., Braswell, B., Churkina, G., Kruijt, B. & Hollinger, D. (2007) Statistical properties of random CO_2 flux measurement uncertainty inferred from model residuals. *Agricultural and Forest Meteorology*, **148**, 38-50.

Richardson, A., Hollinger, D. & Burba, G. (2006) A multi-site analysis of random error in tower-based measurements of carbon and energy fluxes. *Agricultural and Forest Meteorology*, **136**, 1-18.

Rivkin, R. & Legendre, L. (2001) Biogenic Carbon Cycling in the Upper Ocean: Effects of Microbial Respiration. *Science* **23***: 2398-2400.*

Rodeghiero, M., Niinemets, U. & Cescatti, A. (2007) Major diffusion leaks of clamp-on leaf curvettes still unaccounted: how erroneous are the estimates of Farquhar et al. model parameters? *Plant, Cell and Environment*, **30**, 1006-1022.

Rodgers, C. (2000). *Inverse methods for atmospheric sounding*. Singapore: World scientific.

Rougier, J. (2011). Uncertainty Assessment for complex systems [PowerPoint presentation]. Department of Probability and Statistics, University of Sheffield.

Rougier J., Sexton, D., Murphy, J. & Stainforth, D. (2008), Analysing the climate sensitivity of the HadSM3 climate model using ensembles from different but related experiments. *Journal of Climate*, **22(13)**, 3540-3557.

Rougier, J. (2008). Efficient Emulators for Multivariate Deterministic Functions. *Journal of Computational and Graphical Statistics*, **17(4),** 827-843.

Roy, D. Lewis, P. Schaaf, C. Devadiga, S. & Boschetti, L. (2006) The global impact of clouds on the production of MODIS bidirectional reflectance model-based composites for terrestrial monitoring. *IEEE Transactions in Geoscience Remote Sensing.* **3(4)**, 452–456.

Ruimy, A., Saugier, B., & Dedieu, G. (1994) Methodology for the estimation of terrestrial net primary production from remotely sensed data (1994) *Journal of Geophysical Research*, **99,** 5263–5283.

Running, S., Nemani, R., Heinsch, F., Zhao, M., Reeves, M. & Hashimoto, H. (2004) A continuous Satellite-Derived Measure of Global Terrestrial Primary Production. *Bioscience* **54(6)**, 547–560.

Sabine, C., Feely, R., Gruber, N., Key, R., Lee, K., Bullister, J., Wanninkhof, R., Wong, C., Wallace, D., Tilbrook, B., Millero, F., Peng, T., Kozyr, A., Ono, T. & Rios, A. (2004a) The Oceanic Sink for Anthropogenic CO_2. *Science:* **305 (5682)**, 367-371.

Sabine, C., Heimann, M. & Artaxo, P. (2004b) Current status and past trends of the global carbon cycle. In: Field CB and Raupach MR(eds) SCOPE 62, The Global Carbon Cycle: Integrating Humans, Climate, and the Natural World, Chap. 2, pp. 17–44. Washington, DC: Island Press.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., Tarantola, S. (2008) *Global Sensitivity Analysis. The Primer*. Chichesster:Wiley.

Saltelli, A., Chan, K. and Scott, M. (eds) (2000) *Sensitivity Analysis*. New York: Wiley.

Santner, T. (2003) *The Design and Analysis of Computer Experiments*. Springer-Verlag.

Sarmiento, J. & Gruber, N. (2006) *Ocean Biogeochemical Dynamics*, Princeton, NJ: Princeton University Press, 526pp.

Saugier, B., Roy, J. & Mooney, H. (2001) *Terrestrial Global Productivity.* Academic Press, San Diego, CA.

Sea, W., Choler, P., Beringer, J., Weinmann, R., Hutley, L. & Leuning, R. (2011) Documenting improvements in leaf area index estimates from MODIS using hemiscpherical photos from Austrialian savannas. *Agricultural and Forest Meteorology*, **151**, 1453-1461.

Sellers, P. (1997). Modeling the exchange of energy, water, and carbon between continents and atmosphere. *Science*, **275**, 602−609.

Sitch, S., Smith, B., Prentice, I., Arneth, A., Bondeau, A., Cramer, W., Kaplans, J., Levis, S., Lucht, W., Sykes, M., Thonicke, K. & Venevsky, S. (2003) Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model. *Global Change Biology*, **9,** 161–185.

Smith, P., Baines, M., Dance, S., Nichols, N. & Scott, T. (2008) Data Assimilation for Parameter Estimation with Application to a simple Morphodynamic Model. *Mathematics Report 2/2008, Department of Mathematics*, University of Reading.

Smith, F., Sampson, A. & Long, N. (1991). Comparison of leaf area index estimates from tree allometrics and measured light interception. *Forest Science.* **37**, 1682-1688.

Solomon, S., Qin, D., Manning, M, Chen, Z, Marquis, M, Averyt, K.B., Tignor M. & Miller, H. (2007) IPCC, 2007: Climate Change 2007: The Physical Science Basis. Contributions of Working Group I to the Fourth Assessment. *Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.*

Spadavecchia, L., Williams, M. & Law, B. (2011) Uncertainty in predictions of forest carbon dynamics - separating driver error from model error. *Ecological Applications.* **21**, 1506-1522.

Spadavecchia, L., Williams, M. & Law, B. (2008) Estimation of Landscape Budgets: Combining Geostatistical and Data Assimilation Approaches. *Unpublished thesis (PhD).* University of Edinburgh.

Steffen, W., Noble, I. & Canadell, J. (1998) The terrestrial carbon cycle: Implications for the Kyoto Protocol, *Science*, **280**, 1393– 1394.

Stroud, J., Stein, M., Lesht, B., Schwab, D. & Beletsky, D. (2010) An Ensemble Kalman Filter and Smoother for Satellite Data Assimilation. *Journal of the American Statistical Association.* **105:491**, 978-990.

Sus, O., Williams, M., Bernhofer, C., Beziat, P., Buchmann, N., Ceschia, E., Doherty, R., Eugster, W., Gruenwald, T., Kutsch, W., Smith, P., & Wattenbach, M (2010) A linked carbon cycle and crop developmental model Residue, Respiration, and Residuals: Evaluation of a Dynamic 25 Agroecosystem Model Using Eddy Flux Measurements and Biometric Data: Description and evaluation against measurements of carbon fluxes and carbon stocks at several European agricultural sites. *Agriculture, Ecosystem Environment*, **139**, 402–418.

Sus, O., Heuer, M. Meyers, T. & Williams, M. (2012) A data assimilation framework for constraining upscaled cropland carbon flux seasonality and biometry with MODIS. *Biogeosciences Discussions*, **9**, 11139-11177.

Tang, J. & Zhuang Q. (2008) Equifinality in parameterization of process-based biogeochemistry models: A significant uncertainty source to the estimation of regional carbon dynamics. *Journal of Geophysical Research-Biogeosciences*, **113**.

Tans, P. & Conway, T. (2011) Trends in atmospheric carbon dioxide. NOAA/ESRL (www.esrl.noaa.gov/gmd/ccgg/trends/).

Tarantola, A. (1987) Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation, Elsevier, New York.

Tian, X., Xie, Z. & Dai, A. (2008) A land surface soil moisture data assimilation system based on the dual-UKF method and the Community Land Model. *Journal of Geophysical Research*, **113**, DI4127.

Trefethen, L. & Weiderman, J. (1991) Two results on polynomial interpolation in equally spaced points, *Journal of Approximation Theory*, **65**, 247-260.

Trudinger, C., Raupach, M. & Rayner, P. (2007) OptIC project: An intercomparison of optimization techniques for parameter estimation in terrestrial biogeochemical models. *Journal of Geophysical Research-Biogeosciences*, **112**, G02027.

Trudinger, C., Raupach, M., Rayner, P. & Enting, I. (2008) Using the Kalman filter for parameter estimation in biogeochemical models. *Environmetrics*, **19**, 849-870.

Tucker, C. J., Slayback, D. A., Pinzon, J. E., Los, S. O., Myneni, R. B., & Taylor, M. G. (2001). Higher northern latitude normalized difference vegetation index and growing season trends from 1982 to 1999. *International Journal of Biometeorology*, **45(4),** 184-190.

Turner, D., Ritts, W., Cohen, W., Maeirsperger, T., Gower, S. & Kirschbaum, A. (2005) Site-level evaluation of satellite-based global terrestrial gross primary production and net priamary production monitoring. *Global Change Biology*, **11**, 666-684.

Upton, G. & Cook, I. (2004) *Oxford Dictionary of Statistics.* New York: Oxford University Press Inc.

Van der Werf, G., Morton, D., DeFries, R., Olivier, J., Kasibhatla, P., Jackson, R., Collatz G. & Randerson, J. (2009) CO_2 emissions from forest loss. *Nature Geoscience*. **2**, 737-738.

Van Dongon, S. (2006) Prior specification in Bayesian statistics: three cautionary tales. *Journal of Theoretical Biology*. **242(1)**: 90-100.

Van Leeuwen P. (2010) Nonlinear data assimilation in geosciences: an extremely efficient particle filter *Quarterly Journal of the Royal Meteorological Society.* **136**, 1991–1999.

Van Leeuwen, P. (2009) Particle Filtering in Geophysical Systems. *American Meteorological Society* **137**, 4089 – 4114

Van Wijk & M., Bouton, W. (2001) Towards understanding tree root profiles: simulating hydrologically optimal strategies for root distribution. *Hydrology and Earth System Sciences*, **5(4)**, 629-644.

Vickers, D., Thomas, C. & Law, B. (2009) Random and systematic CO2 flux sampling errors for tower measurements over forests in the convective boundary layer. *Agricultural and forest Meteorology*, **149**, 73-89.

Vrugt, J., Diks, C., Gupta, H., Bouten, W. & Verstraten, M. (2005) Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation. *Water resources research*, **41**, 1-17.

Vrugt, J., Gupta, H., Bouten, W. & Sorooshian, S. (2003a). A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. *Water Resources Research.* **39 (8)**, 1201.

Vrugt, J., Gupta, H., Bastidas, A., Bouten.,W. & Sorooshian, S. (2003b) Effective and efficient algorithm for multiobjective optimization of hydrological models. *Water resources research*, **39(8)**, 1214.

Wang, G. (2003) Adaptive Response Surface Method Using Inherited Latin Hypercube Design Points. *Journal of Mechanical Design*, **125**, 210-220.

Wang, Y., Tian, Y., Zhang, Y., El-Saaleous, N., Knyazikhin, Y., Vermote, R., & Myneni, R. (2001). Investigation of product accuracy as a function of input and model uncertainties: Case study with SeaWiFS and MODIS LAI/FPAR algorithm. *Remote Sensing of Environment*, **78**, 296−311.

Wang, Y., Baldocchi, D., Leuning, R., Falge, E. & Vesala, T. (2007) Estimating parameters in a land-surface model by applying nonlinear inversion to eddy covariance flux measurements from eight FLUXNET sites. *Global Change Biology*, **13**, 652-670.

Wardlow, B., Kastens, J., & Egbert, S. (2006) Using USDA Crop Progress Data for the Evaluation of Greenup Onset Date Calculated from MODIS 250-Meter Data, Photogramm. *Engineering Remote Sensing*, **72,** 1225–1234.

Welch, W., Buck, R., Sacks, J., Wynn, H., Mitchell, T. & Morris, M. (1992) Screening, predicting, and computer experiments. *Technometrics*, **34**, 15–25.

Weiss, M., Baret, F., Garrigues, S. & Lacaze, R. (2007) LAI and FAPAR CYCLOPES global products derived from Vegetation. Part 2 : validation and comparison with MODIS C4 products. *Remote Sensing of Environment.* **110**:317-331.

Whittle, M., Quegan, S., Uryu, Y., Stuewe, M. & Yulianto, K. (2012). Detection of tropical deforestation using ALOS-PALSAR: a Sumatran case study, *Remote Sensing Environment*, 83-89.

Williams, M., mat.williams@ed.ac.uk (2012) Leaf Mass per Area [e-mail]. Message to E. Ryan (edmund.ryan@shef.ac.uk) sent 11th October, 2012.

Williams, M, Hill, T. & Ryan, C. (2012). Using biomass distributions to determine probability and intensity of tropical forest disturbance. *Plant Ecology and Diversity* [online only].

Williams, M., mat.williams@ed.ac.uk (2010) *Simulating a drought in DALEC* [e-mail]. Message to E. Ryan (edmund.ryan@shef.ac.uk) sent 4th August, 2010.

Williams, M, Richardson, A. & Reichstein, M. (2009) Improving land surface models with FLUXNET data. *Biogeosciences,* **6**, 1341-1359.

Williams, M, Schwarz, P., Law, B., Irvine, J. & Kurpius, M. (2005) An improved analysis of forest carbon dynamics using data assimilation. *Global Change Biology,* **11**, 89-105.

Williams, M., Rastetter, E. & Fernandes, D. (1996) Modelling the soil-plant-atmosphere continuum in a Quercus-Acer stand at Harvard forest: The regulation of stomatal conductance by light, nitrogen and soil/plant hydraulic properties. *Plant Cell and Environment,* **19**, 911-927.

Williams, M., Rastetter, E., Fernandes, D., Goulden, M., Shaver, G. & Johnson, L. (1997) Predicting gross primary productivity in terrestrial ecosystems. *Ecological Applications,* **7**, 882-894.

Williams, M., Baron, J., Caine, N., Sommerfeld, R. & Sanford, R. (1995) Nitrogen Saturation in the Rocky Mountains. *Environmental Science & Technology*, **30(2)**, 640 – 646.

Wofsy, S., Goulden, M., Munger, J., Fan, S., Bakwin, P., Daube, B., Bassow, S. & Bazzaz, F. (1993) Net exchange of CO_2 in a midlatitude forest. *Science* **260**, 1314-1317.

Woodward, F.I. & Lomas, M. (2004). Vegetation dynamics - simulating responses to climatic change. *Biological Reviews*, **79**, 643-670.

Woodward F.I., Lomas M. & Lee S. (2001) Predicting the future productivity and distribution of global terrestrial vegetation. In: Terrestrial Global Productivity (eds Roy J, Saugier B, Mooney HA), pp. 521–541. Academic Press, New York.

Woodward, F.I., Smith, T. & Emanuel, W. (1995) A global primary productivity and phytogeography model. *Global Biogeochemical Cycles*, **9**, 471–490.

Xu, Q. (2006) Measuring informaton content from observations for data assimilation: relative entropy versus shannon entropy difference. *Tellus*. **59A**, 198-209.

Xu, L., Baldocchi, D. & Tang, J. (2004), How soil moisture, rain pulses, and growth alter the response of ecosystem respiration to temperature. *Global Biogeochemical Cycles*, **18**, GB4002.

Xue, H., Wang, J. & Ping, C. (2011) Validation of collection 5 MODIS LAI product by scaling-up method using field measurements. *Remote Sensing*. International Society for Optics and Photonics, 2011.

Yang, W., Tan, B., Huang, D., Rautiainen, M., Shabanov, N., Wang, Y., Privette, J., Huemmrich, K., Fensholt, R., Sandholt, I., Weiss, M., Ahl, D., Gower, S., Nemani, R., Knyazikhin, Y., & Myneni, R. (2006). MODIS leaf area index products: From validation to algorithm improvement. *IEEE 660 Transactions on Geoscience and Remote Sensing*, **44(7)**, 1885–1898.

Yuan, H., Dai, Y., Xiao, Z., Ji, D. & Shangguan, W. (2011) Reprocessing the MODIS leaf area index products for land surface and climate modelling. *Remote Sensing of Environment*. **115(5),** 1171-1187.

Yuan, W., Liu, S., Zhou, G., Zhou, G., Tieszen, L., Baldocchi, D., Bernhofer, C., Gholz, H., Goldstein H., Goulden, M., Hollinger, D., Hu, Y., Law, B., Stoy, P., Vesala, T. & Wofsy, S. (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**, 183 – 207.

Ziehn, T., Scholze, M. & Knorr, W. (2012) On the capability of Monte Carlo and adjoint inversion techniques to derive posterior parameter uncertainties in terrestrial ecosystem models. *Global Biochemical Cycles,* **26,** GB3025.

Zobitz, J., Desai, A., Moore, D. & Chadwick, M. (2011) A primer for data assimilation with ecological models using Markov Chain Monte Carlo (MCMC). *Oceologia*, **167**, 599-611.

Zobitz, J., Moore, D., Sacks, W., Monson, R., Bowling, D. & Schimel, D. (2008) Integration of process-based soil respiration models with whole-ecosystem CO_2 measurements. *Ecosystems,* **11**, 250-269.