Improvement of forest carbon balance model DALEC2 for flux site Alice Holt using four-dimensional variational data assimilation.

Ewan Pinnington

April 23, 2015

1 Introdution

Four-dimensional variational data assimilation (4D-Var) has been used extensively in numerical weather prediction to improve forecasts REF. Currently efforts to use variational assimilation with carbon balance models have been limited REF, with sequential and Markov Chain Monte Carlo methods being more prevalent REF. In this report we will paramterize a model of forest carbon balance (DALEC2) using a 4D-Var scheme in order to produce better forecasts of forest carbon balance. We will use data from the research site at Alice Holt forest run by Forest Research.

2 Methods

2.1 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,

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

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

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

2.2 The DALEC2 model

The DALEC2 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.

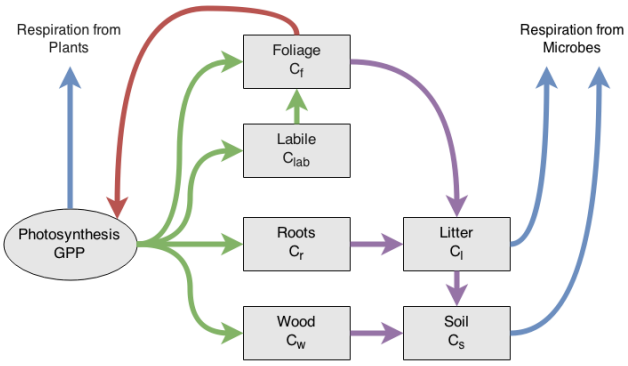

Figure 1: Representation of the fluxes in the DALEC2 carbon balance model. Green arrows represent C allocation, purple arrows represent litter fall and decomposition fluxes, blue arrows represent respiration fluxes and the red arrow represents the feedback of foliar carbon to the GPP function.

The model equations for the carbon pools at day $t+1$ are as follows:

$$GPP^t = ACM(C_f^t, c_{lma}, c_{eff}, \Psi) \tag{6}$$
$$C_{lab}^{t+1} = (1-\Phi_{on})C_{lab}^t + (1-f_{auto})(1-f_{fol})f_{lab}GPP^t, \tag{7}$$
$$C_f^{t+1} = (1-\Phi_{off})C_f^t + \Phi_{on}C_{lab}^t + (1-f_{auto})f_{fol}GPP^t, \tag{8}$$
$$C_r^{t+1} = (1-\theta_{roo})C_r^t + (1-f_{auto})(1-f_{fol})(1-f_{lab})f_{roo}GPP^t, \tag{9}$$
$$C_w^{t+1} = (1-\theta_{woo})C_w^t + (1-f_{auto})(1-f_{fol})(1-f_{lab})(1-f_{roo})GPP^t, \tag{10}$$
$$C_l^{t+1} = (1-(\theta_{lit}+\theta_{min})e^{\Theta T^t})C_l^t + \theta_{roo}C_r^t + \Phi_{off}C_f^t, \tag{11}$$
$$C_s^{t+1} = (1-\theta_{som}e^{\Theta T^t})C_s^t + \theta_{woo}C_w^t + \theta_{min}e^{\Theta T^t}C_l^t, \tag{12}$$

where T^t is the daily mean temperature, Ψ represents the meteorological driving data used in the GPP function and Φ_{on}/Φ_{off} are functions controlling leaf on and leaf off. The model parameters used in equations 6 to 12 and the equations used to calculate GPP, Φ_{on} and Φ_{off} are included in the appendix. The full details of this version of DALEC can be found in Bloom and Williams [2014]. We now see how DALEC2 can be implemented in a 4D-Var framework.

2.3 DALEC2 4D-Var scheme

In our DALEC2 4D-Var scheme the state vector, $\mathbf{x}_0$, corresponds to the vector of the 17 model parameters and 6 initial carbon pool values defined in the appendix. 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, these values can be found in the appendix.

In order to find the tangent linear model (TLM) for DALEC2 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. To do this we use the AlgoPy automatic differentiation package in Python REF. For our $\mathbf{x}_b$ we use a parameter set found by the University of Edinburgh using the CARbon DAta-MOdel fraMework (CARDAMOM) REF. This used Harmonized World Soil Database (HWSD) C_s observations as initial conditions, meteorological driving data from ERA-interim and Markov chain Monte Carlo (MCMC) techniques to assimilate MODIS leaf area index (LAI) observations over a 10 year period. We now have all the tools to create our 4D-Var scheme which is implemented in Python. In sections 2.3.1 to 2.3.3 we will show some tests of our scheme.

2.3.1 Test of tangent linear model

In 4D-Var we assume the tangent linear hypothesis,

$$m_{0\to i}(\mathbf{x}_0 + \delta\mathbf{x}_0) \approx m_{0\to i}(\mathbf{x}_0) + \mathbf{M}_{i,0}\delta\mathbf{x}_0. \tag{13}$$

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 13 to find,

$$\text{percentage error in TLM} = \left|\frac{m_{0\to i}(\mathbf{x}_0+\delta\mathbf{x}_0)-m_{0\to i}(\mathbf{x}_0)}{\mathbf{M}_{i,0}\delta\mathbf{x}_0} - 1\right| \times 100. \tag{14}$$

In figure 2 we can see that our TLM for DALEC2 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

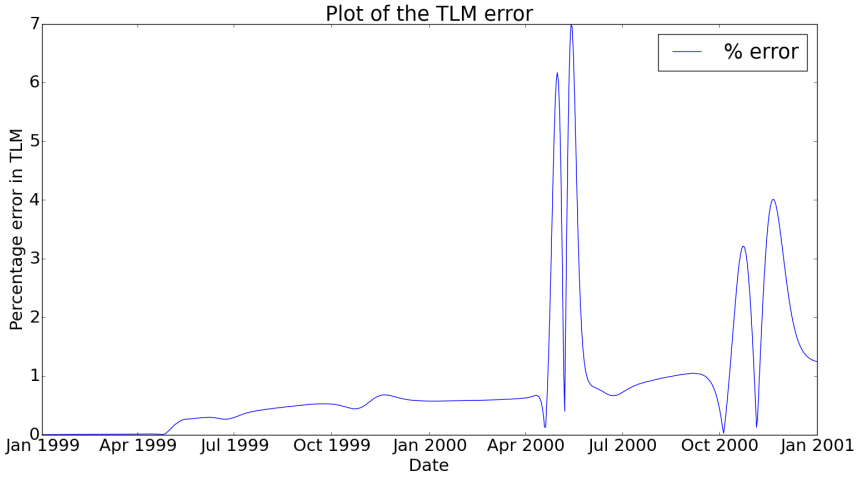

Figure 2: Plot of the percentage error in our tangent linear model for DALEC2 when evolving our state forward over a period of two years.

autumn, this is where our leaf on and leaf off functions in the TLM have gone out of phase with the nonlinear DALEC2. Even at these peaks our error is still reasonable reaching a maximum at 7% and then coming back to around 1%.

Maybe also show tangent linear model behaves as we expect with, for small perturbations $\gamma\delta\mathbf{x}_0$ we have

$$m_{0\to i}(\mathbf{x}_0 + \gamma\delta\mathbf{x}_0) - m_{0\to i}(\mathbf{x}_0) \approx \mathbf{M}_{i,0}\gamma\delta\mathbf{x}_0. \tag{15}$$

If we then plot the relative error

$$E_R = \frac{m_{0\to i}(\mathbf{x}_0 + \gamma\delta\mathbf{x}_0) - m_{0\to i}(\mathbf{x}_0)}{\mathbf{M}_{i,0}\gamma\delta\mathbf{x}_0} \tag{16}$$

as $\gamma \to 0$ we should have that $E_R \to 0$.

PLOT

2.3.2 Test of adjoint model

For our TLM $\mathbf{M}_{i,0}$ and its adjoint $\mathbf{M}_{i,0}^T$ we have the identity

$$< \mathbf{M}_{i,0}\delta\mathbf{x}_0, \mathbf{M}_{i,0}\delta\mathbf{x}_0 > = < \delta\mathbf{x}_0, \mathbf{M}_{i,0}^T\mathbf{M}_{i,0}\delta\mathbf{x}_0 > \tag{17}$$

for any inner product $<,>$ and perturbation $\delta\mathbf{x}_0$. We can use this to check that our adjoint is coded correctly.

2.3.3 Gradient test

For our cost function J and its gradient ∇J we can check that we have coded the ∇J correctly using the identity

$$f(\alpha) = \frac{J(\mathbf{x}_0 + \alpha\mathbf{h}) - J(\mathbf{x}_0)}{\alpha\mathbf{h}^T\nabla J(\mathbf{x}_0)} = 1 + O(\alpha), \tag{18}$$

where $\mathbf{h}$ is a vector of unit length, which we can take to be $\nabla J(\mathbf{x}_0)||\nabla J(\mathbf{x}_0)||^{-1}$. Maybe also consider $\mathbf{h} = \mathbf{x}_0||\mathbf{x}_0||^{-1}$?? In plot REF we can see that as $\alpha \to 0$ we have $f(\alpha) \to 1$ before α gets too close to machine zero.

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/`.

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

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.

Appendix