Fourth Monitoring Committee Meeting

Understanding the Information Content in Diverse Observations of Forest Carbon Stocks and Fluxes for Data Assimilation and Ecological Modeling
NERC case partnership with Forest Research

E. Pinnington

May 14, 2015

1 Introduction

Forest ecosystems play a large role in reducing the impact of anthropogenic CO_2 emissions and therefore understanding their response to climate change is important. Currently, in data assimilation, the optimal set of observations to add the most information to models of forest carbon balance is not known.

In our last monitoring committee we saw some results of information content experiments using the Data Assimilation Linked Ecosystem Carbon model (DALEC) [Williams et al., 2005] taken from a report I had written. The report had focused mainly on Shannon Information Content (SIC) [Rodgers et al., 2000] and Degrees of Freedom for Signal (DFS) [Fowler and van Leeuwen, 2011] as measures for information content in carbon balance observations. It was shown previously that a single observation of the Net Ecosystem Exchange (NEE) of CO_2 taken in summer with warmer temperatures and higher daily irradiances had a higher information content than those taken in winter. This made physical sense as observations of NEE in the summer are of greater magnitude and give more information about the fluxes of carbon throughout the forest. In the report on information content this was explored further and it was shown that it would take 30 days of successive NEE observations in winter to give the same amount of information as a single NEE observation made in summer.

In the last report DALEC was also implemented in a Four-Dimensional Variational assimilation (4D-Var) scheme for state estimation, where our state corresponded to the 5 initial carbon pool values in DALEC. We used this scheme to assimilate observations of NEE with DALEC using data from an evergreen forest in Oregon. It was shown that assimilating NEE observations not only improved our models estimate of NEE but also improved modelled Gross Primary Production (GPP) and Total ecosystem Respiration (RT). The field work I had been doing at the Alice Holt forest with my case partner Forest Research was also outlined in the last report.

Since completing the last report a new version of the DALEC model was released (DALECV2 [Bloom and Williams, 2014]) and I have built this into a 4D-Var scheme for parameter and state estimation. It was decided to begin work on DALECV2 as it can be parameterised for both deciduous and evergreen forests and the Alice Holt research site is a mainly deciduous forest. Whereas the version of DALEC previously used was a model of evergreen forests only. We are currently using data acquired from Alice Holt to run DALECV2 and I have worked on getting the driving data and NEE observations into the right format for use with DALECV2.

I have also now completed taking a set of stem respiration observations for Forest Research at the Alice Holt flux site and am now in the early stages of planning a campaign to take leaf area index measurements at the Alice Holt flux site. I have completed a PhD plan which is included along with the monitoring report.

2 DALECV2

The DALECV2 model is a simple process-based model describing the carbon balance of a forest ecosystem [Bloom and Williams, 2014] and is the new version of the original DALEC [Williams et al., 2005]. The model is constructed of six carbon pools (labile (C_{lab}), foliage (C_f), fine roots (C_r), woody stems and coarse roots (C_w), fresh leaf and fine root litter (C_l) and soil organic matter and coarse woody debris (C_s)) linked via fluxes. The aggregated canopy model (ACM) [Williams et al., 1997] is used to calculate daily gross primary production (GPP) of the forest, taking meteorological driving data and the site's leaf area index (a function of C_f) as arguments.

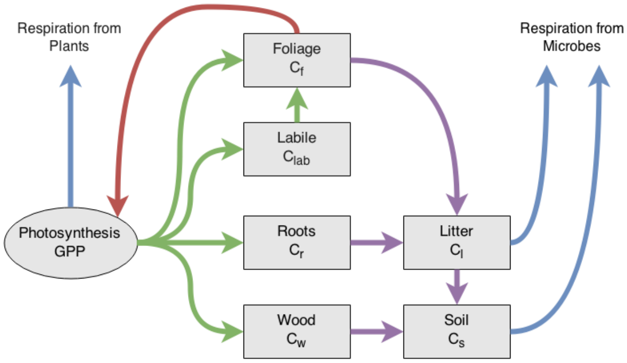

Figure 1: Representation of the fluxes in the DALEC2 carbon balance model. Green arrows represent C allocation, purple arrows represent litter fall and decomposition fluxes, blue arrows represent respiration fluxes and the red arrow represents the feedback of foliar carbon to the GPP function.

3 4D-Var

In 4D-Var we aim to maximise the probability of our initial state $\mathbf{x}_0$ given a set of observations $\mathbf{y}$, $P(\mathbf{x}_0|\mathbf{y})$, over some time window, N. $P(\mathbf{x}_0|\mathbf{y})$ is maximised by minimising a cost function $J(\mathbf{x})$ derived from Baye's Theorem [Lewis et al., 2006]. The cost function is given as,

$$J(\mathbf{x}_0) = \frac{1}{2}(\mathbf{x}_0 - \mathbf{x}_b)^T\mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}_b) + \frac{1}{2}\sum_{i=0}^{N}(\mathbf{y}_i - h_i(\mathbf{x}_i))^T\mathbf{R}_i^{-1}(\mathbf{y}_i - h_i(\mathbf{x}_i)), \qquad (1)$$

where $\mathbf{x}_b$ is our background and acts as our initial guess to our state $\mathbf{x}_0$, $\mathbf{B}$ is the background error covariance matrix and quantifies our knowledge of the error in our background, h_i is our observation operator at time t_i and maps our state vector evolved by our nonlinear model ($m_{0\rightarrow i}(\mathbf{x}_0) = \mathbf{x}_i$) to the observations at this time ($\mathbf{y}_i$) and $\mathbf{R}_i$ is the observation error covariance matrix at time t_i and represents our knowledge of the uncertainty in the observations. The state that minimises the cost function is called the analysis and is denoted as $\mathbf{x}_a$, this state is found using a minimisation routine that takes the cost function, our initial guess ($\mathbf{x}_b$) and also the gradient of the cost function defined as,

$$\nabla J(\mathbf{x}_0) = \mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}_b) - \sum_{i=0}^{N} \mathbf{M}_{i,0}^T \mathbf{H}_i^T \mathbf{R}_i^{-1}(\mathbf{y}_i - h_i(\mathbf{x}_i)), \tag{2}$$

where $\mathbf{H}_i = \frac{\delta h_i(\mathbf{x}_i)}{\delta \mathbf{x}_i}$ is our linearized observation operator and $\mathbf{M}_{i,0} = \mathbf{M}_{i-1}\mathbf{M}_{i-2}\cdots\mathbf{M}_0$ is our tangent linear model with $\mathbf{M}_i = \frac{\delta m_i(\mathbf{x}_i)}{\delta \mathbf{x}_i}$. We can rewrite the cost function and its gradient to avoid the sum notation as,

$$J(\mathbf{x}_0) = \frac{1}{2}(\mathbf{x}_0 - \mathbf{x}_b)^T \mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}_b) + \frac{1}{2}(\hat{\mathbf{y}} - \hat{h}(\mathbf{x}_0))^T \hat{\mathbf{R}}^{-1}(\hat{\mathbf{y}} - \hat{h}(\mathbf{x}_0)) \tag{3}$$

and

$$\nabla J(\mathbf{x}_0) = \mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}_b) - \hat{\mathbf{H}}^T \hat{\mathbf{R}}^{-1}(\hat{\mathbf{y}} - \hat{h}(\mathbf{x}_0)), \tag{4}$$

where,

$$\hat{\mathbf{y}} = \begin{pmatrix} \mathbf{y}_0 \\ \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_N \end{pmatrix}, \ \hat{h}(\mathbf{x}_0) = \begin{pmatrix} h_0(\mathbf{x}_0) \\ h_1(m_{0\to 1}(\mathbf{x}_0)) \\ \vdots \\ h_N(m_{0\to N}(\mathbf{x}_0)) \end{pmatrix}, \ \hat{\mathbf{R}} = \begin{pmatrix} \mathbf{R}_0 & 0 & 0 & 0 \\ 0 & \mathbf{R}_1 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \mathbf{R}_N \end{pmatrix} \text{ and } \ \hat{\mathbf{H}} = \begin{pmatrix} \mathbf{H}_0 \\ \mathbf{H}_1\mathbf{M}_0 \\ \vdots \\ \mathbf{H}_N\mathbf{M}_{N,0} \end{pmatrix}. \tag{5}$$

4 4D-Var with DALECV2

In our DALECV2 4D-Var scheme the state vector, $\mathbf{x}_0$, corresponds to the vector of the 17 model parameters and 6 initial carbon pool values. We use a diagonal approximation to our background and observational error covariance matrices so that, $\mathbf{B} = \underline{\sigma}_b^2\mathbf{I}$ and $\hat{\mathbf{R}} = \underline{\sigma}_o^2\mathbf{I}$, where $\underline{\sigma}_b$ and $\underline{\sigma}_o$ are the vectors of the background and observational standard deviations respectively.

In order to find the tangent linear model (TLM) for DALECV2 we need to find the derivative of the model at each time step with respect to the 17 model parameters and the 6 carbon pools. Previously the DALEC TLM was calculated by hand, however, now that we are working with a more complex model and an extra 18 state members we use the AlgoPy automatic differentiation package in Python. It is important to test the tangent linear hypothesis as we did with the original DALEC. In 4D-Var we assume the tangent linear hypothesis,

$$m_{0\to i}(\mathbf{x}_0 + \delta\mathbf{x}_0) \approx m_{0\to i}(\mathbf{x}_0) + \mathbf{M}_{i,0}\delta\mathbf{x}_0. \tag{6}$$

The validity of this assumption depends on how nonlinear the model is, the length of the assimilation window and the size of the perturbation $\delta\mathbf{x}_0$. We can test the validity for DALECV2 by taking an initial state $\mathbf{x}_0$ and a 5% perturbation for $\delta\mathbf{x}_0$. We then rearrange equation 6 to find,

$$\text{percentage error in TLM} = \left| \frac{m_{0\to i}(\mathbf{x}_0 + \delta\mathbf{x}_0) - m_{0\to i}(\mathbf{x}_0)}{\mathbf{M}_{i,0}\delta\mathbf{x}_0} - 1 \right| \times 100. \tag{7}$$

In figure 2 we can see that our TLM for DALECV2 performs very well after being run forward a year with less than a 1% error. By the second year we see some peaks in our error in spring and autumn, this is where functions controlling leaf on and leaf off in the nonlinear DALECV2 have gone out of phase with the TLM. Even at these peaks our error is still reasonable reaching a maximum at 7% and then coming back to around 1%.

In figure 3 we see a 4D-Var run using NEE observations and meteorological driving data from Alice Holt. Here we are using a truncated Newton method [Nocedal and Wright, 1999] from the

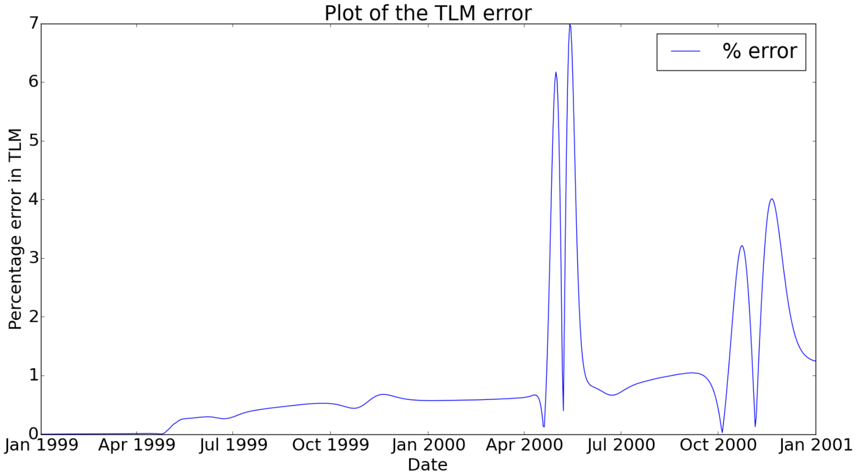

Figure 2: Plot of the percentage error in our tangent linear model for DALECV2 when evolving our state forward over a period of two years.

Python package Scipy.optimize to find the minimum of our 4D-Var cost function. Our $\mathbf{x}_b$ is a parameter set found by the University of Edinburgh using the CARbon DAta-MOdel fraMework (CARDAMOM) [Exbrayat et al., 2015]. This used Harmonised World Soil Database C_s observations as initial conditions, meteorological driving data from ERA-interim and Markov chain Monte Carlo techniques to assimilate MODIS leaf area index observations over a 10 year period.

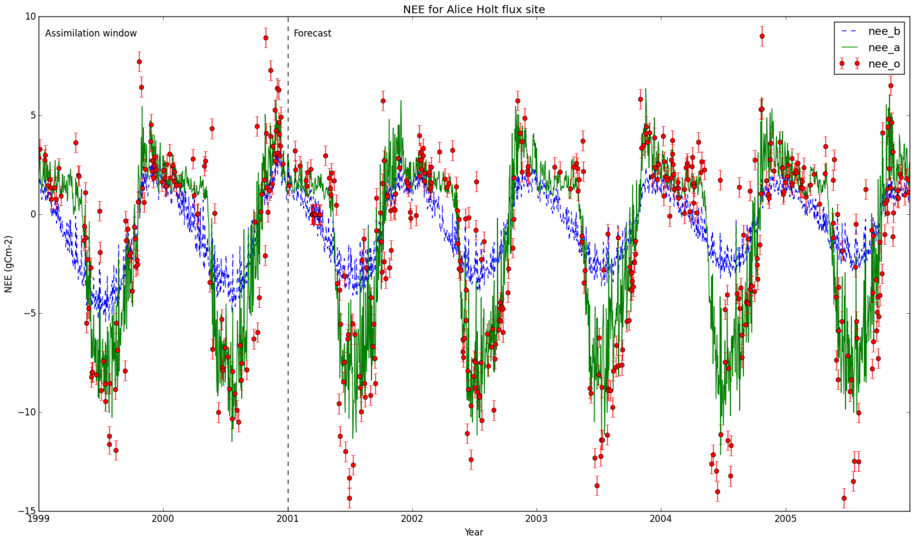

Figure 3: 4D-Var run for Alice Holt flux site, assimilating NEE observations from 1999-2001. The red dots are our observations, the blue dotted line our background estimate and the green line is our analysis.

From figure 3 we can see that after assimilating NEE observations from 1999-2001 and then running our analysis forward from 2001-2006 we have a much better forecast of NEE than we did from our initial background guess (when judging against observations).

I am currently working on getting DALECV2 coded up with a 4D-Var scheme in Python, with a state made up of the initial carbon pool values and the 17 model parameters so that the model can be parameterised for the Alice Holt research site.

5 Current Work and Future Plans

I am currently repeating the information content experiments conducted with DALECV2. While applying the measures previously used (SIC and DFS) I am also starting to work with the influence matrix [Cardinali et al., 2004] and the adjoint technique proposed by Langland and Baker [2004]. The influence matrix is defined as,

$$\mathbf{S} = \frac{\delta\hat{\mathbf{H}}\mathbf{x}_a}{\delta\hat{\mathbf{y}}} = \hat{\mathbf{R}}^{-1}\hat{\mathbf{H}}\mathbf{A}\hat{\mathbf{H}}^T, \tag{8}$$

where $\mathbf{A}$ is our analysis error covariance matrix. Using these measures and an observing system simulation experiment I have developed I will produce another report investigating the best set of observations for understanding the carbon balance of a forest and apply the results to the available data from Alice Holt.

I am also in the very early stages of planning a fieldwork campaign at Alice Holt. In this campaign I will be taking measurements of leaf area index using a ceptometer and hemispherical photographs. Tristan and I have meeting with Forest Research on the 28^{th} May to further plan the fieldwork. These measurements will be used in the information content experiments but also to address site specific science questions. Half of the Alice Holt forest has been thinned and the other half left unmanaged. The flux tower measuring NEE is placed directly on the boundary of the two halves. The leaf area index measurements will allow us to compare the difference between the two sides and also evaluate DALECV2's estimate of leaf area index. We will also parameterise two versions of DALECV2 for the thinned/unthinned forest using NEE observations and a tower foot print model to split the NEE observations when the signal is coming from the thinned or unthinned side of the forest. We will then be able to see if by just assimilating NEE we can pick out a difference in LAI between the two sides.

Other future work will be based on the attached PhD outline. This includes attempting to improve the representation of background and observational error covariance matrices in DALECV2. This is important as improving our representation of these matrices could lead to different results in the information content experiments.

6 Professional and Academic Development

6.1 Masters Courses

- MAMB10 (Data Assimilation) - 85%
- MAMNSO (Numerical Solutions to Ordinary Differential Equations) - 79%
- MTMG02 (Atmospheric Physics) - 66%
- MTMG49 (Boundary Layer) - 72%
- MTMD01 (Environmental Data Visualization) - 78%
- MTMD02 (Operational Data Assimilation) - 70%

6.2 Transferable Skills

During my PhD I have taken part in the following courses, workshops and activities:

- 28/01/2014 - Basic Statistics Refresher - RRDP
- 31/03/2014-01/04/2014 - Land Data Assimilation workshop at UCL - ESA
- 23/04/2014-25/03/2014 - Correlated Observation Errors in Data Assimilation Workshop - ESA
- 13/05/2014 - Social Media - Bloggs, Twitter and Your Online Presence - RRDP
- 29/05/2014 - How to Write a Paper - RRDP
- 25/06/2014-26/06/2014 - Software Carpentry Course - Git and Python
- 10/07/2014-11/07/2014 - Forest Research - Helped with field work LiDAR
- 21/07/2014-01/08/2014 - Fluxcourse Summer School - University of Colorado
- 29/09/2014-03/10/2014 - NERC course - Software Development for Environmental Scientists Level 1
- 08/10/2014-10/10/2014 - Environment YES - NERC "dragon's den" type competition at Syngenta, Jesops Hill
- 17/12/2014 - Presentation at Maths for Planet Earth Industry day
- 24/02/2015 - Reading Soil Centre Workshop - What can Land Surface Models do for you?
- 23/03/2015-27-03/2015 - NERC course - Software Development for Environmental Scientists Level 2

6.3 Demonstrating

During my PhD I have helped demonstrate on the following courses:

- 15/09/2014-1909/2014 - NERC Data assimilation for environmental scientists training course
- 16/02/2015-20/02/2015 - NERC Software Development for Environmental Scientists Level 1
- 20/04/2015-23/04/2015 - MT26E Surface Energy Exchange Practicals

References

Anthony Bloom and Mathew Williams. Constraining ecosystem carbon dynamics in a data-limited world: integrating ecological "common sense" in a model-data-fusion framework. *Biogeosciences Discussions*, 11(8):12733–12772, 2014. URL http://www.biogeosciences-discuss.net/11/12733/2014/.

C. Cardinali, S. Pezzulli, and E. Andersson. Influence-matrix diagnostic of a data assimilation system. pages 2767–2786, 2004. ISSN 00359009. doi: 10.1256/qj.03.205. URL http://centaur.reading.ac.uk/9477/.

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

Alison Fowler and Peter Jan van Leeuwen. Measures of observation impact in gaussian data assimilation. *Reading University*, 2011.

Rolf H. Langland and Nancy L. Baker. Estimation of observation impact using the NRL atmospheric variational data assimilation adjoint system. *Tellus, Series A: Dynamic Meteorology and Oceanography*, 56(3):189–201, 2004. ISSN 02806495. doi: 10.1111/j.1600-0870.2004.00056.x.

John M Lewis, Sivaramakrishnan Lakshmivarahan, and Sudarshan Dhall. *Dynamic data assimilation: a least squares approach*, volume 13. Cambridge University Press, 2006.

Jorge Nocedal and Stephen J Wright. *Numerical Optimization.* Springer Science & Business Media, 1999. ISBN 0387987932. URL http://books.google.co.uk/books/about/Numerical_Optimization.html?id=epc5fX0lqRIC&pgis=1.

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

Mathew Williams, Edward B Rastetter, David N Fernandes, Michael L Goulden, Gaius R Shaver, and Loretta C Johnson. Predicting gross primary productivity in terrestrial ecosystems. *Ecological Applications*, 7(3):882–894, 1997.

Mathew Williams, Paul A Schwarz, Beverly E Law, James Irvine, and Meredith R Kurpius. An improved analysis of forest carbon dynamics using data assimilation. *Global Change Biology*, 11 (1):89–105, 2005.