Second 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

3$^{\text{rd}}$ June 2014, Room 1L36

Project Background

A large amount of data is currently being gathered that is relevant to the carbon balance of forests, with much of this data coming from Eddy covariance flux towers [1]. Attempts are also being made to combine this data with models of forest carbon stocks and fluxes, such as the Data Assimilation Linked Ecosystem Carbon model (DALEC) [10], in a data assimilation scheme. Currently, however, there are limitations with such schemes as there is a lack of understanding about the additional information provided by different observations. There is also a lack of understanding about observation error correlations. Current data assimilation (DA) schemes for ecosystem carbon flux only specify the diagonal elements of the observation error covariance matrix, $\mathbf{R}$, which correspond to the individual uncertainties in particular observations. As such, these DA schemes do not specify observation error correlations or covariances, corresponding to the off diagonal elements of $\mathbf{R}$. In numerical weather prediction (NWP) it has been shown that the inclusion of observation error correlations can increase the information content from a given data set and reduce analysis error [7, 8, 9]. Better understanding of the information content and error correlations of carbon balance observations will form two of the main aims of the project. We begin by introducing the DALEC model which will initially be used to look at the information content in different observations.

The DALEC model

The DALEC model is a simple process-based model describing the carbon balance of an evergreen forest ecosystem [10]. The model is constructed of five carbon pools (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 gross primary production function (GPP) uses meteorological driving data and the site's leaf area index (a function of C_f) to calculate the total amount of carbon to be allocated at a daily time step. The model equations for the carbon pools at day $t+1$ are as follows:

$$C_f(t+1) = (1-p_5)C_f(t) + p_3(1-p_2)GPP(C_f(t),\phi), \tag{1}$$
$$C_r(t+1) = (1-p_7)C_r(t) + p_4(1-p_3)(1-p_2)GPP(C_f(t),\phi), \tag{2}$$
$$C_w(t+1) = (1-p_6)C_w(t) + (1-p_4)(1-p_3)(1-p_2)GPP(C_f(t),\phi), \tag{3}$$
$$C_l(t+1) = (1-(p_1+p_8)T(t))C_l(t) + p_5C_f(t) + p_7C_r(t), \tag{4}$$
$$C_s(t+1) = (1-p_9T(t))C_s + p_6C_w(t) + p_1T(t)C_l(t), \tag{5}$$

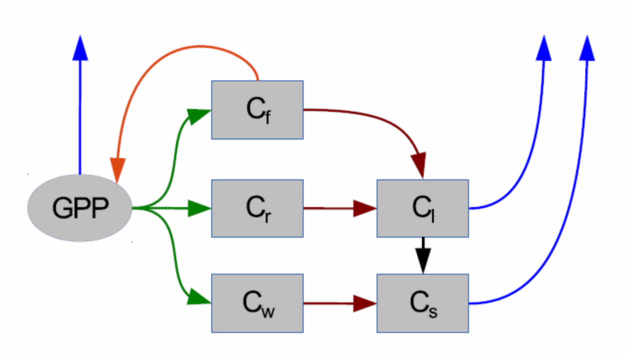

Figure 1: DALEC carbon balance model [3]

where $T(t) = \frac{1}{2}exp(p_{10}T_m(t))$, T_m is daily mean temperature, $p_1, \ldots, p_{10}$ are rate parameters and ϕ represents the meteorological driving data used in the GPP function. The full details of this version of DALEC can be found in [10]. We now introduce Shannon Information Content as one method to assess the information content in different carbon balance observations.

Shannon Information Content and Observation Sensitivity

In DA Shannon Information Content (SIC) is a measure of the reduction in entropy given a set of observations. Entropy physically corresponds to the volume in state space taken up by the probability density function (pdf) describing the knowledge of the state [6]. Assuming all pdfs are Gaussian we have,

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

where $\mathbf{B}$ is the background error covariance matrix and $\mathbf{A}$ is the analysis error covariance matrix. For a larger reduction in uncertainty in our analysis we have a larger value of SIC. I began by using SIC to understand the information content for different sets of observations at one time when being assimilated with the DALEC model. We specify the state vector for the assimilation as,

$$\underline{x}_b = (C_f, C_r, C_w, C_l, C_s)^T,$$

where the elements of the state vector have variances, $\sigma^2_{cf,b}, \ldots, \sigma^2_{cs,b}$, respectively. We then have the following background error covariance matrix,

$$\mathbf{B} = \begin{pmatrix} \sigma^2_{cf,b} & 0 & 0 & 0 & 0 \\ 0 & \sigma^2_{cr,b} & 0 & 0 & 0 \\ 0 & 0 & \sigma^2_{cw,b} & 0 & 0 \\ 0 & 0 & 0 & \sigma^2_{cl,b} & 0 \\ 0 & 0 & 0 & 0 & \sigma^2_{cs,b} \end{pmatrix},$$

here we assume a diagonal background error covariance matrix. In order to calculate the SIC we need $\left|\mathbf{A}^{-1}\right|$ we have,

$$\mathbf{A}^{-1} = \mathbf{J}'' = \mathbf{B}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H},$$

where $\mathbf{J}''$ is the Hessian of $\mathbf{J}$, the cost function to be minimized in Three-Dimensional Variational Data Assimilation (3D-Var), and $\mathbf{H}$ is the linearized observation operator. One of the main observations made of the carbon balance of a forest at flux tower sites is the net ecosystem exchange (NEE)

of CO$_2$, which can be estimated by DALEC as the difference between GPP and the respiration of C_l and C_s, giving,

$$NEE(t) = -(1-p_2)GPP(C_f(t), \phi) + p_8 C_l T(t) + p_9 C_s T(t).$$

For a single observation of NEE at one time, t_0, an analytical expression for the SIC can be derived using,

$$\mathbf{H}_0 = \begin{pmatrix} -(1-p_2)\zeta_0 & 0 & 0 & p_8 T_0 & p_9 T_0 \end{pmatrix},$$

where $\zeta_0 = GPP'(C_f(t_0), \phi)$, $T_0 = T(t_0)$ and $\mathbf{H}_0 = \frac{\delta NEE(t_0)}{\delta \underline{x}}$ is the linearized observation operator at time t_0. As we have a single observation at one time our observation error covariance matrix, $\mathbf{R}$, is just the variance of our observation of NEE, $\sigma_{nee,0}^2$, at time t_0. Therefore,

$$\mathbf{R} = \sigma_{nee,0}^2$$

and

$$\begin{aligned} \mathbf{J}'' &= \mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} \\ &= \begin{pmatrix} \sigma_{cf,b}^{-2} + \sigma_{nee,0}^{-2}(1-p_2)^2\zeta_0^2 & 0 & 0 & \sigma_{nee,0}^{-2}(1-p_2)\zeta_0 p_8 T_0 & \sigma_{nee,0}^{-2}(1-p_2)\zeta_0 p_9 T_0 \\ 0 & \sigma_{cr,b}^{-2} & 0 & 0 & 0 \\ 0 & 0 & \sigma_{cw,b}^{-2} & 0 & 0 \\ \sigma_{nee,0}^{-2}(1-p_2)\zeta_0 p_8 T_0 & 0 & 0 & \sigma_{cl,b}^{-2} + \sigma_{nee,0}^{-2} p_8^2 T_0^2 & \sigma_{nee,0}^{-2} p_8 p_9 T_0^2 \\ \sigma_{nee,0}^{-2}(1-p_2)\zeta_0 p_9 T_0 & 0 & 0 & \sigma_{nee,0}^{-2} p_8 p_9 T_0^2 & \sigma_{cs,b}^{-2} + \sigma_{nee,0}^{-2} p_9^2 T_0^2 \end{pmatrix}. \end{aligned}$$

We then have,

$$SIC = \frac{1}{2} ln \frac{|\mathbf{B}|}{|\mathbf{A}|} = \frac{1}{2} ln \, |\mathbf{B}| \, |\mathbf{J}''| \, .$$

Hence,

$$SIC = \frac{1}{2} ln \frac{(p_2-1)^2 \zeta_0^2 \sigma_{cf,b}^2 + \sigma_{nee,0}^2 + T_0^2 (p_9^2 \sigma_{cs,b}^2 + p_8^2 \sigma_{cl,b}^2)}{\sigma_{nee,0}^2}.$$

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_0, and the square of the first derivative of GPP, ζ_0^2. Generally, the value of GPP (and its first derivative) is highest in summer with higher total daily irradiance and higher temperatures. 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 of carbon fluxes**. I have also derived analytical forms for the SIC using different sets of observations at a single time, which all have a similar form.

Following this work, based at a single time, I started looking at the SIC when successive observations are added over a period of time. The model was now built into a Four-Dimensional Variational Data Assimilation (4D-Var) framework where our observation operator, $\mathbf{H}$, and observation error covariance matrix, $\mathbf{R}$, are now,

$$\mathbf{H} = \begin{pmatrix} \mathbf{H}_0 \\ \mathbf{H}_1 \mathbf{M}_0 \\ \vdots \\ \mathbf{H}_n \mathbf{M}_{n,0} \end{pmatrix} \quad \text{and} \quad \mathbf{R} = \begin{pmatrix} \mathbf{R}_0 & 0 & 0 & 0 \\ 0 & \mathbf{R}_1 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \mathbf{R}_n \end{pmatrix},$$

where $\mathbf{H}_i$ is our linearized observation operator at time t_i, $\mathbf{M}_{i,0} = \mathbf{M}_{i-1}\mathbf{M}_{i-2}\cdots\mathbf{M}_0$ is our linearized model evolving the state vector, $\underline{x}_b$, at time t_0 to time t_i and $\mathbf{R}_i$ is the observation error covariance matrix corresponding to $\mathbf{H}_i$ at time t_i [5]. I first calculated the adjoint model for DALEC analytically as $\mathbf{M}_i = \frac{\delta \underline{m}_i}{\delta \underline{x}_i}$. I then wrote a code in Python that calculates $\mathbf{H}$ and $\mathbf{R}$ to find a value of SIC when successive observations of NEE are made each day over a chosen period (To calculate the SIC we do not need the actual observation value). The meteorological driving data used in the model is taken from a Ponderosa pine forest in central Oregon for which the DALEC model in [10] is parameterized. Below we have a plot when different starting points and periods are chosen, day 0 represents the start of the year so that day 200 is the 19th of July.

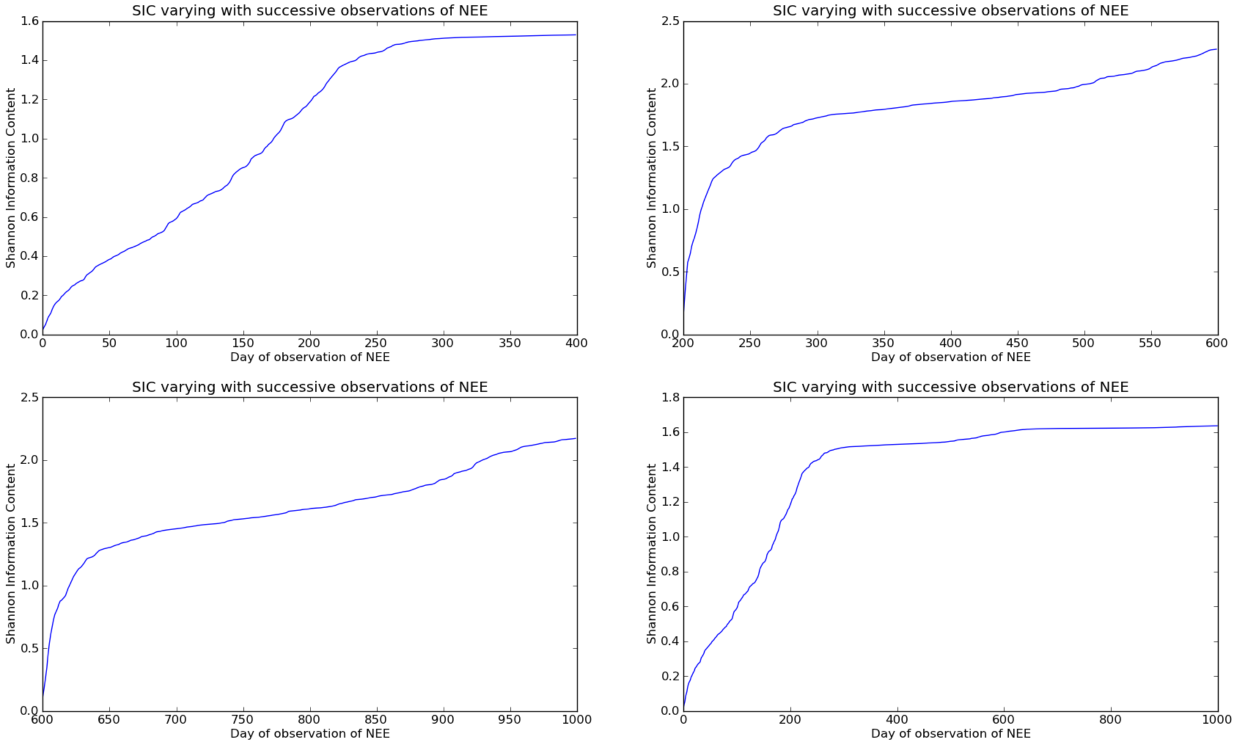

Figure 2: SIC varying as successive observations of NEE are added using driving data from Oregon pine forest.

When starting at day 0 we can see a leveling off in the SIC after 200 days of observations when the forest in Oregon will be experiencing some of the highest temperatures of the year. When starting our observations from day 200 we see that the SIC levels off much quicker and achieves a higher value than the first period. This agrees with the analytic form of SIC found previously for a single observation that is a function of temperature. The reason for the leveling off of the SIC when yearly peak temperatures are reached needs to be investigated further. **As temperatures decrease very little information is added to the system, see graph of single observations of NEE as opposed to successive observations of NEE (make this graph if this hasnt been done!).**

Error in the Linear Model

Redo this with correct GPP' It is important to understand how well our linearized model performs. In 4D-Var we use our linearized model in the approximation,

$$m(\underline{x} + \delta\underline{x}) \approx m(\underline{x}) + \mathbf{M}\delta\underline{x},$$

where $m(\underline{x})$ is our non-linear model applied to the state $\underline{x}$. To see if this approximation is reasonable we can plot,

$$\delta\underline{x}_N = m(\underline{x} + \delta\underline{x}) - m(\underline{x}) \quad \text{and} \quad \delta\underline{x}_L = \mathbf{M}\delta\underline{x}$$

using the initial conditions for the carbon pools given in [10] as our state, $\underline{x}$, and a value of 10% of each pool for our perturbation $\delta\underline{x}$. If we then run our linear and nonlinear models forward over 1000 days using these initial conditions and plot $\delta\underline{x}_N$ and $\delta\underline{x}_L$ we will be able to see how well our linear DALEC model performs. Here then we see that although the linear model is a very good approximation for C_s the other pools diverge considerably by day 1000, with approximately a 70% error in C_f and a 15% error in C_w. This shows that past day 100 our linear approximation for the DALEC model does not perform well.

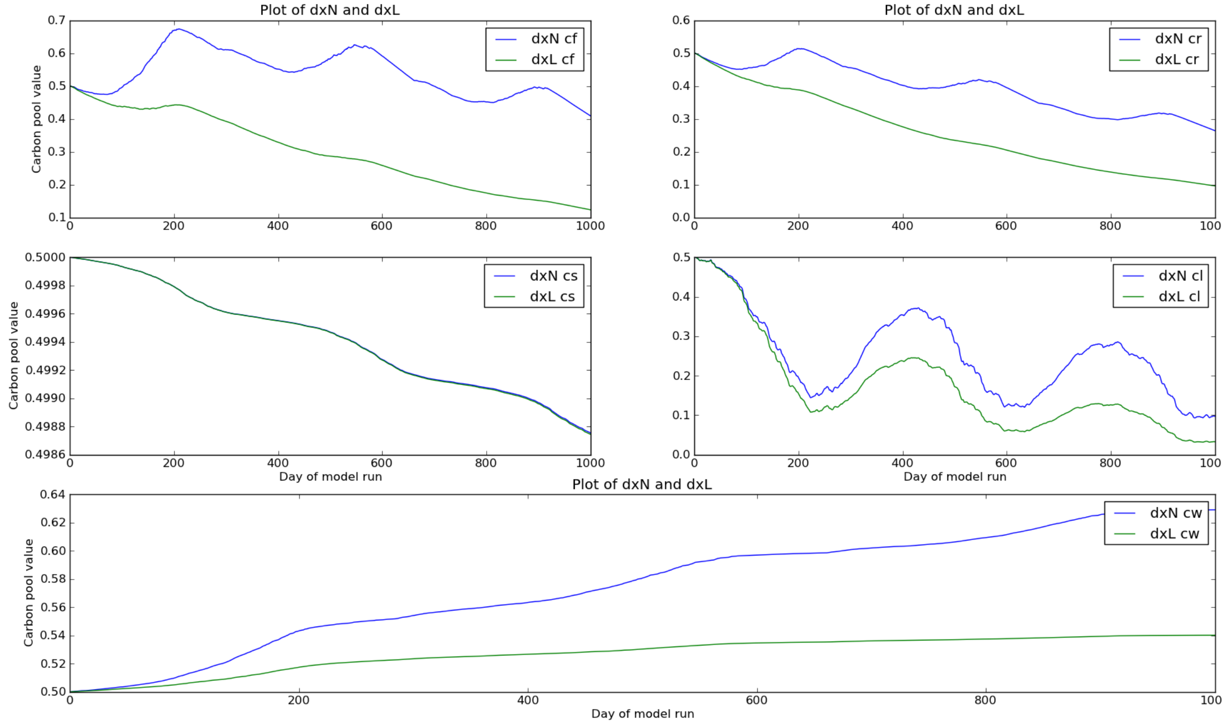

Figure 3: Plot of $\delta\underline{x}_N$ and $\delta\underline{x}_L$ for each of the carbon pools in the state $\underline{x}$

Future Plans

My plans for the future are as follows:

- I will continue to investigate the reason behind the leveling off in SIC for successive observations of NEE and its relation to temperature. When this is understood I will start investigating the SIC for different sets of observations over a period of time to understand if there is an optimal set of observations or an optimal time to make a set of observations for the DALEC model. I will also begin looking at the 'degrees of freedom for signal' [6] as another measure of information content.

- I have begun building DALEC into an operational DA scheme using the meteorological driving data from the forest in Oregon. The data also contains observations of forest carbon balance to be assimilated with the DALEC model. Once I have a fully working DA scheme for DALEC I shall start using the Desroziers diagnostics in an attempt to find a correlated $\mathbf{R}$ [4]. Once we have a better specified $\mathbf{R}$ I will repeat the information content experiments to see what effect using this new $\mathbf{R}$ has. I will also undertake more reading in the area of correlated observation errors and hopefully find a way to build some temporally correlated errors into $\mathbf{R}$ as opposed to just correlations between observations taken at the same time.
- In the next academic year I will begin spending time at the Forest Research site Alice Holt, this will allow me to see how different forest carbon balance observations are actually taken. I will also develop an understanding of the constraints on when and how many observations can be made so that I do not conduct unrealistic information content experiments.
- I will start reading up on other forest carbon balance models and possibly begin working with the the Simplified Photosynthesis and Evapo-Transpiration model (SiPNET) [2]. I will then try to apply the techniques I will have built up with DALEC to these more complicated models.

Professional and Academic Development

Masters Courses

I have now completed the seven masters courses in which I was enrolled at the beginning of the first term. In my first term module I received the following marks:

- MAMB10 (Data Assimilation) - 85%
- MAMNSO (Numerical Solutions to Ordinary Differential Equations) - 79%
- MTMG02 (Atmospheric Physics) - 66%

Computing techniques and projects was taken as just a formative module with no assessment. I have taken all the exams for my three second term modules (MTMG49 - Boundary Layer, MTMD01 - Environmental Data Visualization and Exploration, MTMD02 - Operational Data Assimilation) and am awaiting the results.

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

In the coming months I will be attending:

- 29/05/2014 - How to Write a Paper - RRDP
- 21/07/2014-01/08/2014 - Fluxcourse - University of Colorado
- 25/06/2014-26/06/2014 - Software Carpentry - Git and Python
- 10/07/2014-11/07/2014 - Forest Research - Helped with field work LiDAR

The Fluxcourse is held in Boulder Colorado and I hope it will help me see the problems in my project from a different view point as I have come from a mathematics degree and the course is taught by some of the leading scientists in the field of forest carbon balance.

References

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

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

[3] Sylvain Delahaies, Ian Roulstone, and Nancy Nichols. A regularization of the carbon cycle data-fusion problem. In *EGU General Assembly Conference Abstracts*, volume 15, page 4087, 2013.

[4] Gérald Desroziers, Loic Berre, Bernard Chapnik, and Paul Poli. Diagnosis of observation, background and analysis-error statistics in observation space. *Quarterly Journal of the Royal Meteorological Society*, 131(613):3385–3396, 2005.

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

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

[7] Laura M Stewart, Sarah Dance, and Nancy K Nichols. Data assimilation with correlated observation errors: experiments with a 1-d shallow water model. *Tellus A*, 65, 2013.

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

[9] LM Stewart, J Cameron, SL Dance, S English, J Eyre, and NK Nichols. Observation error correlations in iasi radiance data. *Mathematics report series*, 1, 2009.

[10] 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.