Investigating the role of **R** and **B** in improving a model of forest carbon balance using 4D-Var.

Ewan Pinnington

September 25, 2015

Abstract

Efforts to implement functional ecology and land surface models in variational data assimilation routines have been limited with sequential and Markov chain Monte Carlo data assimilation methods being more prevalent. When data assimilation has been used with models of carbon balance, background errors (describing our knowledge of error in our prior model estimates before data assimilation) and observation errors have largely been treated as independent and uncorrelated. In numerical weather prediction it has been shown that including correlations in these errors can considerably improve data assimilation results and forecasts. In this paper we implement a simple model of forest carbon balance in a Four-Dimensional Variational (4D-Var) data assimilation scheme, for parameter and state estimation, assimilating observations of Net Ecosystem Exchange (NEE) taken at the Alice Holt flux site in Hampshire managed by Forest Research. We then investigate the effect of specifying correlations between parameter and state variables in background errors and the effect of specifying correlations in time between observation errors. We do this by moving away from a diagonal representation of the background error covariance matrix, **B**, and the observation error covariance matrix, **R**. We outline novel methods for creating a correlated **B** and **R** and show that using these new correlated matrices can almost half the root mean square error in our models forecast of NEE in comparison to the results when using an uncorrelated diagonal **B** and **R**.

1 Introdution

1.1 Tree blurb

Terrestrial ecosystems and oceans are responsible for removing around half of all human emitted carbon-dioxide from the atmosphere and therefore greatly reduce the effect of anthropogenic induced climate change. Terrestrial ecosystem carbon uptake is the least understood process in the global carbon cycle [Ciais et al., 2014]. It is therefore vital that we improve understanding of the carbon uptake of terrestrial ecosystems and their response to climate change in order to better constrain predictions of future carbon budgets. Observations of the Net Ecosystem Exchange (NEE) of CO_2 between terestrial ecosystems and the atmosphere are now routinely made at flux tower sites world-wide [Baldocchi, 2008] providing a great resource for model validation and data assimilation.

1.2 DA paragraph

Data assimilation is the process of combining a mathematical model with observations in order to improve the estimate of the state of a system. Data assimilation has been used in many applications

with great results for improving models and forecast. One such application has been In numerical weather prediction where the impact of data assimilation has been vast with the four day forecast in 2014 having the same level of accuracy as the one day forecast when data assimilation was first introduced operationally in 1979 [Kalnay, 2003, Rabier, 2005]. This increase in forecast skill is obviously not solely due to data assimilation but also increased quality and resolution of observations along with improvements in model structure. However the introduction and evolution of data assimilation has played a large part. The current method implemented at several leading operational numerical weather prediction centres is Four-Dimensional Variational data assimilation (4D-Var), which has been shown to be a significant improvement over its predecessor three-dimensional variational assimilation [Lorenc and Rawlins, 2005]. Variational assimilation techniques minimise a cost function to find an improved state of a system. The minimisation routine requires the derivative of the model, this can sometimes prove a problem to find. However using techniques such as automatic-differentiation can reduce the time taken to implement the derivative of a model and also increase the accuracy of the models derivative.

1.3 DALEC and ecosystem models with data assimilation (draw out gaps):

Many different observations relevant to the carbon balance of forests have now been combined with functional ecology models, using data assimilation, in order to improve our knowledge ecological systems [Fox et al., 2009, Zobitz et al., 2011]. Two such models that have been used extensively with data assimilation are the Data Assimilation Linked Ecosystem Carbon (DALEC) model [Williams et al., 2005] and the Simplified Photosynthesis and Evapo-Transpiration (SIPNET) model [Braswell et al., 2005], nearly all data assimilation routines built with these models have used sequential and Monte Carlo Markov chain (MCMC) data assimilation methods with the exception of DALEC being implemented in a variational routine by Delahaies et al. [2013]. There have been examples of global land surface models being implemented with variational methods such as the Carbon Cycle Data Assimilation System (CCDAS) [Kaminski et al., 2013] and the global model ORganizing Carbon and Hydrology In Dynamic EcosystEms (ORCHIDEE) [Krinner et al., 2005]. These examples have mainly been used to assimilate data from satellite observations with a few examples where site level data has been assimilated [Bacour et al., 2015, Verbeeck et al., 2011].

Currently background and observation errors have been treated as uncorrelated and independent in ecosystem model data assimilation schemes. In many data assimilation schemes background and observation errors are represented by the error covariance matrices $\mathbf{B}$ and $\mathbf{R}$ respectively. The off-diagonal elements of these matrices indicate the correlations between the parameter and state variables for $\mathbf{B}$ and the correlations between observation errors for $\mathbf{R}$. Including correlations in both $\mathbf{B}$ and $\mathbf{R}$ has been shown to significantly improve data assimilation results in numerical weather prediction, with much more research ongoing in this field [Fisher, 2003, Stewart et al., 2013]. Although correlations in $\mathbf{R}$ have been explored in numerical weather prediction, currently these have been mainly between observations made at the the same time rather than correlations between observations throughout time. In Richardson et al. [2010] the problem of data streams with many more observations having more impact on assimilation results than those with fewer observations is discussed and an assimilation method put forward to deal with this. Specifying serial correlations between observations of the same quantity decreases the impact of these observations [Järvinen et al., 1999] and represents another way to tackle this problem, whilst also adding valuable information to our data assimilation routine.

1.4 What does this paper do/results:

In this paper we implement the new version of DALEC (DALEC2 [Bloom and Williams, 2015]) in a 4DVAR data assimilation scheme for state and parameter estimation, assimilating daily NEE observations from the Alice Holt flux site in Hampshire managed by Forest Research. This assimilation scheme is then subjected to rigorous testing to ensure correctness. We then outline a new method for including parameter and state correlations in our background error covariance matrix and a method for including serial correlations in our observation error covariance matrix. These matrices are then used in a series of experiments in order to examine the effect that including correlations in our assimilation scheme has on our results. We show that specifying parameter and state correlations in our prior knowledge and serial correlations between observation errors can drastically improve the results from our assimilation.

2 Model and Data Assimilation Methods

2.1 Alice Holt research forest

Alice Holt is a well established research forest located in Hampshire, England with observational data spanning 50 years, the site is managed by Forest Research. The flux tower site is situated in the Straits Inclosure which is a mainly deciduous part of the forest comprising of mostly oak trees with a hazel understory, although there is a bank of conifer approximately 1km north west of the flux tower site. Flux tower records of Net Ecosystem Exchange (NEE) are available from 1999 up to present day, meaning that Alice Holt has one of the longest records of flux tower data in the UK.

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, 2015] 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.

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

$$GPP^t = ACM(C_f^t, c_{lma}, c_{eff}, \Psi) \tag{1}$$

$$C_{lab}^{t+1} = (1-\Phi_{on})C_{lab}^t + (1-f_{auto})(1-f_{fol})f_{lab}GPP^t, \tag{2}$$

$$C_f^{t+1} = (1-\Phi_{off})C_f^t + \Phi_{on}C_{lab}^t + (1-f_{auto})f_{fol}GPP^t, \tag{3}$$

$$C_r^{t+1} = (1-\theta_{roo})C_r^t + (1-f_{auto})(1-f_{fol})(1-f_{lab})f_{roo}GPP^t, \tag{4}$$

$$C_w^{t+1} = (1-\theta_{woo})C_w^t + (1-f_{auto})(1-f_{fol})(1-f_{lab})(1-f_{roo})GPP^t, \tag{5}$$

$$C_l^{t+1} = (1-(\theta_{lit}+\theta_{min})e^{\Theta T^t})C_l^t + \theta_{roo}C_r^t + \Phi_{off}C_f^t, \tag{6}$$

$$C_s^{t+1} = (1-\theta_{som}e^{\Theta T^t})C_s^t + \theta_{woo}C_w^t + \theta_{min}e^{\Theta T^t}C_l^t, \tag{7}$$

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

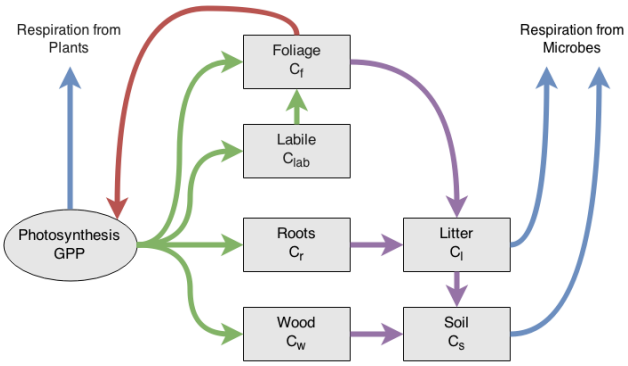

Figure 1: Representation of the fluxes in the DALEC2 carbon balance model. Green arrows represent C allocation, purple arrows represent litter fall and decomposition fluxes, blue arrows represent respiration fluxes and the red arrow represents the feedback of foliar carbon to the GPP function.

parameters used in equations 1 to 7 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 [2015].

2.3 4D-Var

In 4D-Var we aim to maximise the probability of the 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 Bayes Theorem [Lewis et al., 2006]. The cost function is given as,

$$J(\mathbf{x}_0) = \frac{1}{2}(\mathbf{x}_0 - \mathbf{x}_b)^T\mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}_b) + \frac{1}{2}\sum_{i=0}^{N}(\mathbf{y}_i - h_i(\mathbf{x}_i))^T\mathbf{R}_i^{-1}(\mathbf{y}_i - h_i(\mathbf{x}_i)), \tag{8}$$

where $\mathbf{x}_b$ is the background and acts as the initial guess to the state $\mathbf{x}_0$, $\mathbf{B}$ is the background error covariance matrix and quantifies our knowledge of the error in the background, h_i is the observation operator at time t_i and maps the state vector evolved by the nonlinear model ($m_{0\rightarrow i}(\mathbf{x}_0) = \mathbf{x}_i$) 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 (here we use a truncated Newton method [Nocedal and Wright, 1999] from the Python package Scipy.optimize) that takes the cost function, the initial guess ($\mathbf{x}_b$) and also the gradient of the cost function given as,

$$\nabla J(\mathbf{x}_0) = \mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}_b) - \sum_{i=0}^{N}\mathbf{M}_{i,0}^T\mathbf{H}_i^T\mathbf{R}_i^{-1}(\mathbf{y}_i - h_i(\mathbf{x}_i)), \tag{9}$$

where $\mathbf{H}_i = \frac{\partial h_i(\mathbf{x}_i)}{\partial \mathbf{x}_i}$ 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 $\mathbf{M}_i = \frac{\partial m_i(\mathbf{x}_i)}{\partial \mathbf{x}_i}$. We can rewrite the cost function and its gradient to avoid the sum notation as,

$$J(\mathbf{x}_0) = \frac{1}{2}(\mathbf{x}_0 - \mathbf{x}_b)^T\mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}_b) + \frac{1}{2}(\hat{\mathbf{y}} - \hat{h}(\mathbf{x}_0))^T\hat{\mathbf{R}}^{-1}(\hat{\mathbf{y}} - \hat{h}(\mathbf{x}_0)) \tag{10}$$

and

$$\nabla J(\mathbf{x}_0) = \mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}_b) - \hat{\mathbf{H}}^T \hat{\mathbf{R}}^{-1}(\hat{\mathbf{y}} - \hat{h}(\mathbf{x}_0)), \tag{11}$$

where,

$$\hat{\mathbf{y}} = \begin{pmatrix} \mathbf{y}_0 \\ \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_N \end{pmatrix}, \ \hat{h}(\mathbf{x}_0) = \begin{pmatrix} h_0(\mathbf{x}_0) \\ h_1(m_{0\to 1}(\mathbf{x}_0)) \\ \vdots \\ h_N(m_{0\to N}(\mathbf{x}_0)) \end{pmatrix}, \ \hat{\mathbf{R}} = \begin{pmatrix} \mathbf{R}_0 & 0 & 0 & 0 \\ 0 & \mathbf{R}_1 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \mathbf{R}_N \end{pmatrix} \text{ and } \hat{\mathbf{H}} = \begin{pmatrix} \mathbf{H}_0 \\ \mathbf{H}_1\mathbf{M}_0 \\ \vdots \\ \mathbf{H}_N\mathbf{M}_{N,0} \end{pmatrix}. \tag{12}$$

Solving the cost function in this form will also allow us to build serial correlations into the observation error covariance matrix $\hat{\mathbf{R}}$.

2.4 Testing of 4D-Var system

In our DALECV2 4D-Var scheme the state vector, $\mathbf{x}_0$, corresponds to the vector of the 17 model parameters and 6 initial carbon pool values, which can be found in the appendix. In the following tests we use a diagonal approximation to our background and observational error covariance matrices so that, $\mathbf{B} = \text{diag}(\underline{\sigma}_b^2)$ and $\hat{\mathbf{R}} = \text{diag}(\underline{\sigma}_o^2)$, where $\underline{\sigma}_b$ and $\underline{\sigma}_o$ are the vectors of the background and observational standard deviations respectively.

In order to find the tangent linear model (TLM) for DALECV2 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. We use the AlgoPy automatic differentiation package in Python to calculate the TLM at each time step. This package uses forward mode automatic differentiation to calculate the derivative of our model. AlgoPy was selected after testing other automatic differentiation packages (PyAutoDiff and ad.py) and finding that AlgoPy could compute the TLM in the fastest time. We now have all the tools to create our 4D-Var scheme. In sections 2.4.1 to 2.4.3 we will show some tests of our scheme.

2.4.1 Test of tangent linear model

We can have confidence that our implementation of the TLM for DALEC2 is correct as it passes relevant tests. In 4D-Var we assume the tangent linear hypothesis,

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

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 this by rearranging equation 13 to find the relative error,

$$E_R = \frac{||m_{0\to i}(\mathbf{x}_0 + \gamma\delta\mathbf{x}_0) - m_{0\to i}(\mathbf{x}_0)||}{||\mathbf{M}_{i,0}\gamma\delta\mathbf{x}_0||}, \tag{14}$$

where we should have $E_R \to 0$ as $\gamma \to 0$. In figure 2 we have plotted equation 14 for DALEC2 with a TLM evolving our state 731 days forward in time for a 5% perturbation $\delta\mathbf{x}_0$. Figure 2 shows that our TLM behaves as expected for values of γ approaching 0.

It is also useful to show how our TLM behaves over a time window to see how the error in our TLM grows as we evolve our state further forward in time. We again rearrange equation 13 to find,

$$\text{percentage error in TLM} = \left| \frac{||m_{0\to i}(\mathbf{x}_0+\delta\mathbf{x}_0) - m_{0\to i}(\mathbf{x}_0)||}{||\mathbf{M}_{i,0}\delta\mathbf{x}_0||} - 1 \right| \times 100. \tag{15}$$

In figure 3 we can see that our TLM for DALEC2 performs very well after being run forward a year with less than a 3% error for all values of $\delta\mathbf{x}_0$. By the second year we see some peaks in our

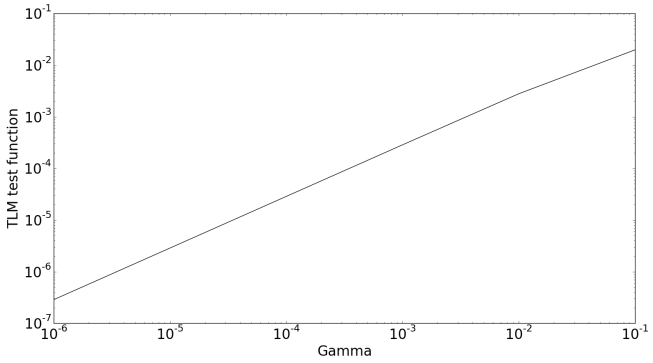

Figure 2: Plot of the tangent linear model test function for DALEC2, for a TLM evolving our state 731 days forward in time and a 5% perturbation, $\delta\mathbf{x}_0$.

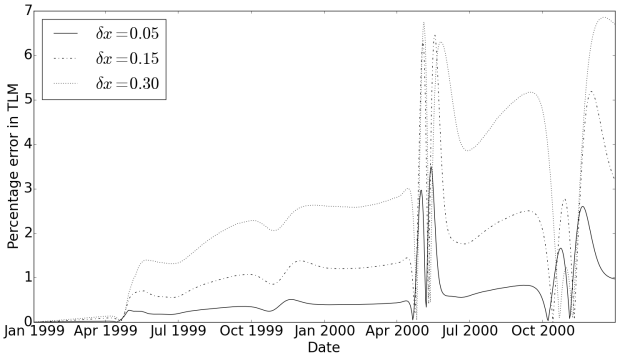

Figure 3: Plot of the percentage error in our tangent linear model for DALEC2 when evolving our state forward over a period of two years with three differing values of perturbation, $\delta\mathbf{x}_0$.

error in spring and 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%. For this reason we present results using a one year assimilation window in this paper.

2.4.2 Test of adjoint model

The adjoint model we have implemented for DALEC2 passes correctness tests. For our TLM $\mathbf{M}_{i,0}$ and its adjoint $\mathbf{M}_{i,0}^T$ we have the identity

$$< \mathbf{M}_{i,0}\delta\mathbf{x}_0, \mathbf{M}_{i,0}\delta\mathbf{x}_0 >=< \delta\mathbf{x}_0, \mathbf{M}_{i,0}^T\mathbf{M}_{i,0}\delta\mathbf{x}_0 > \tag{16}$$

for any inner product $<,>$ and perturbation $\delta\mathbf{x}_0$. This identity has been used with differing values of $\delta\mathbf{x}_0$ and i to show that our adjoint model is implemented correctly.

2.4.3 Gradient test

The 4D-Var system we have developed passes tests for the gradient of the cost function. For our cost function J and its gradient ∇J we can show that we have implemented ∇J correctly using

the identity,

$$f(\alpha) = \frac{J(\mathbf{x}_0 + \alpha\mathbf{h}) - J(\mathbf{x}_0)}{\alpha\mathbf{h}^T \nabla J(\mathbf{x}_0)} = 1 + O(\alpha), \tag{17}$$

where $\mathbf{h}$ is a vector of unit length. For small values of α not too close to machine zero we should have $f(\alpha)$ close to 1. In figure 4 we have plotted $f(\alpha)$ for a 731 day assimilation window with $\mathbf{h} = \mathbf{x}_0||\mathbf{x}_0||^{-1}$, we can see that $f(\alpha) \to 1$ as $\alpha \to 0$, as expected.

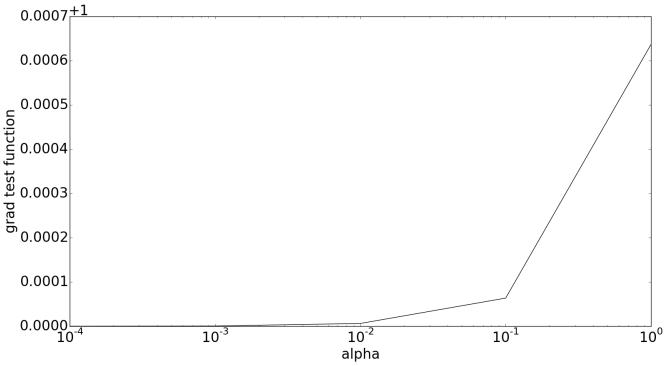

Figure 4: Test of the gradient of the cost function for a 731 day assimilation window with $\mathbf{h} = \mathbf{x}_0||\mathbf{x}_0||^{-1}$.

We can also plot $|f(\alpha) - 1|$, where we expect $|f(\alpha) - 1| \to 0$ as $\alpha \to 0$. In figure 5 we have plotted $|f(\alpha) - 1|$ for the same conditions as in figure 4, we can see that $|f(\alpha) - 1| \to 0$ as $\alpha \to 0$, as expected (before $|f(\alpha) - 1|$ gets too close to machine zero at $O(\alpha) = 10^{-5}$). This gives us confidence that the gradient of our cost function is implemented correctly.

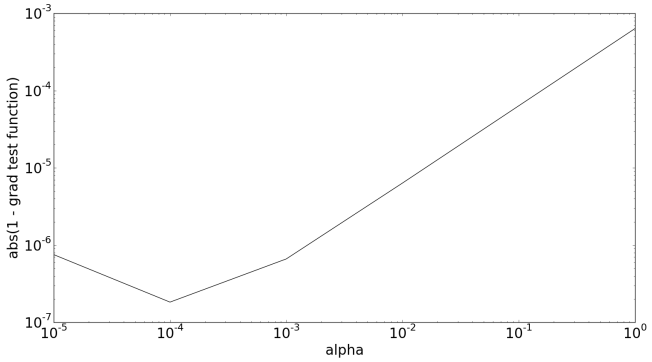

Figure 5: Test of the gradient of the cost function, $|f(\alpha) - 1|$. As $\alpha \to 0$ we have $abs(1 - f(\alpha)) \to 0$ up to $O(\alpha) = 10^{-4}$ where we have gone past the precision of the computer.

2.5 Including correlations in the background error covariance matrix

The background vector ($\mathbf{x}_b$) and its corresponding variances used in this paper were found by the University of Edinburgh using the CARbon DAta-MOdel fraMework (CARDAMOM) [Exbrayat et al., 2015]. This used Harmonised World Soil Database C_s observations as initial conditions, meteorological driving data from ERA-interim and MCMC techniques to assimilate MODIS leaf

area index observations covering the Alice Holt forest over a 10 year period. So without correlations we just have the diagonal matrix of the variances, denoted as $\mathbf{B}_{diag}$.

Including correlations in $\mathbf{B}$ impacts how information from assimilated observations is spread between different types of analysis variables [Singh et al., 2011]. We explored a number of different methods in order to include parameter-state correlations in $\mathbf{B}$, in this paper we present a method using a set of ecological dynamical constraints on model parameters and state variables from Bloom and Williams [2015]. In this paper implementing these constraints in a Metropolis Hastings MCMC data assimilation routine are shown to improve results significantly. The constraints impose conditions on carbon pool turnover and allocation ratios, steady state proximity and growth and decay of model carbon pools.

In order to create a correlated background error covariance matrix, $\mathbf{B}_{corr}$, using these constraints we first create an ensemble using the following procedure:

1. Draw a random state vector, $\mathbf{x}_i$, from the multivariate truncated normal distribution described by our $\mathbf{x}_b$, associated variances and parameter-state ranges given in table REF (in appendix).
2. Test this $\mathbf{x}_i$ with the ecological dynamical constraints.
3. If $\mathbf{x}_i$ passes it is added to to our ensemble, else it is discarded.

Once we have a full ensemble we then take the covariance of the ensemble to find our $\mathbf{B}_{corr}$. In figure 6 we have plotted the correlation matrix or normalised error covariance matrix of $\mathbf{B}_{corr}$. We see that this matrix includes both positive and negative correlations between parameter and state variables, with correlations of 1 down the diagonal between variables of the same quantity as expected. The largest positive off-diagonal correlation being between f_{lab} and C_{lab}, this makes physical sense as f_{lab} is the parameter controlling the amount of GPP allocated to the labile carbon pool, C_{lab}.

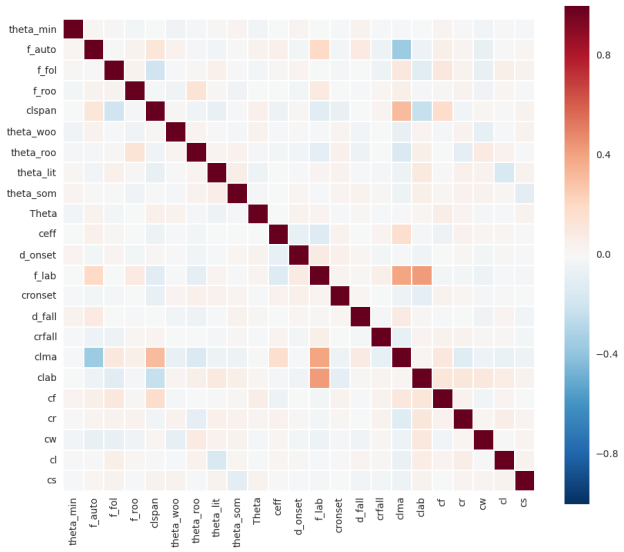

Figure 6: Background error correlation matrix created using method in section 2.5

2.6 Specifying serial correlations in the observational error covariance matrix

In this paper we use observations of NEE averaged daily. The flux towers actual sampling time is half-hourly, we take the sum over the 48 measurements made each day. We only select days

where there is no missing data and over 90% of observations pass quality control tests. Errors in NEE observations come from different sources such as instrument errors, sampled ecosystem structure and turbulent conditions (when we have low turbulence and limited air mixing NEE is underestimated) [Papale et al., 2006]. Due to this dependance on atmospheric conditions we expect the errors in observations of NEE to be serially correlated, as the atmospheric signal itself is serially correlated [Daley, 1992]. If we were assimilating half hourly observations of NEE we would expect stronger correlations between observation errors, as atmospheric conditions are more constant at this time scale, with correlations between observation errors getting weaker with lower frequency observations.

In section 2.3 we have re-written the 4D-Var cost function in equation 10 in order to allow the specification of serial observation error correlations in our assimilation scheme. These serial correlations are represented by the off-diagonal terms of $\hat{\mathbf{R}}$. As we do not have a practical method of estimating the serial correlation in NEE observation error we adapt the simple Gaussian model found in Järvinen et al. [1999] (a second order autoregressive correlation function was also tested but not presented here). Here the correlation r between 2 observations at times t_1 and t_2 is given as,

$$r = \begin{cases} a\exp\left[\frac{-(t_1-t_2)^2}{\tau^2}\right] + (1-a)\delta_{t_1-t_2} & |t_1 - t_2| \leq \eta \\ 0 & \eta < |t_1 - t_2| \end{cases}, \tag{18}$$

where τ is the e-folding time, a is the correlation, δ is the Kronecker delta and η is the cut off time after which the correlation between two observation errors is zero. We have incorporated a cut off for correlations between observation errors as we believe the correlation length scale to be quite short for our assimilated observations but also because this makes $\hat{\mathbf{R}}$ better conditioned and therefore easier to invert in the assimilation process.

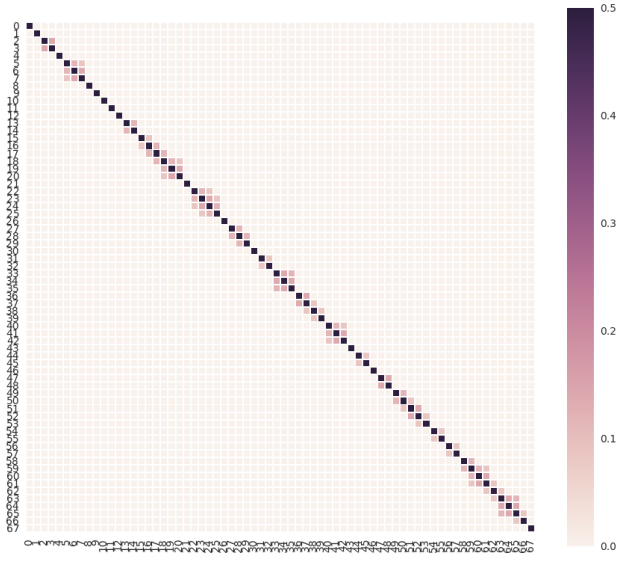

Figure 7: Observation error covariance matrix created using method in section 2.6 with $\tau = 4$, $a = 0.3$ and $\eta = 4$.

In figure 2.6 we show $\hat{\mathbf{R}}$ created using equation 18, we can see that we have 67 NEE observations in our years assimilation window, these observations are obviously not all on adjacent days and this is evident in the structure of $\hat{\mathbf{R}}$. We also see the effect of the short e-folding time chosen here

($\tau = 4$) giving us the desired structure as correlations are believed to be on the scale of a day.

2.7 Experiments

In this paper we will present the results of 4 experiments where we vary the representations of $\mathbf{B}$ and $\hat{\mathbf{R}}$ while assimilating the same NEE observations in the years window from 1999-2000. These experiments are outlined in table 1 where $\mathbf{B}_{diag}$ and $\hat{\mathbf{R}}_{diag}$ are the diagonal matrices of the parameter-state variances and the observations variances respectively and $\mathbf{B}_{corr}$ and $\hat{\mathbf{R}}_{corr}$ are the matrices as specified in section 2.5 and section 2.6 respectively.

Experiment	$\mathbf{B}_{diag}$	$\hat{\mathbf{R}}_{diag}$	$\mathbf{B}_{corr}$	$\hat{\mathbf{R}}_{corr}$
A	×	×		
B		×	×	
C	×			×
D			×	×

Table 1: The combination of error covariance matrices used in each data assimilation experiment.

In section 3 we will show the results of each of the experiments and see the impact that using these different representations of $\mathbf{B}$ and $\hat{\mathbf{R}}$ has on our assimilation scheme.

3 Results

In all the following experiments we present results when assimilating NEE observations in the period 1999-2000 from the Alice Holt flux site, other assimilation windows were also tested but are not shown here.

3.1 Experiment A

In this experiment $\mathbf{B}_{diag}$ and $\hat{\mathbf{R}}_{diag}$ were used in our assimilation as described in section 2.7. This experiment will form the standard by which the subsequent results from assimilation experiments will be judged.

In figur 8 we have plotted assimilation and forecast results for NEE. We can see that our analysis trajectory (green line) fits well with the observations during the the assimilation window (1999-2000) and then diverges in the forecast (2000-2014). One reason for this is that our analysis is being over constrained by the observations and under constrained by the background. Using a diagonal background error covariance matrix ($\mathbf{B}_{diag}$) we are not adding much background information to our system and therefore the observations dominate and result in the over fitting of our analysis to the assimilated observations of NEE.

To see how well our forecast performs after assimilation we show a scatter plot of modelled NEE against observed NEE in figure 13. Here we have a Root-Mean-Square Error (RMSE) of 4.22gCm^{-2} and a bias of -0.3gCm^{-2} for our forecast of NEE, whereas our analysis (1999-2000) has a RMSE of 1.36gCm^{-2} and a bias of -0.03gCm^{-2}. The background trajectory meanwhile has a RMSE of 3.86gCm^{-2} and a bias of -1.60gCm^{-2} in the analysis window (1999-2000) and the same RMSE of 3.86gCm^{-2} but a bias of -1.36gCm^{-2} during the forecast period (2000-2014). So we see that although using $\mathbf{B}_{diag}$ and $\hat{\mathbf{R}}_{diag}$ in our assimilation has considerably reduced the RMSE in our analysis period, it has also increased the RMSE in our forecast of NEE. However it has reduced the bias in the models forecast considerably from -1.36gCm^{-2} to -0.3gCm^{-2}. The bias in our

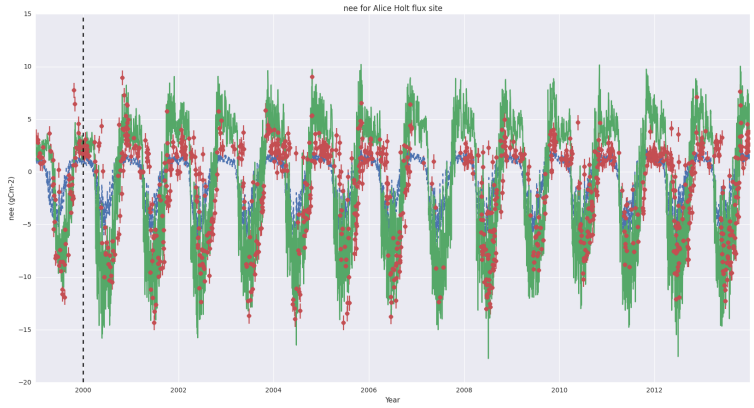

Figure 8: One year assimilation and fourteen year forecast of Alice Holt NEE with DALEC2, blue dotted line: background model trajectory, green line: analysis and forecast after assimilation, red dots: observations from Alice Holt flux site with error bars. Number of function evaluations needed for minimisation to converge: 571.

Figure 9: Forecast scatter plot of modelled NEE vs. observations for 2000-2014 (green dots), forecasted NEE has a root-mean squared error of 4.22gCm^{-2} between model and observations. Blue line represents 1-1 line where all green dots would lie if the model predicted the observations perfectly.

background comes from a constant under prediction of the more extreme negative values of NEE and this leads to considerably worse results than our analysis and its forecast for total forest carbon uptake.

Maybe include something about the estimated reduction in error for parameter and state variables using the diagonal terms of **B and **A** (analysis error covariance matrix). How best to compare and present these?

3.2 Experiment B

Here $\mathbf{B}_{corr}$ (as defined in section 2.5) and $\hat{\mathbf{R}}_{diag}$ are used in our assimilation. In figure 10 we can see that our forecast performs considerably better than in experiment A and indeed from figure 11 we see that our forecasts RMSE has almost halved (now 2.56gCm^{-2}) with a reduction in bias also, now -0.2gCm^{-2}. In comparison using $\mathbf{B}_{corr}$ in our assimilation slightly degrades the fit for our analysis (1999-2000), with a RMSE of 1.42gCm^{-2} and a bias of -0.04gCm^{-2}, this is because our

assimilation scheme is now more constrained by the background than in experiment A. Therefore using $\mathbf{B}_{corr}$ in our assimilation reduces the problem of overfitting to our assimilated observations of NEE as seen in experiment A. Another improvement made by using $\mathbf{B}_{corr}$ in our assimilation is that our minimisation routine converges to a solution more quickly, taking 218 fewer function iterations than experiment A to converge.

Figure 10: One year assimilation and fourteen year forecast of Alice Holt NEE with DALEC2, blue dotted line: background model trajectory, green line: analysis and forecast after assimilation, red dots: observations from Alice Holt flux site with error bars.Number of function evaluations needed for minimisation to converge: 353.

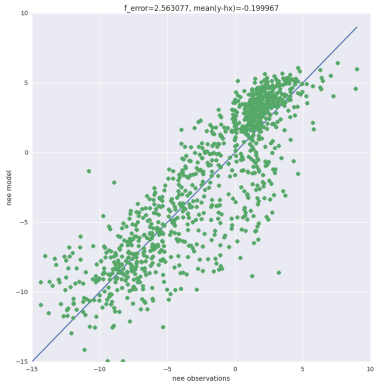

Figure 11: Forecast scatter plot of modelled NEE vs. observations for 2000-2014 (green dots), forecasted NEE has a root-mean squared error of 2.56gCm^{-2} between model and observations. Blue line represents 1-1 line where all green dots would lie if the model predicted the observations perfectly.

3.3 Experiment C

Here we use $\mathbf{B}_{diag}$ and $\hat{\mathbf{R}}_{corr}$ (as defined in section 2.6) in our assimilation. We can see from figure 12 that our results look similar to those in section 3.1 however there are some differences. Using $\hat{\mathbf{R}}_{corr}$ in our assimilation has also reduced our minimisation routines convergence time, taking 127 less function iterations to converge than experiment A. We also have a slight reduction in RMSE for our forecast (now 4.09gCm^{-2}) from experiment A. As in experiment B our analysis gets slightly worse

as using $\hat{\mathbf{R}}_{corr}$ acts to reduce the overfitting of the analysis to the assimilated observations. The changes seen when using $\hat{\mathbf{R}}_{corr}$ in our assimilation are less extreme than when using $\mathbf{B}_{corr}$ as the correlations we have specified in $\hat{\mathbf{R}}_{corr}$ are quite small, we have therefore added less information to our system when using $\hat{\mathbf{R}}_{corr}$ in our assimilation. We would expect the effect to be larger if using a model with a half-hourly time step assimilating half-hourly observations of NEE as we would then have much stronger correlations in $\hat{\mathbf{R}}_{corr}$ (as discussed in section 2.6). We also expect that specifying these correlations in $\hat{\mathbf{R}}$ will help when assimilating other less frequently sampled data streams along with NEE as the serial correlations reduce the weight given to the mean of the observations and also reduce the information content of the data streams with more observations [Daley, 1992, Järvinen et al., 1999].

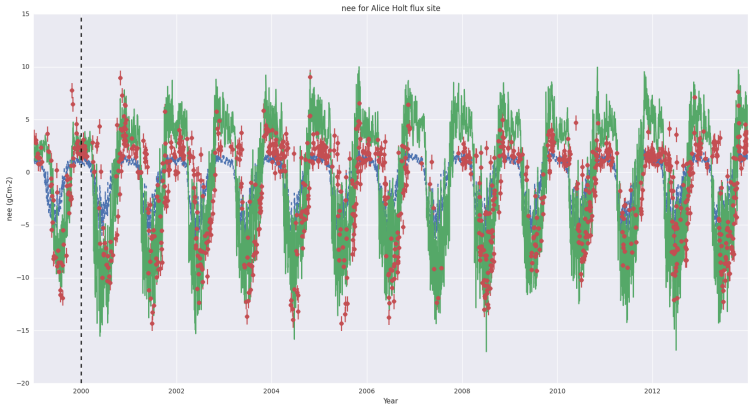

Figure 12: One year assimilation and fourteen year forecast of Alice Holt NEE with DALEC2, blue dotted line: background model trajectory, green line: analysis and forecast after assimilation, red dots: observations from Alice Holt flux site with error bars. Number of function evaluations needed for minimisation to converge: 444.

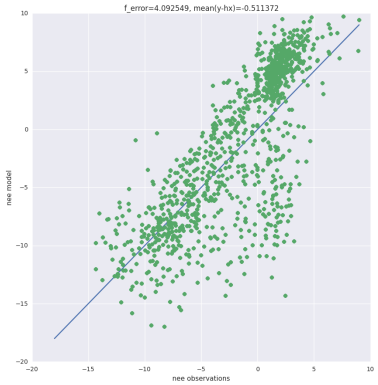

Figure 13: Forecast scatter plot of modelled NEE vs. observations for 2000-2014 (green dots), forecasted NEE has a root-mean squared error of 4.09gCm^{-2} between model and observations. Blue line represents 1-1 line where all green dots would lie if the model predicted the observations perfectly.

3.4 Experiment D

In the final experiment we use $\mathbf{B}_{corr}$ and $\hat{\mathbf{R}}_{corr}$ in our assimilation. From figure 14 we see that using both correlated matrices gives similar results as experiment B when $\mathbf{B}_{corr}$ is used with $\hat{\mathbf{R}}_{diag}$. However using $\hat{\mathbf{R}}_{corr}$ in addition to $\mathbf{B}_{corr}$ provides similar improvements as in experiment C. The number of function evaluations taken for our minimisation routine to converge is the lowest yet at 316. Also our forecast RMSE is reduced again still from results in experiment B to 2.38gCm^{-2}. Using both matrices appears to combine the beneficial effects described in both section 3.2 and section 3.3.

From our results in experiment D the forecasted (2000-2014) total carbon uptake for our research site is 5.04×10^6kgC (try to find better units for this?). In comparison our less accurate results from experiment A (using $\mathbf{B}_{diag}$ and $\hat{\mathbf{R}}_{diag}$) predict a total carbon uptake of 5.26×10^6kgC, a difference of 2.25×10^5kgC. This is quite a substantial difference as we are considering the carbon uptake of a small (~ 0.93km^2) research site only.

Figure 14: One year assimilation and fourteen year forecast of Alice Holt NEE with DALEC2, blue dotted line: background model trajectory, green line: analysis and forecast after assimilation, red dots: observations from Alice Holt flux site with error bars. Number of function evaluations needed for minimisation to converge: 316.

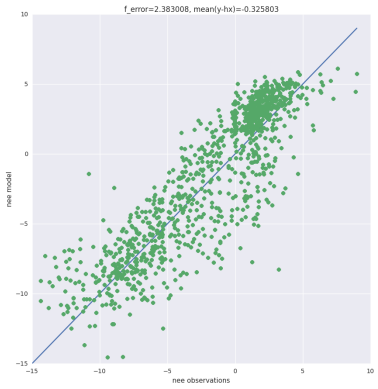

Figure 15: Forecast scatter plot of modelled NEE vs. observations for 2000-2014 (green dots), forecasted NEE has a root-mean squared error of 2.38gCm^{-2} between model and observations. Blue line represents 1-1 line where all green dots would lie if the model predicted the observations perfectly.

4 Discussion

- In this paper we have implemented the DALEC2 functional ecology model in a 4D-Var data assimilation scheme, building an adjoint of the DALEC2 model and applying rigorous tests to our scheme. Using 4D-Var can provide much quicker assimilation results than MCMC techniques as we have knowledge of the derivative of the model, however we do also assume that the problem is Gaussian whereas MCMC techniques do not. We have shown that 4D-Var is a valid tool for improving the DALEC2 model estimate of NEE.

- We then considered the nature of background and observational errors. The effect of specifying parameter-state correlations in our background information and serial correlations between our observation errors was explored.

- The technique presented here to specify $\mathbf{B}_{corr}$ has been shown to significantly improve forecasts of NEE. Other techniques were also tested (not presented here) to create a correlated $\mathbf{B}$, one of which added correlations to $\mathbf{B}$ using the DALEC2 model. Here we evolved an ensemble of state vectors over the length of the chosen assimilation window, then taking the covariance of the evolved ensemble. This gave us a $\mathbf{B}$ with parameter-state and state-state correlations, but no parameter-parameter correlations as the parameters are not updated by the model. Using the $\mathbf{B}$ created with this method also improved assimilation results considerably in some cases. Many different tests were run using different background vectors and variances and it was found that specifying some form of correlations in $\mathbf{B}$ always made some improvement to the results of our assimilation.

- Here the $\hat{\mathbf{R}}_{corr}$ used in our experiments has improved our forecast of NEE, however many other choices of $\hat{\mathbf{R}}_{corr}$ tested for this paper degraded the forecast. This is probably due to specifying unrealistic correlations and suggests a more diagnostic approach is needed for the calculation of serial correlations in $\hat{\mathbf{R}}$. One option would be to adapt the Deroziers diagnostic [Desroziers et al., 2005], which has been used successfully in NWP for diagnosing observation error correlations for observations taken at the same time, and extending this technique to diagnose serial correlations.

- Specifying serial correlations also allows us to address the issues discussed by Richardson et al. [2010], that when assimilating multiple data streams more frequently sampled observation types (such as NEE) have much more impact on the assimilation than data streams sampled less frequently. Specifying serial correlations between observations of the same type has the effect of reducing the weight given to the mean of the observations [Järvinen et al., 1999], thus allowing less frequent data streams to have more impact on the assimilation.

  Using the form of $\hat{\mathbf{R}}$ given in this paper for specifying serial correlations will also allow us to specify serial correlations between different observation types. When running the model with a day-night time step this will allow us to build in the type of correlations investigated by Baldocchi et al. [2015] between ecosystem respiration and canopy photosynthesis.

- We have also seen an improvement in convergence times for our minimisation routine when specifying correlations in our $\mathbf{B}$ and $\hat{\mathbf{R}}$ matrices, this has been the case in most other tests not presented here. This is due to the correlated matrices constraining the assimilation more and making the problem less ill-posed.

5 Conclusion

- 4DVar is a valid tool for improving the DALEC2 model estimate of NEE.
- Including correlations in the background error covariance matrix improves our forecast after assimilation (sometimes greatly) in comparison to using a diagonal representation of $\mathbf{B}$.
- Specifying serial correlations between observation errors in $\hat{\mathbf{R}}$ can improve our forecast, however more work is needed to find a way to properly diagnose these correlations.

References

C. Bacour, P. Peylin, N. MacBean, P. J. Rayner, F. Delage, F. Chevallier, M. Weiss, J. Demarty, D. Santaren, F. Baret, D. Berveiller, E. Dufrêne, and P. Prunet. Joint assimilation of eddy-covariance flux measurements and FAPAR products over temperate forests within a process-oriented biosphere model. *Journal of Geophysical Research: Biogeosciences*, pages n/a–n/a, 2015. ISSN 21698953. doi: 10.1002/2015JG002966. URL `http://doi.wiley.com/10.1002/2015JG002966`.

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.

Dennis Baldocchi, Cove Sturtevant, and Fluxnet Contributors. Does day and night sampling reduce spurious correlation between canopy photosynthesis and ecosystem respiration? *Agricultural and Forest Meteorology*, 207:117–126, 2015. ISSN 01681923. doi: 10.1016/j.agrformet.2015.03.010. URL `http://linkinghub.elsevier.com/retrieve/pii/S016819231500088X`.

A. A. Bloom and M. Williams. Constraining ecosystem carbon dynamics in a data-limited world: integrating ecological "common sense" in a modeldata fusion framework. *Biogeosciences*, 12(5):1299–1315, 2015. ISSN 1726-4189. doi: 10.5194/bg-12-1299-2015. URL `http://www.biogeosciences.net/12/1299/2015/`.

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.

Philippe Ciais, Christopher Sabine, Govindasamy Bala, Laurent Bopp, Victor Brovkin, Josep Canadell, Abha Chhabra, Ruth DeFries, James Galloway, Martin Heimann, et al. Carbon and other biogeochemical cycles. In *Climate change 2013: the physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, pages 465–570. Cambridge University Press, 2014.

Roger Daley. The Effect of Serially Correlated Observation and Model Error on Atmospheric Data Assimilation, 1992. ISSN 0027-0644.

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.

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.

Jean-françois Exbrayat, T Luke Smallman, A Anthony Bloom, and Mathew Williams. Using a data-assimilation system to assess the influence of fire on simulated carbon fluxes and plant traits for the Australian continent. *EGU General Assembly*, 17:6421, 2015.

Mike Fisher. Background error covariance modelling. In *Seminar on Recent Development in Data Assimilation for Atmosphere and Ocean*, pages 45–63, 2003.

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.

Heikki Järvinen, Erik Andersson, and François Bouttier. Variational assimilation of time sequences of surface observations with serially correlated errors. *Tellus A*, 51(4):469–488, 1999.

Eugenia Kalnay. *Atmospheric modeling, data assimilation, and predictability*. Cambridge university press, 2003.

T. Kaminski, W. Knorr, G. Schürmann, M. Scholze, P. J. Rayner, S. Zaehle, S. Blessing, W. Dorigo, V. Gayler, R. Giering, N. Gobron, J. P. Grant, M. Heimann, a. Hooker-Stroud, S. Houweling, T. Kato, J. Kattge, D. Kelley, S. Kemp, E. N. Koffi, C. Köstler, P. P. Mathieu, B. Pinty, C. H. Reick, C. Rödenbeck, R. Schnur, K. Scipal, C. Sebald, T. Stacke, a. Terwisscha Van Scheltinga, M. Vossbeck, H. Widmann, and T. Ziehn. The BETHY/JSBACH Carbon Cycle Data Assimilation System: Experiences and challenges. *Journal of Geophysical Research: Biogeosciences*, 118 (4):1414–1426, 2013. ISSN 21698961. doi: 10.1002/jgrg.20118.

G. Krinner, Nicolas Viovy, Nathalie de Noblet-Ducoudré, Jérôme Ogée, Jan Polcher, Pierre Friedlingstein, Philippe Ciais, Stephen Sitch, and I. Colin Prentice. A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system. *Global Biogeochemical Cycles*, 19(1):1–33, 2005. ISSN 08866236. doi: 10.1029/2003GB002199.

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

Andrew C Lorenc and F Rawlins. Why does 4d-var beat 3d-var? *Quarterly Journal of the Royal Meteorological Society*, 131(613):3247–3257, 2005.

Jorge Nocedal and Stephen J Wright. *Numerical Optimization*. Springer Science & Business Media, 1999. ISBN 0387987932. URL `http://books.google.co.uk/books/about/Numerical_Optimization.html?id=epc5fX0lqRIC&pgis=1`.

D. Papale, M. Reichstein, M. Aubinet, E. Canfora, C. Bernhofer, W. Kutsch, B. Longdoz, S. Rambal, R. Valentini, T. Vesala, and D. Yakir. Towards a standardized processing of Net Ecosystem Exchange measured with eddy covariance technique: algorithms and uncertainty estimation. *Biogeosciences*, 3(4):571–583, 2006. ISSN 1726-4189. doi: 10.5194/bg-3-571-2006.

Florence Rabier. Overview of global data assimilation developments in numerical weather-prediction centres. *Quarterly Journal of the Royal Meteorological Society*, 131(613):3215–3233, 2005.

Andrew D Richardson, Mathew Williams, David Y Hollinger, David JP Moore, D Bryan Dail, Eric A Davidson, Neal A Scott, Robert S Evans, Holly Hughes, John T Lee, et al. Estimating parameters of a forest ecosystem c model with measurements of stocks and fluxes as joint constraints. *Oecologia*, 164(1):25–40, 2010.

K. Singh, M. Jardak, a. Sandu, K. Bowman, M. Lee, and D. Jones. Construction of non-diagonal background error covariance matrices for global chemical data assimilation. *Geoscientific Model Development*, 4(2):299–316, 2011. ISSN 1991959X. doi: 10.5194/gmd-4-299-2011.

Laura M. Stewart, Sarah L. Dance, and Nancy K. Nichols. Data assimilation with correlated observation errors: Experiments with a 1-D shallow water model. *Tellus, Series A: Dynamic Meteorology and Oceanography*, 65(1):1–14, 2013. ISSN 02806495. doi: 10.3402/tellusa.v65i0.19546.

Hans Verbeeck, Philippe Peylin, Cédric Bacour, Damien Bonal, Kathy Steppe, and Philippe Ciais. fluxes in Amazon forests: Fusion of eddy covariance data and the ORCHIDEE model. *Journal of Geophysical Research*, 116(G2):1–19, 2011. ISSN 0148-0227. doi: 10.1029/2010JG001544.

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.

JM Zobitz, AR Desai, DJP Moore, and MA Chadwick. A primer for data assimilation with ecological models using markov chain monte carlo (mcmc). *Oecologia*, 167(3):599–611, 2011.

Appendix

Write appendix!