Third 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

3rd December 2014, Maths Room 100

1 Introduction

In our last monitoring committee the Data Assimilation Linked Ecosystem Carbon model (DALEC) [8] was introduced and we showed how this model could be implemented in a variational data assimilation (DA) framework in order to look at the information content of different forest carbon balance observations. In particular, Shannon Information Content [6] was used to investigate the level of information provided by single and successive observations of Net Ecosystem Exchange (NEE) and it was shown that observations of NEE taken in summer with warmer temperatures and higher daily irradiances had a higher information content than those taken in winter with lower temperatures and lower daily irradiance. This makes 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. The error in our linearized version of DALEC was also considered in the last report, this is important as in four-dimensional variational data assimilation (4D-Var) the linearized model is used in the tangent linear hypothesis, the results from this experiment showed the linearized DALEC to perform poorly. However since the last report I have found that the linear DALEC was not derived correctly and repeating these experiments shows the tangent linear hypothesis to be a good approximation for DALEC.

Since our last meeting I have continued the work on the information content of forest carbon balance observations. I have looked at the degrees of freedom for signal as another measure for information content and have written a first draft of a report on my findings (I have included this alongside my monitoring report). I have built DALEC into a 4D-Var scheme assimilating data from a young pine stand in Oregon in order to estimate the five carbon pools which make up the state vector for DALEC. I have started visiting the Forest Research site at Alice Holt on a weekly basis and am helping with current measurement campaigns. A new version of the DALEC model has been released since our last meeting (DALECV2 [1]) which can be parameterized for both evergreen and deciduous forests, I have started working with this model now as the research site at Alice Holt is largely deciduous and the previous DALEC with which I have been working only models evergreen forests. Currently, I have coded my own version of DALECV2 in Python and have built this into a 4D-Var scheme similar to that which I had written for the state estimation with DALEC but now estimating the 17 model parameters as well as the initial value for the carbon pools. I hope this will allow me to parameterize DALECV2 for the Alce Holt research site. In this report we begin by showing some of the results from the information content experiments.

2 Information Content

In the last monitoring report the Shannon Information Content (SIC) was used as a measure of the reduction in entropy of a system given a set of observations [7]. The SIC can be derived as,

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

where $\left|\mathbf{B}\right|$ is the determinant of the background error covariance matrix and $\left|\mathbf{A}\right|$ is the determinant of the analysis covariance matrix. The NEE of a forest in the DALEC model is expressed as the difference between the Gross Primary Production (GPP) of the forest and the respirations of the forest at time t,

$$\text{NEE}(t) = -(1-p_2)\text{GPP}(C_f(t),\phi) + p_8 C_l T(t) + p_9 C_s T(t), \tag{2}$$

where p_2, p_8, p_9 are model parameters, C_f, C_l, C_s are carbon pools, ϕ represents the meteorological driving data used by the GPP function and $T = \frac{1}{2}\exp(\Theta T_{\text{mean}}(t))$ is an exponential function of the mean daily temperature. Combining equations 1 and 2 and using the 4D-Var framework for DALEC we can derive the SIC for an observation of NEE at time t as,

$$\text{SIC for NEE}(t) = \frac{1}{2}\ln\frac{(p_2-1)^2\text{GPP}'^2(C_f(t),\phi)\sigma^2_{cf,b} + \sigma^2_{nee,o} + T^2(t)(p_9^2\sigma^2_{cs,b} + p_8^2\sigma^2_{cl,b})}{\sigma^2_{nee,o}}. \tag{3}$$

where $\sigma^2_{cf,b}, \sigma^2_{cl,b}, \sigma^2_{cs,b}$ are the background variances for the carbon pools and $\sigma^2_{nee,o}$ is the variance for an observation of NEE. If we assume that the variances and parameters here are fixed, we can see that the size of the SIC is dependent on the temperature term, $T(t)$, and the square of the first derivative of GPP. Generally, the value of GPP (and its first derivative) is highest in summer with higher temperatures and higher total daily irradiance. We therefore have that there will be more information content in observations that are taken when temperatures are higher. Physically this makes sense as more NEE takes place when temperatures are higher (to a point), so measurements are of greater magnitude and give us more information about carbon fluxes.

By plotting the SIC for a single observation of NEE, varying with three years of meteorological driving data, next to the temperature term ($T(t)$) for the same data we can see that both are closely linked, figure 1. This shows that observations of NEE made in the summer are much more valuable

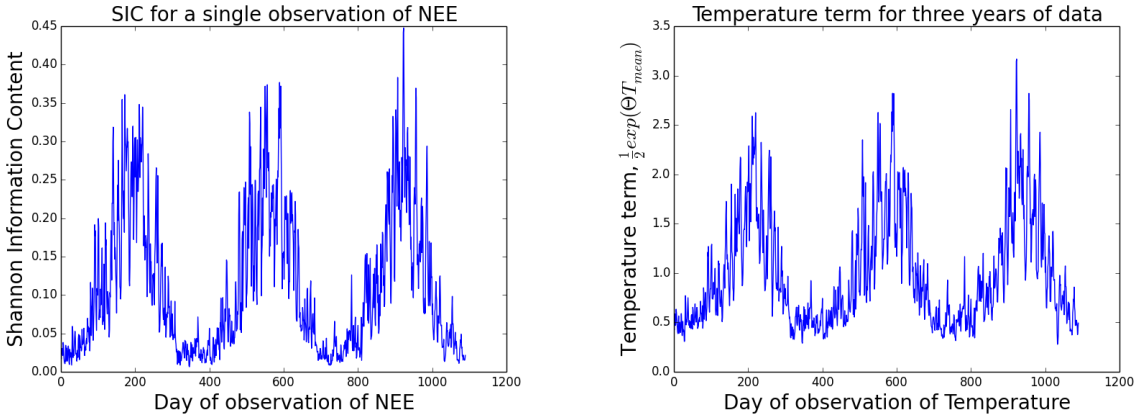

Figure 1: SIC and temperature varying over three years using driving data from Oregon pine forest.

than those made in the winter assuming warmer temperatures, higher daily irradiance and a higher amount of foliar carbon in the summer.

In the report this idea is also explored for the degrees of freedom for signal (DFS) measure for information content. The DFS indicates the number of elements of the state that have been measured by the observations. If we consider a state vector $\mathbf{x}$ with n elements (or n degrees of freedom) then the maximum value the DFS could obtain would be n, in this case all elements of the state would have been measured. Conversely if DFS $= 0$ then no elements of the state would have been measured by our observations [2].The DFS can be written as,

$$\text{DFS} = n - \text{trace}(\mathbf{B}^{-1}\mathbf{A}). \tag{4}$$

The result in figure 1 can be reproduced using the DFS showing that we measure more elements of our state with observations of NEE made in summer than observations made in winter. To understand how much more value the summer observations of NEE have compared to the winter observations we have plotted the increasing SIC and DFS for 50 observations made from December 21^{st} onwards and also the the constant line of SIC and DFS for one observation of NEE made in summer at a mean daily temperature of 26^{o}C on July 12^{th}. This is shown in figure 2.

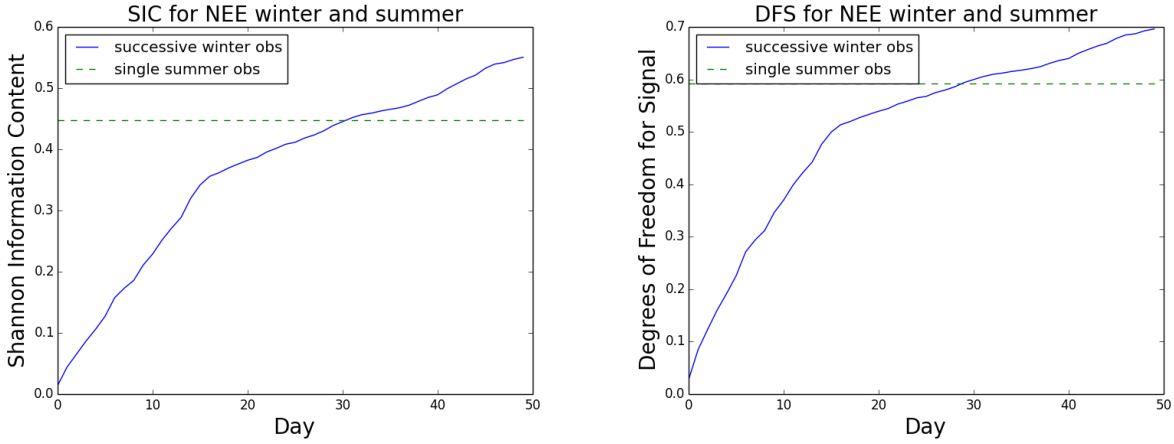

Figure 2: SIC (left) and DFS (right) for a single summer observation of NEE and successive winter observations of NEE using driving data from Oregon pine forest.

Figure 2 shows us that, for the driver data used in this experiment, we require 31 days of NEE observations to gain the same reduction in entropy for our analysis as 1 observation made in the summer and 30 days of NEE observations to achieve the same DFS as 1 observation made in the summer. For more detail on the information content experiments with DALEC please see the attached report [5].

3 4D-Var with DALEC

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 [3]. The cost function is given as,

$$J(\mathbf{x}_0) = \frac{1}{2}(\mathbf{x}_0 - \mathbf{x}_b)^T \mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}_b) + \frac{1}{2}\sum_{i=0}^{N}(\mathbf{y}_i - h_i(\mathbf{x}_i))^T \mathbf{R}_i^{-1}(\mathbf{y}_i - h_i(\mathbf{x}_i)), \tag{5}$$

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 ($m_{0\to i}(\mathbf{x}_0) = \mathbf{x}_i$) to the observations at this time $\mathbf{y}_i$ and $\mathbf{R}_i$ is the observation error covariance matrix 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,

$$\nabla J(\mathbf{x}_0) = \mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}_b) - \sum_{i=0}^{N} \mathbf{M}_{i,0}^T \mathbf{H}_i^T \mathbf{R}_i^{-1}(\mathbf{y}_i - h_i(\mathbf{x}_i)), \tag{6}$$

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 $\mathbf{M}_i = \frac{\delta m_i(\mathbf{x}_i)}{\delta \mathbf{x}_i}$.

For DALEC our state $\mathbf{x}_0$ corresponds to the initial values of the five carbon pools, $\mathbf{x}_0 = (C_f(t_0), C_r(t_0), C_w(t_0), C_l(t_0), C_s(t_0))^T$. We can calculate the linearized model for DALEC as,

$$\mathbf{M}_i = \begin{pmatrix} (1-p_5) + p_3(1-p_2)\zeta_i & 0 & 0 & 0 & 0 \\ p_4(1-p_3)(1-p_2)\zeta_i & (1-p_7) & 0 & 0 & 0 \\ (1-p_4)(1-p_3)(1-p_2)\zeta_i & 0 & (1-p_6) & 0 & 0 \\ p_5 & p_7 & 0 & (1-(p_1+p_8)T_i) & 0 \\ 0 & 0 & p_6 & p_1 T_i & (1-p_9 T_i) \end{pmatrix}, \tag{7}$$

where $\zeta_i = \mathrm{GPP}'(C_f(t_i), \phi)$ and $T_i = T(t_i)$. The cost function and gradient of the cost function were coded in Python along with the DALEC model. Meteorological driving data and observations for assimilation were initially taken from a young pine forest in Oregon. Our analysis ($\mathbf{x}_a$) was found by passing our cost function and its gradient function along with $\mathbf{x}_b$ to the Broyden-Fletcher-Goldfarb-Shanno minimization algorithm (BFGS) [4] found in the scipy.optimize package for Python. The results for our analysis when assimilating a years worth of NEE observations can be seen in figure 3.

In figure 3 we can see a better fit to the observations after assimilation of the NEE observations when our $\mathbf{x}_a$ is run forward in figure 3. Assimilating the NEE data not only improves the fit of the model trajectory to the observations of NEE but also to the observations of GPP and total respiration (RT). This shows that by assimilating NEE observations we are also gaining information about GPP and RT. This is expected as NEE is a product of RT and GPP (NEE = RT - GPP) so improving our estimate of NEE should also improve our estimate of RT and GPP. We can also see how simple the DALEC model is in some areas by looking at its prediction of litter fall (LF) which is just modelled as a constant multiplied by C_f in this evergreen version of DALEC.

More experiments had been planned with DALEC in a 4D-Var scheme however with the release of DALECV2 [1] it was decided that it would be best to begin working with this new model. This new model DALECV2 can be paramterized for either deciduous or evergreen forests where as the DALEC being used for previous experiments only models evergreen forests. Paramterizing the model for a deciduous forest is important as the Forest Research site at Alice Holt forest is mainly decidous and this is where I will be based for my field work.

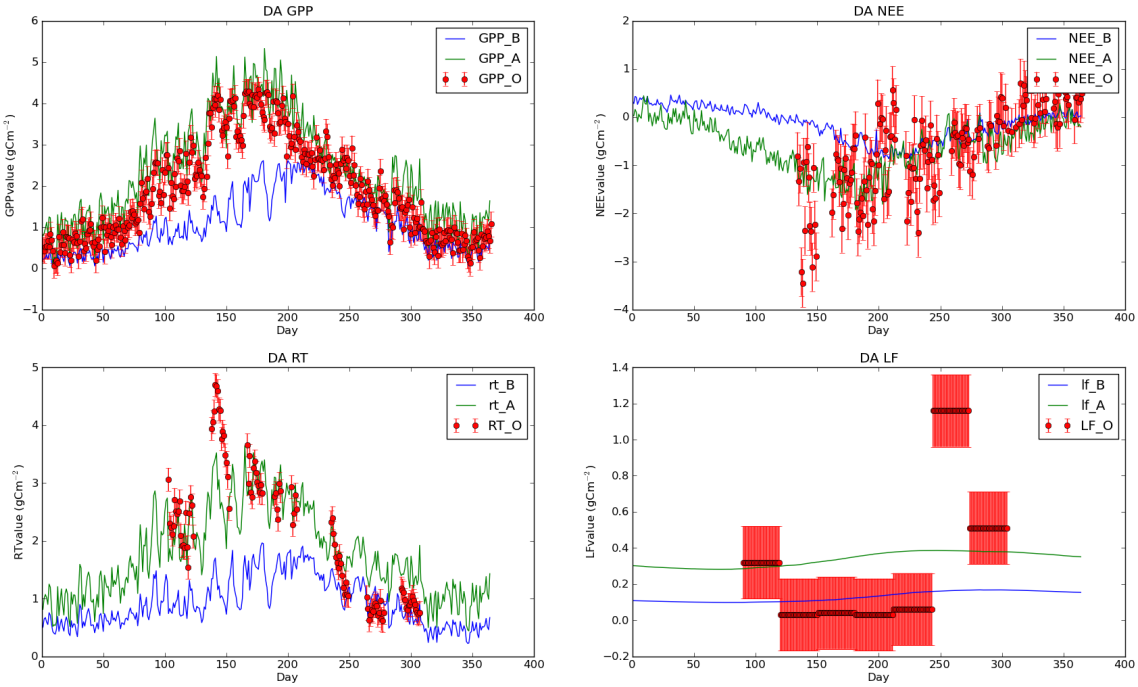

Figure 3: 4D-Var with DALEC assimilting observations of NEE over a year from Oregon pine stand. Here the blue line is our $\mathbf{x}_b$ run forward, the green line is our $\mathbf{x}_a$ run forward and the red dots are diferent forest carbon balance observations corresponding to GPP (top left), NEE (top right), total respiration (RT) (bottom left) and litter fall (LF) (bottom right).

4 Current Work

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 parameterized for the Alice Holt research site. I have started to work with meteorological driving data and forest carbon balance observations from Alice Holt and am converting this data into the correct format to be used with the DALECV2 model.

I have recently started working at Forest Research once a week on Tuesdays. I am currently helping out conducting stem and soil respiration measurements, the equipment used for the stem respiration measurements can be seen in figure 4. These measurements form part of an existing measurement campaign that Forest Research are conducting. The measurements of stem respiration may not be directly used in my own work, however they are valuable for my understanding of conducting fieldwork measurements and learning my way around the Straits enclosure where the bulk of the carbon balance measurements are made at the Alice Holt site.

Figure 4: Tree respiration chamber used at Alice Holt research site.

5 Future Plans

I shall continue to work on parameterizing DALECV2 for the Alice Holt research site. I also plan on repeating the information content experiments with DALECV2 to see if we have similar results as achieved with the original DALEC model.

6 Professional and Academic Development

6.1 Masters Courses

Since our last meeting I have now received the grades for my final three assessed modules:

- 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:

- 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 - 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 - Git and Python
- 10/07/2014-11/07/2014 - Forest Research - Helped with field work LiDAR
- 29/09/2014-03/10/2014 - NERC course - Software Developement for Environmental Scientists
- 08/10/2014-10/10/2014 - Environment YES - NERC competition at Syngenta, Jesops Hill

I have also demonstrated on the NERC 'Data assimilation and visualization for environmental sciences' course held from 15/09/2014 to 19/09/2014 organised by Dr. Amos Lawless.

On 17/12/14 I will be giving a presentation at the Mathematics for Planet Earth industry day at Imperial College about my project and working with a case partner.

6.3 Summer School

I attended the Fluxcourse summer school held at the University of Colorado from 21/07/2014 to 01/08/2014. During the course we stayed at the University's mountain research station near Boulder, Colorado. The course covered many different aspects of measuring and modelling the CO_2 flux of forests and plants. I found spending time with other early career scientists studying similar problems very helpful in the understanding of my own research topic. At the end of the course we had to present a piece of group work using a carbon balance model with a data assimilation scheme.

References

[1] A. Bloom and M. 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, August 2014.

[2] Alison Fowler and Peter Jan van Leeuwen. Measures of observation impact in Gaussian data assimilation. 2011.

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

[4] Jorge Nocedal and Stephen J Wright. *Numerical Optimization*. Springer Science & Business Media, 1999.

[5] Ewan Pinnington. Information content for observations of forest carbon stocks and fluxes when assimilated with the DALEC carbon balance model. 2014.

[6] Clive D Rodgers and Others. *Inverse methods for atmospheric sounding: Theory and practice*, volume 2. World scientific Singapore, 2000.

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

[8] 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.