Information content for observations of forest carbon stocks and fluxes when assimilated with the DALEC carbon balance model DRAFT

E. Pinnington

March 3, 2016

1 Introduction

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, 2, 10, 13]. 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), in a data assimilation scheme [7, 15]. Currently, however, there are limitations with such schemes as there is a lack of understanding about the additional information provided by different observations. Better understanding of the information content of carbon balance observations will help inform measurement campaigns of when and which observations to take, in order to gain the most information possible about the system. In this report, we will look at different information content measures which have been used in meteorological data assimilation [5, 9, 11] and apply these to carbon balance observations assimilated with DALEC. Although we use DALEC in this report the results will be similar for other carbon balance models which use similar driving data and equations. We begin by introducing the DALEC model which will initially be used to look at the information content in different observations.

2 The DALEC Model

The DALEC model is a simple process-based model describing the carbon balance of an evergreen forest ecosystem [15]. 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:

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

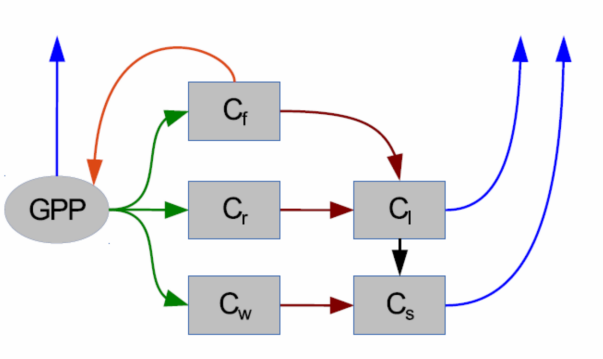


Figure 1: Representation of the carbon fluxes in the DALEC carbon balance model. Green arrows represent C allocation, dark red and black arrows represent litterfall and decomposition fluxes, blue arrows represent respiration fluxes and the light red arrow represents the feedback of foliar carbon to the GPP function. [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 [15]. It is parameterized for data from a young pine stand in Ponderossa, Oregon. The model parameters and the equations used to calculate GPP are included in the appendix. We now see how DALEC can be implemented in a four-dimensional variational data assimilation (4D-Var) framework.

2.1 DALEC in a variational assimilation framework

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. For DALEC our state $\mathbf{x}_0$ corresponds to the initial values of the five carbon pools, $\mathbf{x}_0 = (C_f(t_0), C_r(t_0), C_w(t_0), C_l(t_0), C_s(t_0))^T$. $P(\mathbf{x}_0|\mathbf{y})$ is maximised by minimising a cost function $J(\mathbf{x})$ derived from Baye's Theorem [8]. The cost function is given as,

$$J(\mathbf{x}_0) = \frac{1}{2}(\mathbf{x}_0 - \mathbf{x}_b)^T\mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}_b) + \frac{1}{2}\sum_{i=0}^{N}(\mathbf{y}_i - h_i(\mathbf{x}_i))^T\mathbf{R}_i^{-1}(\mathbf{y}_i - h_i(\mathbf{x}_i)), \tag{6}$$

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,

$$\nabla J(\mathbf{x}_0) = \mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}_b) - \sum_{i=0}^{N}\mathbf{M}_{i,0}^T\mathbf{H}_i^T\mathbf{R}_i^{-1}(\mathbf{y}_i - h_i(\mathbf{x}_i)), \tag{7}$$

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 calculate the linearized model for DALEC from

equations 1 to 5 as,

$$\mathbf{M}_i = \begin{pmatrix} (1-p_5)+p_3(1-p_2)\zeta_i & 0 & 0 & 0 & 0 \\ p_4(1-p_3)(1-p_2)\zeta_i & (1-p_7) & 0 & 0 & 0 \\ (1-p_4)(1-p_3)(1-p_2)\zeta_i & 0 & (1-p_6) & 0 & 0 \\ p_5 & p_7 & 0 & (1-(p_1+p_8)T_i) & 0 \\ 0 & 0 & p_6 & p_1 T_i & (1-p_9 T_i) \end{pmatrix}, \tag{8}$$

where $\zeta_i = GPP'(C_f(t_i), \phi)$ and $T_i = T(t_i)$. In 4D-Var we assume the tangent linear hypothesis,

$$m_{0\to i}(\mathbf{x}_0 + \delta\mathbf{x}_0) \approx m_{0\to i}(\mathbf{x}_0) + \mathbf{M}_{i,0}\delta\mathbf{x}_0. \tag{9}$$

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 DALEC by taking our initial states from the appendix for $\mathbf{x}_0$ and a 5% perturbation for $\delta\mathbf{x}_0$. We then rearrange equation 9 to find,

$$\text{Percentage error in tangent linear model} = \left| \frac{m_{0\to i}(\mathbf{x}_0+\delta\mathbf{x}_0) - m_{0\to i}(\mathbf{x}_0)}{\mathbf{M}_{i,0}\delta\mathbf{x}_0} - 1 \right| \times 100. \tag{10}$$

The percentage error in our approximation at each time step for each carbon pool when the model is run forward for 1000 days is shown in figure 2.

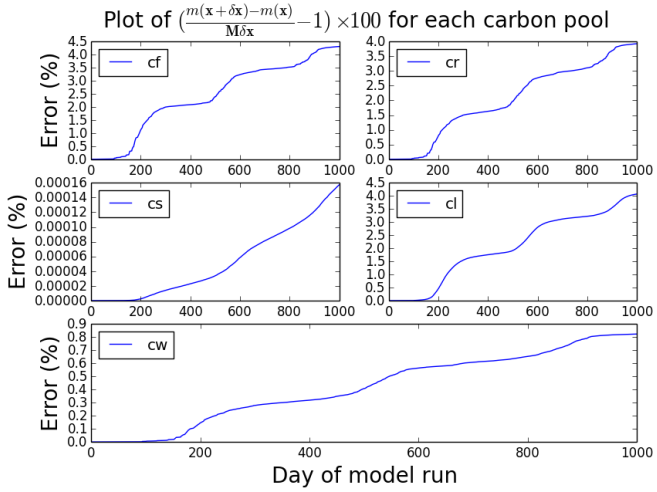


Figure 2: Percentage error in linear model for each carbon pool when run forward for 1000 days.

We see that the error grows as we get further away from the start of our run. We also see that the carbon pools with the most error (C_f, C_r and C_l) are the ones most dependent on GPP, the only nonlinear term in our model. Figure 2 shows us that for DALEC the tangent linear hypothesis is not a bad approximation.

Once we have performed the minimisation of the cost function and determined $\mathbf{x}_a$, we can also calculate the analysis error covariance matrix, $\mathbf{A}$, to quantify the uncertainty in our new estimate of the state. We define the analysis error covariance matrix as,

$$\mathbf{A} = \left(\mathbf{J}''\right)^{-1} = \left(\mathbf{B}^{-1} + \hat{\mathbf{H}}^T\hat{\mathbf{R}}^{-1}\hat{\mathbf{H}}\right)^{-1}, \tag{11}$$

where $\hat{\mathbf{H}}$ is the matrix of linearized observation operators evolved by the tangent linear model and $\hat{\mathbf{R}}$ is the block diagonal matrix of observation error covariance matrices,

$$\hat{\mathbf{H}} = \begin{pmatrix} \mathbf{H}_0 \\ \mathbf{H}_1\mathbf{M}_0 \\ \vdots \\ \mathbf{H}_N\mathbf{M}_{N,0} \end{pmatrix} \quad \text{and} \quad \hat{\mathbf{R}} = \begin{pmatrix} \mathbf{R}_0 & 0 & 0 & 0 \\ 0 & \mathbf{R}_1 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \mathbf{R}_N \end{pmatrix}. \tag{12}$$

It is important to note that we do not need any actual observations to calculate $\mathbf{A}$, $\mathbf{J}''$ or $\mathbf{B}$, only the values of the standard deviations for the observations and the standard deviations for our background values.

3 Information Content Measures

Information content measures are already being used to quantify the different levels of information provided by observations in the development of satellite instruments [4, 12] and in operational data assimilation schemes [5, 11]. In these fields, two of the more widely used measures are Shannon Information Content (also known as entropy reduction) and the degrees of freedom for signal. We will apply both methods for observations assimilated with DALEC.

3.1 Shannon Information Content

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. When a measurement is made, the volume of this *pdf* decreases. The SIC of the measurement is a measure of the factor by which the *pdf* decreases [9]. If $P_b(x)$ is our knowledge of the state before an observation and $P_o(x|y)$ is our knowledge after an observation then we have entropies,

$$S[P_b(x)] = -\int P_b(x)\log_2[P_b(x)]dx \quad \text{and} \quad S[P_o(x|y)] = -\int P_o(x|y)\log_2[P_o(x|y)]dx.$$

The entropy reduction, or SIC, due to the observation is then,

$$SIC = S[P_b(x)] - S[P_o(x|y)]. \tag{13}$$

If we assume all *pdfs* are Gaussian and use the natural logarithm as opposed to $\log_2$ (for algebraic convenience) [9], the entropy of a multivariate Gaussian distruibution for a vector $\mathbf{x}$ with n elements (before and after observations) can be derived as,

$$S[P_b(\mathbf{x})] = n\ln(2\pi e)^{\frac{1}{2}} + \frac{1}{2}\ln\left|\mathbf{B}\right| \tag{14}$$

and

$$S[P_o(\mathbf{x}|\mathbf{y})] = n\ln(2\pi e)^{\frac{1}{2}} + \frac{1}{2}\ln\left|\mathbf{A}\right| \tag{15}$$

where $\mathbf{B}$ is the background error covariance matrix and $\mathbf{A}$ is the analysis error covariance matrix. Combining equations 13, 14 and 15 we can write the SIC as,

$$SIC = \frac{1}{2}\ln\frac{\left|\mathbf{B}\right|}{\left|\mathbf{A}\right|}. \tag{16}$$

From equation 11 we can see the SIC can also be written as,

$$SIC = \frac{1}{2}\ln\left|\mathbf{B}\right|\left|\mathbf{J}''\right|. \tag{17}$$

3.2 Degrees of Freedom for Signal

The degrees of freedom for signal (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 [6].

We have symmetric positive definite background and analysis error covariance matrices $\mathbf{B}$ and $\mathbf{A}$. The eigenvalues of each matrix gives a representation for the uncertainty in the direction of the associated eigenvector, thus, by comparing the eigenvalues of both matrices we can determine the reduction in uncertainty given a set of observations [12].

In order to do this we take $\mathbf{B}^{\frac{-1}{2}}$ such that $\mathbf{B}^{-1} = \mathbf{B}^{\frac{-1}{2}}\mathbf{B}^{\frac{-1}{2}}$. We now take $\mathbf{Q}$ to be the orthogonal matrix composed of the eigenvectors of $\mathbf{B}^{\frac{-1}{2}}\mathbf{A}\mathbf{B}^{\frac{-1}{2}}$ we have,

$$\mathbf{Q}^T\left(\mathbf{B}^{\frac{-1}{2}}\mathbf{A}\mathbf{B}^{\frac{-1}{2}}\right)\mathbf{Q} = \mathbf{\Lambda}, \tag{18}$$

$$\mathbf{Q}^T\left(\mathbf{B}^{\frac{-1}{2}}\mathbf{B}\mathbf{B}^{\frac{-1}{2}}\right)\mathbf{Q} = \mathbf{I}_{n\times n} \tag{19}$$

where $\mathbf{\Lambda}$ is a diagonal matrix. Each diagonal element of our transformed $\mathbf{B}$ is equal to one and corresponds to one degree of freedom. The diagonal elements of $\mathbf{\Lambda}$ correspond to the matrix's eigenvalues and can be interpreted as the relative reduction in variance for each of the n degrees of freedom [11]. We can then define the DFS as,

$$\begin{aligned} DFS &= \text{trace}(\mathbf{I}_{n\times n} - \mathbf{\Lambda}) \\ &= n - \text{trace}(\mathbf{\Lambda}) \\ &= n - \text{trace}(\mathbf{B}^{\frac{-1}{2}}\mathbf{A}\mathbf{B}^{\frac{-1}{2}}) \\ &= n - \text{trace}(\mathbf{B}^{-1}\mathbf{A}). \end{aligned} \tag{20}$$

4 Shannon Information Content for DALEC

We begin by using SIC to understand the information content for different observations when being assimilated with the DALEC model. For these experiments the model is set up as in section 2.1 where the elements of the state vector have standard deviations, $\sigma_{cf,b}, \ldots, \sigma_{cs,b}$, respectively. These standard deviations represent the uncertainty in our inital background estimate, and are taken as a percentage of the initial carbon pools (these values can be found in the appendix). The background error covariance matrix is then taken as the diagonal matrix of the variances of the carbon pools,

$$\mathbf{B} = \begin{pmatrix} \sigma^2_{cf,b} & 0 & 0 & 0 & 0 \\ 0 & \sigma^2_{cr,b} & 0 & 0 & 0 \\ 0 & 0 & \sigma^2_{cw,b} & 0 & 0 \\ 0 & 0 & 0 & \sigma^2_{cl,b} & 0 \\ 0 & 0 & 0 & 0 & \sigma^2_{cs,b} \end{pmatrix}. \tag{21}$$

Our initial experiments look at the SIC in observations taken at a single time; the 4D-Var data assimilation then becomes three-dimensional variational assimilation (3D-Var) as we are not summing over a time window and have no time component.

4.1 SIC for a single observation at one time

If we first consider one observation of C_f (the first element of our state vector $\mathbf{x}$) at time t_0, we can derive an analytical expression for the SIC using,

$$\mathbf{H}_0 = \frac{\delta C_f(t_0)}{\delta \mathbf{x}_0} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \end{pmatrix}, \tag{22}$$

where $\mathbf{H}_0$ is our linearized observation operator at time t_0. As we have a single observation at one time, our observation error covariance matrix, $\mathbf{R_0}$, is just the variance of our observation of C_f at time t_0 ($\sigma^2_{cf,o}$). Therefore,

$$\mathbf{R}_0 = \sigma^2_{cf,o}. \tag{23}$$

We then have from equation 11,

$$\begin{aligned} \mathbf{J}'' &= \mathbf{B}^{-1} + \hat{\mathbf{H}}^T \hat{\mathbf{R}}^{-1} \hat{\mathbf{H}} \\ &= \mathbf{B}^{-1} + \mathbf{H}_0^T \mathbf{R}_0^{-1} \mathbf{H}_0 \\ &= \begin{pmatrix} \sigma^{-2}_{cf,b} + \sigma^{-2}_{cf,o} & 0 & 0 & 0 & 0 \\ 0 & \sigma^{-2}_{cr,b} & 0 & 0 & 0 \\ 0 & 0 & \sigma^{-2}_{cw,b} & 0 & 0 \\ 0 & 0 & 0 & \sigma^{-2}_{cl,b} & 0 \\ 0 & 0 & 0 & 0 & \sigma^{-2}_{cs,b} \end{pmatrix}. \end{aligned} \tag{24}$$

We can now derive the SIC using equation 17 as,

$$SIC = \frac{1}{2} \ln \left|\mathbf{B}\right| \left|\mathbf{J}''\right| = \frac{1}{2} \ln \frac{(\sigma^2_{cf,o} + \sigma^2_{cf,b})}{\sigma^2_{cf,o}} = \frac{1}{2} \ln \left(1 + \frac{\sigma^2_{cf,b}}{\sigma^2_{cf,o}} \right). \tag{25}$$

We see the SIC for an observation of a single observation of C_f is dependent on the ratio between the observation and background variances; the SIC will have the same form for all other direct observations of carbon pools contained in the state. As our background standard deviations are set to being the same percentage for each carbon pool (see appendix), the carbon pool observation which will give us the highest SIC is the pool that we can measure most accurately as this will maximise the ratio $\frac{\sigma^2_{c,b}}{\sigma^2_{c,o}}$ by minimising $\sigma^2_{c,o}$.

One of the main carbon balance observations made at forest 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 respirations of C_l and C_s, giving,

$$NEE(t) = -(1 - p_2) GPP(C_f(t), \phi) + p_8 C_l T(t) + p_9 C_s T(t). \tag{26}$$

For a single observation of NEE at one time, t_0, we can again derive an analytical expression for the SIC using,

$$\mathbf{H}_0 = \frac{\delta NEE(t_0)}{\delta \mathbf{x}_0} = \begin{pmatrix} -(1 - p_2)\zeta_0 & 0 & 0 & p_8 T_0 & p_9 T_0 \end{pmatrix}, \tag{27}$$

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. Again our observation error covariance matrix, $\mathbf{R_0}$, is just the variance of our observation of NEE, $\sigma^2_{nee,o}$, at time t_0. Therefore,

$$\mathbf{R}_0 = \sigma^2_{nee,o} \tag{28}$$

and again,

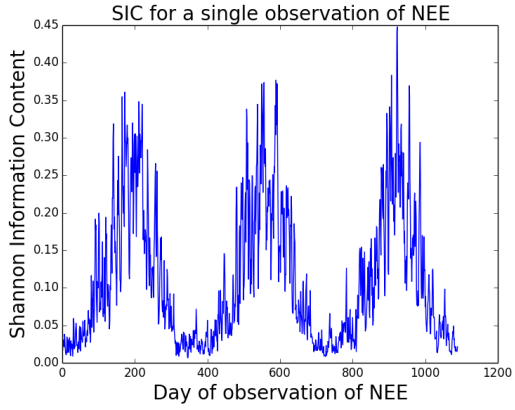
$$\begin{aligned}
\mathbf{J}'' &= \mathbf{B}^{-1} + \mathbf{H}_0^T \mathbf{R}_0^{-1} \mathbf{H}_0 \\
&= \begin{pmatrix}
\sigma^{-2}_{cf,b} + \sigma^{-2}_{nee,o}(1-p_2)^2\zeta_0^2 & 0 & 0 & \sigma^{-2}_{nee,o}(1-p_2)\zeta_0 p_8 T_0 & \sigma^{-2}_{nee,o}(1-p_2)\zeta_0 p_9 T_0 \\
0 & \sigma^{-2}_{cr,b} & 0 & 0 & 0 \\
0 & 0 & \sigma^{-2}_{cw,b} & 0 & 0 \\
\sigma^{-2}_{nee,o}(1-p_2)\zeta_0 p_8 T_0 & 0 & 0 & \sigma^{-2}_{cl,b} + \sigma^{-2}_{nee,o} p_8^2 T_0^2 & \sigma^{-2}_{nee,o} p_8 p_9 T_0^2 \\
\sigma^{-2}_{nee,o}(1-p_2)\zeta_0 p_9 T_0 & 0 & 0 & \sigma^{-2}_{nee,o} p_8 p_9 T_0^2 & \sigma^{-2}_{cs,b} + \sigma^{-2}_{nee,o} p_9^2 T_0^2
\end{pmatrix}.
\end{aligned} \tag{29}$$

We then have,

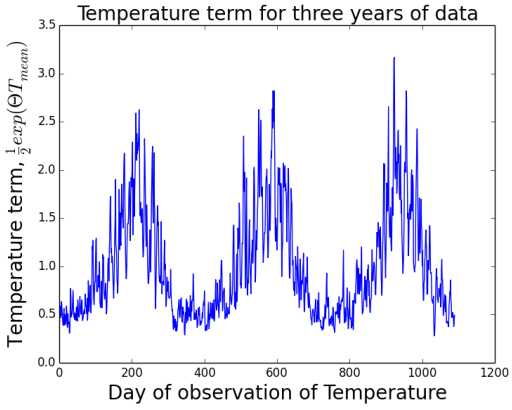
$$SIC = \frac{1}{2}\ln\left|\mathbf{B}\right|\left|\mathbf{J}''\right| = \frac{1}{2}\ln\frac{(p_2-1)^2\zeta_0^2\sigma^2_{cf,b} + \sigma^2_{nee,o} + T_0^2(p_9^2\sigma^2_{cs,b} + p_8^2\sigma^2_{cl,b})}{\sigma^2_{nee,o}}. \tag{30}$$

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 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 ($T(t_i)$) for the same data we can see that both are closely linked, figure 3.



(a) SIC for a single observation of NEE.



(b) Temperature term, $\frac{1}{2}exp(\Theta T_{mean})$.

Figure 3: SIC and temperature varying over three years using driving data from Oregon pine forest.

However the relationship is not linear as the magnitude of GPP's first derivative is also dependent on daily irradiance and the value of the foliar carbon pool (C_f). This shows that observations

of NEE made in the summer are much more valuable than those made in the winter assuming warmer temperatures, higher daily irradiance and a higher amount of foliar carbon in the summer. In the next section we consider a series of observations over some time window. We will see that it takes up to 31 observations of NEE in winter to gain the same level of information as 1 observation of NEE in the summer.

4.2 SIC for successive observations over a time window

Following the results for SIC based at a single time, we now consider the SIC when successive observations are added over a period of time. We begin by considering two successive observations of Cf in time. From section 2.1 and 4.1 we have our linearized observation operator and observation error covariance matrix at time t_i,

$$\mathbf{H}_i = \frac{\delta C_f(t_i)}{\delta \mathbf{x}_i} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{R}_i = \sigma^2_{cf,o}. \tag{31}$$

Then for two successive observations of Cf we have,

$$\hat{\mathbf{H}} = \begin{pmatrix} \mathbf{H}_0 \\ \mathbf{H}_1\mathbf{M}_0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ (1-p_5) + p_3(1-p_2)GPP'(C_f(t_0),\phi) & 0 & 0 & 0 & 0 \end{pmatrix} \tag{32}$$

and

$$\hat{\mathbf{R}} = \begin{pmatrix} \mathbf{R}_0 & 0 \\ 0 & \mathbf{R}_1 \end{pmatrix} = \begin{pmatrix} \sigma^2_{cf,o} & 0 \\ 0 & \sigma^2_{cf,o} \end{pmatrix}. \tag{33}$$

We then have,

$$SIC = \frac{1}{2}\ln\left|\mathbf{B}\right|\left|\mathbf{J}''\right| = \frac{1}{2}\ln\left(1 + \frac{\sigma^2_{cf,b}}{\sigma^2_{cf,o}} + \frac{\sigma^2_{cf,b}\eta_0^2}{\sigma^2_{cf,o}}\right), \tag{34}$$

where $\eta_i = (1-p_5) + p_3(1-p_2)GPP'(C_f(t_i),\phi)$. We see this is similar to equation 25 but with an extra term, η_0^2, which is the square of the first element of our linearized model $\mathbf{M}_0$. We can continue adding more observations at successive times and we start to see a pattern. For three observations at successive times we have,

$$SIC = \frac{1}{2}\ln\left(1 + \frac{\sigma^2_{cf,b}}{\sigma^2_{cf,o}} + \frac{\sigma^2_{cf,b}\eta_0^2}{\sigma^2_{cf,o}} + \frac{\sigma^2_{cf,b}\eta_0^2\eta_1^2}{\sigma^2_{cf,o}}\right), \tag{35}$$

for four,

$$SIC = \frac{1}{2}\ln\left(1 + \frac{\sigma^2_{cf,b}}{\sigma^2_{cf,o}} + \frac{\sigma^2_{cf,b}\eta_0^2}{\sigma^2_{cf,o}} + \frac{\sigma^2_{cf,b}\eta_0^2\eta_1^2}{\sigma^2_{cf,o}} + \frac{\sigma^2_{cf,b}\eta_0^2\eta_1^2\eta_2^2}{\sigma^2_{cf,o}}\right). \tag{36}$$

Using a simple proof by induction we find that for n observations we have,

$$SIC \text{ for } n \text{ observations of } C_f = \frac{1}{2}\ln\left(1 + \frac{\sigma^2_{cf,b}}{\sigma^2_{cf,o}}\left(1 + \sum_{k=0}^{n-2}\prod_{i=0}^{k}\eta_i^2\right)\right). \tag{37}$$

We have plotted the SIC for increasing numbers of observations of C_f, using three years of meteorological driving data from the Oregan pine forest, as seen in figure 4.

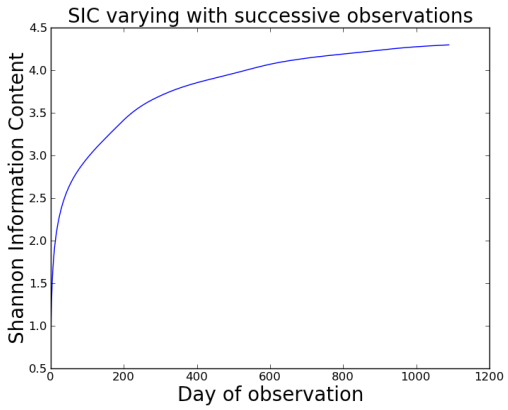


Figure 4: SIC varying as successive observations of Cf are added using driving data from Oregon pine forest.

We can see after around 250 observations of C_f have been made, our SIC starts to level off as we are getting a smaller and smaller reduction in entropy for the extra observations of C_f being made. We see this effect when making successive observations of any of the carbon pools, eventually we cannot reduce the entropy of the system anymore by just adding more of the same observation.

As before with a single observation at one time we can repeat this with successive observations of NEE instead of C_f, this is plotted in figure 5. Here we see a similar levelling off in SIC as in figure 4, we can also see the seasonal cycle of information content in NEE as in figure 3. Once the SIC has levelled off no information is added during the winter months and the SIC is only increased with observations made during the summer.

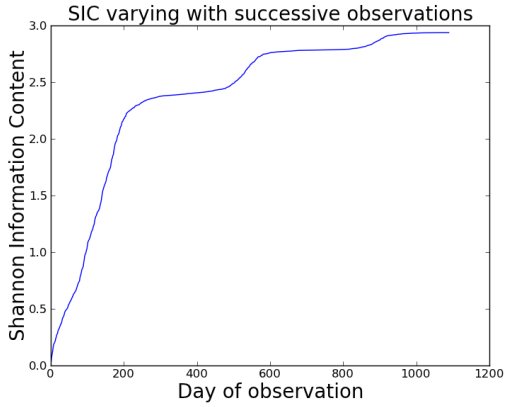


Figure 5: SIC varying as successive observations of NEE are added using driving data from Oregon pine forest.

We can also repeat this with successive observations of both NEE and C_f, this can be seen in figure 6. We attain a higher value of SIC more quickly than before and reach a higher overall value of SIC than in figure 5 or 4. We reduce the entropy of our system very little when adding extra observations of C_f and NEE by day 600.

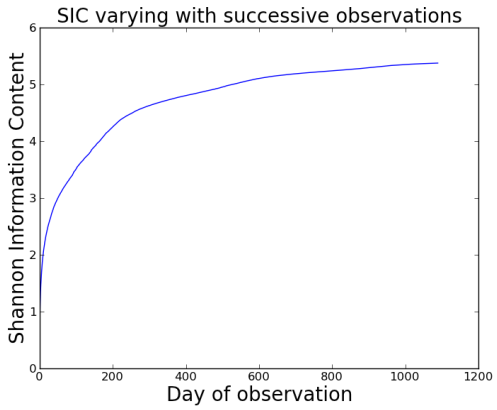


Figure 6: SIC varying as successive observations of NEE and Cf are added using driving data from Oregon pine forest.

In order to see the difference in information content between NEE observations made in the summer and those made in the winter, we have plotted the increasing SIC for 50 observations made from December 21$^{\text{st}}$ onwards and also the the constant line of SIC for one observation of NEE made in summer at a mean daily temperature of 26^{o}C on July 12$^{\text{th}}$. This is shown in figure 7

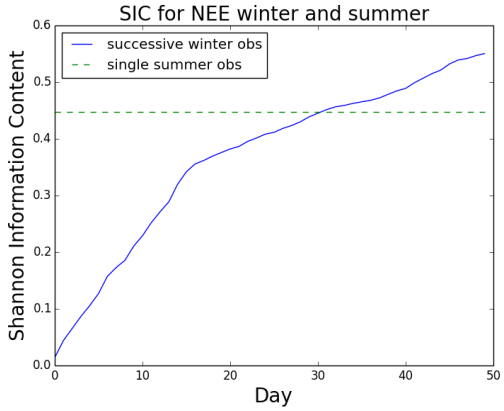


Figure 7: SIC varying as successive observations of NEE are added during winter using driving data from Oregon pine forest. The dashed line shows the SIC for a single observation of NEE made in summer.

Figure 7 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. In the next section (section 5) we will see that we can reproduce all of the results from sections 4.1 and 4.2 using the degrees of freedom for signal measure of information content. We see that although the DFS correspond to the number of elements of the state measured by the observations and SIC represents the reduction in entropy given a set of observations we still see similar results for both measures.

5 Degrees of freedom for signal for DALEC

In this section we repeat some of the experiments from sections 4.1 and 4.2 using the degrees of freedom for signal as a measure for information content, as described in section 3.2. We use the same $\mathbf{B}$ as in equation 21 and calculate $\mathbf{A}$ by finding and inverting $\mathbf{J}''$ (as in section 4.1 and 4.2). We begin by plotting the DFS for a single observation of NEE varying with the three years of driving data as in figure 3 for SIC.

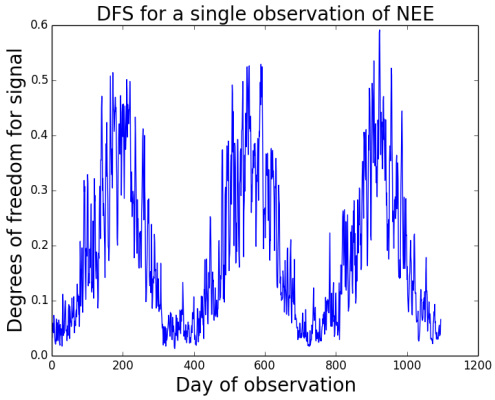


Figure 8: DFS varying for a single observation of NEE over three years of driving data from Oregon pine forest.

We see in figure 8 a very similar result as found in figure 3 with much higher values for the DFS of NEE observations in the summer months, with higher temperatures and higher daily net irradiances. We can repeat the experiment from figure 7 to show that even when successive observations of NEE are added to our data assimilation window in winter it takes 30 daily winter NEE observations to get the same DFS as 1 observation of NEE made in the summer.

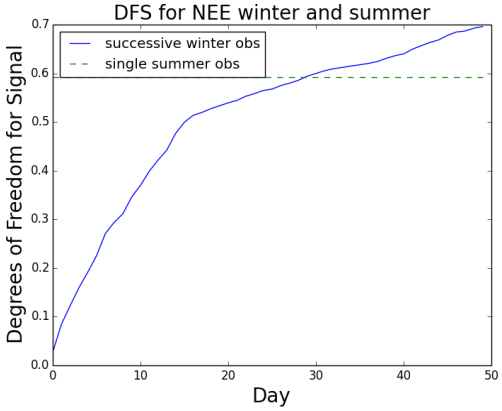


Figure 9: DFS varying as successive observations of NEE are added during winter using driving data from Oregon pine forest. The dashed line shows the DFS for a single observation of NEE made in summer.

The value for the DFS corresponds to the number of state elements deemed measured by the observations. Hence when we make a single observation of any of the carbon pools, our DFS value is close to 1 (it is not 1 however as our observations have error) and as successive observations

of the same carbon pool are added our DFS value starts tending to 1. However with enough observations over a time window of a single carbon pool we can get a DFS value of greater than 1. This is because the model propagates the state forward in time and we start gaining information about the other connected carbon pools from just observing a single carbon pool. The amount of observations it takes to achieve a DFS of more than 1 corresponds to how dependent the observed carbon pool is on other carbon pools. We can see this effect in figure 10 where we have plotted the DFS when observing each carbon pool and also NEE for 100 days of observations.

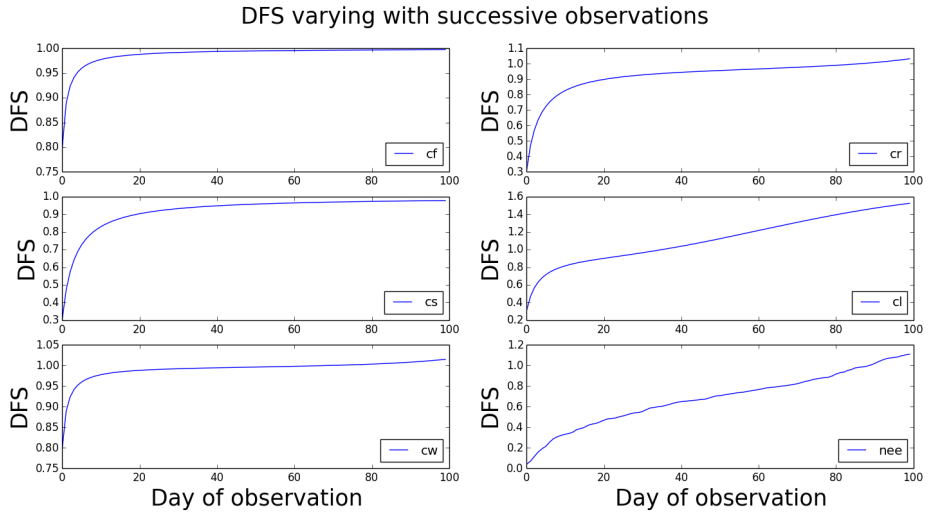


Figure 10: DFS varying as successive observations added using 100 days of driving data from Oregon pine forest. Legend in bottom right corner of each plot denotes which observation is being made.

In order to get a higher DFS it is best to observe a variety of different quantities related to different members of the state. We next repeat the experiment from figure 5 again for DFS instead of SIC.

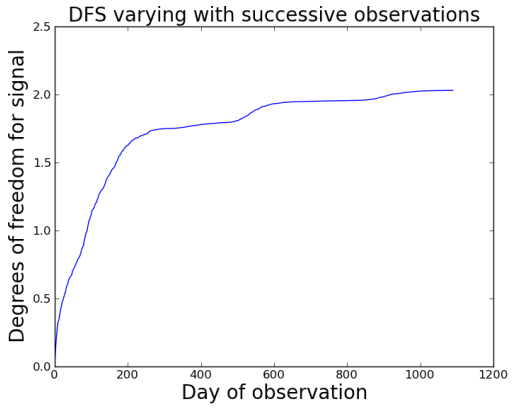


Figure 11: DFS varying as successive observations of NEE are added using three years of driving data from Oregon pine forest.

We see a similar result in figure 11 as in 5. From equation 26 we see that NEE is a function of 3 of our 5 state elements (C_f, C_l and C_s), and in figure 11 it appears that with more observations

the DFS may tend to 3. As NEE is not a direct measurement of the carbon pools we do not have our DFS tending to 3 as quickly as an observation of one of the carbon pools tends to 1.

6 Conclusion

In this report we have taken information content measures previously used in satellite and meteorological data assimilation schemes [5, 11, 12] and applied them to the DALEC carbon balance model [15] implemented in a 4D-Var framework. We have looked at the effects of using different observations and different assimilation time windows, using meteorological driving data from a young pine stand in Ponderossa, Oregon for which our version of DALEC is parameterized.

We have seen that after continually adding the same observation, the information we get from that observation starts to get smaller and eventually we can reduce the entropy of our system no more (or constrain no more elements of our state) with more of the same observation. In order to constrain the most elements of our state (or achieve the maximum possible reduction in entropy for our analysis) it is desirable to take a variety of observations corresponding to different members of the state, if possible. This result is as expected, however, it is important to consider what your state will be when using a forest carbon balance model in a data assimilation framework so that you can identify which observations will correspond to which members of your state and plan your measurement campaign accordingly.

We have also found that NEE observations made in summer are much more valuable than those made in winter, in terms of reducing the entropy of our problem. Although we have used the DALEC carbon balance model for these experiments, NEE is a product of GPP and the respiration fluxes; the GPP function used in DALEC is similar to other such functions from alternative forest carbon balance models, so we would expect to see similar results for these models. NEE observations being more valuable with higher temperatures makes physical sense as this is when NEE observations are of greatest magnitude and give us most information about the fluxes of carbon.

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] Dennis Baldocchi, Eva Falge, Lianhong Gu, Richard Olson, David Hollinger, Steve Running, Peter Anthoni, Ch Bernhofer, Kenneth Davis, Robert Evans, et al. 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(11):2415–2434, 2001.

[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] RJ Engelen and GL Stephens. Information content of infrared satellite sounding measurements with respect to co2. *Journal of Applied Meteorology*, 43(2):373–378, 2004.

[5] Michael Fisher. *Estimation of entropy reduction and degrees of freedom for signal for large variational analysis systems.* European Centre for Medium-Range Weather Forecasts, 2003.

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

[7] Andrew Fox, Mathew Williams, Andrew D Richardson, David Cameron, Jeffrey H Gove, Tristan Quaife, Daniel Ricciuto, Markus Reichstein, Enrico Tomelleri, Cathy M Trudinger, et al. 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(10):1597–1615, 2009.

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

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

[10] SW Running, DD Baldocchi, DP Turner, ST Gower, PS Bakwin, and KA Hibbard. A global terrestrial monitoring network integrating tower fluxes, flask sampling, ecosystem modeling and eos satellite data. *Remote Sensing of Environment*, 70(1):108–127, 1999.

[11] Adrian Sandu, Kumaresh Singh, Mohamed Jardak, Kevin Bowman, and Meemong Lee. A practical method to estimate information content in the context of 4d-var data assimilation. i: Methodology. 2012.

[12] 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.

[13] R Valentini, G Matteucci, AJ Dolman, E-D Schulze, CJMEAG Rebmann, EJ Moors, A Granier, P Gross, NO Jensen, K Pilegaard, et al. Respiration as the main determinant of carbon balance in european forests. *Nature*, 404(6780):861–865, 2000.

[14] 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.

[15] 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.

Appendix

The Aggregated Canopy Model

The aggregated canopy model (ACM) is used in DALEC to calculate GPP. The ACM uses the following equations (for more detail please see [14]),

$$g_c = \frac{|\psi_d|^{a_{10}}}{\frac{1}{2}T_r + a_6 R_{tot}}, \tag{38}$$

$$p = \frac{a_1 NL}{g_c} \exp(T_{max} a_8), \tag{39}$$

$$q = a_3 - a_4, \tag{40}$$

$$C_i = \frac{1}{2}\left[C_a + q - p + \sqrt{(C_a + q + p)^2 - 4(C_a q - p a_3)}\right], \tag{41}$$

$$E_0 = \frac{a_7 L^2}{L^2 + a_9}, \tag{42}$$

$$\delta = -0.408 \arccos\left(\frac{360(D+10)}{365}\frac{\pi}{180}\right), \tag{43}$$

$$s = 24 \arccos(-\tan(lat)\tan(\delta))/\pi, \tag{44}$$

$$GPP = \frac{E_0 I g_c (C_a - C_i)}{E_0 I + g_c (C_a - C_i)}(a_2 s + a_5). \tag{45}$$

The symbols have the following meanings ($a_1, \ldots, a_{10}$ set parameters),

Symbol	Description
g_c	Canopy conductance ($gCm^{-2}day^{-1}$)
ψ_d	Max soil-leaf water potential difference (MPa)
T_r	Daily temperature range (^oC)
R_{tot}	Total plant-soil hydraulic resistance ($MPam^2smmol^{-1}$)
N	Foliar nitrogen (gNm^{-2} leaf area)
L	Leaf area index (m^2m^{-2})
T_{max}	Maximum daily temperature (^oC)
C_a	Atmospheric CO_2 concentration ($\mu molmol^{-1}$)
C_i	CO_2 concentration at site of carboxylation ($\mu molmol^{-1}$)
E_0	Canopy level quantum yield ($gCMJ^{-1}m^{-2}day^{-1}$)
δ	Solar declination ($rads$)
D	Day of year
s	Day length (hrs)
lat	Site latitude (^{o})
I	Irradiance ($MJm^{-2}day^{-1}$)

Table 1: Symbols used in ACM.

DALEC rate parameters and initial pool values

Carbon pool	Initial value (gCm^{-2})	Background standard deviation (as a % of initial pool)
C_f	58	20%
C_r	102	20%
C_w	770	20%
C_l	40	20%
C_s	9897	20%

Table 2: Inital carbon pool values for the DALEC model and the standard deviations expressed as a percentage of the initial value.

Parameter	Description	Value
p_1	decomposition rate	4.41×10^{-6}
p_2	fraction of GPP respired	0.47
p_3	fraction of GPP allocated to foliage	0.31
p_4	fraction of GPP allocated to roots	0.43
p_5	turnover rate of foliage	2.7×10^{-3}
p_6	turnover rate of woods	2.06×10^{-6}
p_7	turnover rate of roots	2.48×10^{-3}
p_8	mineralisation rate of C_l	2.28×10^{-2}
p_9	mineralisation rate of C_s	2.65×10^{-6}

Table 3: Parameters for evergreen DALEC.