# Second Monitoring Committee Meeting

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

E. Pinnington

$16^{th}$ May 2014, Room TBC

## 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) [6], in a data assimilation (DA) scheme. Currently, however, there are limitations with such schemes as there is a lack of understanding about the additional information provided by different observations. Current 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 [5]. Better understanding the information content and error correlations of carbon balance observations will form one of the main aims of the project.

## The DALEC model

The DALEC model is a simple process-based model describing the carbon balance of an evergreen forest ecosystem. 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:

$$\begin{aligned}
C_f(t+1) &= (1-p_5)C_f(t) + p_3(1-p_2)GPP(C_f(t),\phi),\\
C_r(t+1) &= (1-p_7)C_r(t) + p_4(1-p_3)(1-p_2)GPP(C_f(t),\phi),\\
C_w(t+1) &= (1-p_6)C_w(t) + (1-p_4)(1-p_3)(1-p_2)GPP(C_f(t),\phi),\\
C_l(t+1) &= (1-(p_1+p_8)T(t))C_l(t) + p_5C_f(t) + p_7C_r(t),\\
C_s(t+1) &= (1-p_9T(t))C_s + p_6C_w(t) + p_1T(t)C_l(t).
\end{aligned}$$

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 $\phi$ represents the meteorological driving data used in the $GPP$ function. The full details of this version of DALEC can be found in [6].

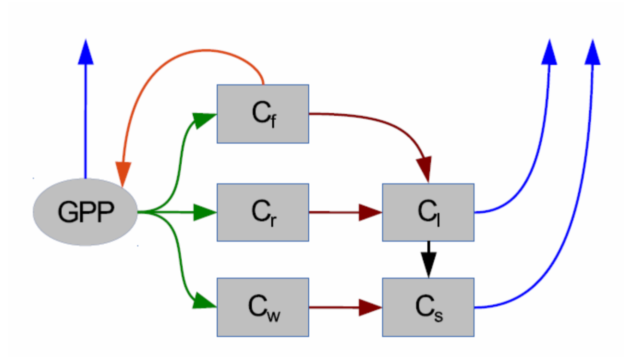


Figure 1: DALEC carbon balance model [2]

## 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 [4]. Assuming all pdfs are Gaussian we have,

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

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, specifying 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. Giving a background error covariance,

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

Here I have assumed a diagonal background error covariance matrix. For $\mathbf{A}^{-1}$ 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 forest carbon balance 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's 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,

$$\mathbf{H}_0 = \begin{pmatrix}(1-p_2)\zeta_0 & 0 & 0 & p_8T_0 & p_9T_0\end{pmatrix},$$

where $\zeta_0 = GPP'(C_f(t_0), \phi)$, $T_0 = T(t_0)$ and $\mathbf{H}_0$ 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 $NEE$, $\sigma^2_{nee,0}$, at time $t_0$. So that,

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

and

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

We then have,

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

So that,

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

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 $GPP$, $\zeta_0^2$. 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. 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 $t_i$ [3]. I first calculated the adjoint model for DALEC analytically as $\mathbf{M}_i = \frac{\delta m_i}{\delta \underline{x}_i}$. 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 [6] 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 during the summer,

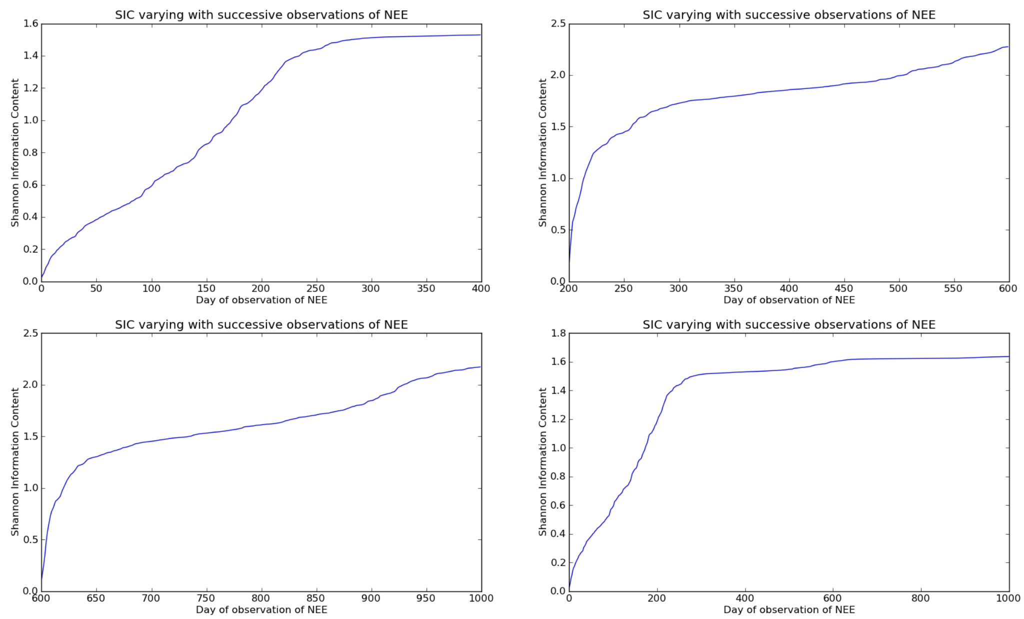


Figure 2: $SIC$ varying as successive observations of $NEE$ are added using driving data from Oregon pine forest.

# 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] 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.

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

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

[5] 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.

[6] 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.