State-Parameter Estimation of Ecosystem Models Using a Smoothed Ensemble Kalman Filter

M. Chen[a], S. Liu[a], and L. Tieszen[b]

[a]Science Applications International Corp. (SAIC)[*], U.S. Geological Survey (USGS) Center for Earth Resources Observation and Science (EROS), Sioux Falls, South Dakota, USA 57198
[b]USGS/EROS, Sioux Falls, South Dakota, USA 57198

[*]Work performed under USGS contract 03CRCN0001

Abstract: Much of the effort in data assimilation methods for carbon dynamics analysis has focused on estimating optimal values for either model parameters or state variables. The main weakness of estimating parameter values alone (i.e., without considering state variables) is that all errors from input, output, and model structure are attributed to model parameter uncertainties. On the other hand, the accuracy of estimating state variables may be reduced if the temporal evolution of parameter values is not incorporated. This research develops a smoothed ensemble Kalman filter (SEnKF) to estimate simultaneously the system states and model parameters of an eddy flux partition model. The approach is used to assimilate observed fluxes of carbon and major driving forces at an AmeriFlux forest station: Howland, Maine, USA. The aim of applying a kernel-smoothing algorithm to an ensemble Kalman filter is to overcome the dramatic, sudden change of parameter values in time and the loss of continuity between two consecutive points in time. Our analysis demonstrates that model parameters, such as light use efficiency, respiration coefficients, minimum and optimum temperatures for photosynthetic activity, and so on, are highly constrained by eddy flux data at daily-to-seasonal time scales. The SEnKF stabilizes parameter values quickly regardless of the initial values of the parameters. Potential ecosystem light use efficiency demonstrates a strong seasonality. Results show that the simultaneous parameter estimation procedure significantly improves model predictions. Results also show that the SEnKF can dramatically reduce variance in state variables stemming from the uncertainty of parameters and driving variables. The SEnKF is a robust and effective algorithm in evaluating and developing ecosystem models and in improving understanding and quantification of carbon cycle parameters and processes.

Keywords: Gross primary production; ecosystem respiration; net ecosystem exchange; smoothed ensemble Kalman filter; AmeriFlux data

1. INTRODUCTION

Inherent limitations exist in the measurement and modelling of ecosystem carbon dynamics. Measurement is usually patchy in space and discontinuous in time, while modelling is always built on some principles with assumptions and imperfectly defined parameters. Advanced data assimilation techniques based on statistics or optimization theory can overcome these limitations by combining a series of measurements with dynamic models. So far, much of the effort in data assimilation methods for carbon dynamics analysis has focused on estimating optimal values for either model parameters [Braswell et al., 2005] or state variables [Bond-Lamberty et al., 2005]. Methods that focus on estimating parameter values alone (i.e., without considering state variables) generally minimize long-term prediction error by using a historical batch of data that assumes time-invariant parameters. The procedures used to process the historical data as a whole lack the flexibility to investigate possible temporal evolution of the model parameters. Although there is now some attempt to partition data into a number of subsets in time order, the partition is inevitably subjective [Reichstein et al., 2005b]. In particular, the main weakness is that all errors from input, output, and model structure are attributed to model parameter uncertainties. Sequential data assimilation procedures such as the ensemble Kalman filter (EnKF) have the potential to overcome this drawback by explicitly taking all sources of uncertainty into account [Evensen, 2003; Nichol et al., 2002]. However, the successful application of the EnKF primarily focuses on estimating time-varying state variables under the typical presumption that the parameters are to be specified in advance. Because the ecosystem is too complex to guarantee that the model parameters do not change over time, model adjustment through the temporal variation of parameters with the state variables is desirable.

We have developed a smoothed ensemble Kalman filter (SEnKF) to estimate simultaneously system states and model parameters of a simple carbon cycle model. The aim of applying a kernel-smoothing algorithm to an ensemble Kalman filter is to overcome the dramatic, sudden change of parameter values in time and the loss of continuity between two consecutive points in time.

2. METHODS

2.1. SEnKF

The SEnKF is a sequential data assimilation method that includes three components: (1) a dynamic model used to forecast estimates, (2) observation data and the relationship between the data and the model state used to update the estimates, and (3) an assimilation scheme for model-data synthesis [Evensen and Van Leeuwen, 2000; Evensen, 2003; Raupach et al., 2005].

2.1.1. Dynamic Model

A dynamic model can be expressed as a discrete-time nonlinear stochastic process:

$$X^{k+1} = f(X^k, U^k, \theta^k) + \varepsilon^k \qquad (1)$$

where k denotes time step, X^k is a vector of random state variables or object variables (such as carbon flux, store attributes and related entities), f is the model operator as a propagation of model state (such as rates of change about net carbon fluxes), U^k is a set of externally specified time-dependent forcing variables (such as meteorological variables and soil properties), θ^k is a set of model parameters or auxiliary variables (such as light use efficient and partition ratios), and the noise term ε^k accounts for both imperfections in model formulation and stochastic variability in forcing variables and parameters.

To extend the applicability of the EnKF to simultaneous state-parameter estimation, we need to build an evolution of the parameters similar to that of the state variables:

$$\theta^{k+1} = g(\theta^k) + \tau^k \qquad (2)$$

Where g is a transition operator (such as linear function, $g(\theta^k) = \theta^k$), and τ is an random error. We will discuss their definitions below. Now we define $Y^k = (X^k, \theta^k)^T$, $M = (f, g)^T$, and $q^k = (\varepsilon^k, \tau^k)^T$, where T denotes transposition. Then (1) and (2) are changed into a standard state model

$$Y^{k+1} = M(Y^k, U^k) + q^k \qquad (3)$$

2.1.2. Observation Data

The observation (Z^k) is related to the system state, external forcing variables, and parameters through the expression of the form

$$Z^k = H(X^k, U^k, \theta^k) + \delta^k \qquad (4)$$

or

$$Z^k = H(Y^k, U^k) + \delta^k \qquad (5)$$

where the operator H specifies the deterministic relationship between the observation data and the model states. The noise term δ^k accounts for both measurement error (instrumental and processing errors in the measurements) and representation error (errors in the model representation of Z, introduced by shortcomings in the observation model H), which is assumed to be Gaussian and independent of model error.

2.1.3. Assimilation Scheme

The EnKF is based on the Monte Carlo method and the Kalman filter formulation and mimics the probability distribution of the model state conditioned on a series of observations of the model state. The probability density of the model state is represented by a large ensemble of model states, and these are integrated forward in time by the model with a stochastic forcing term representing the model errors [Evensen, 1994]. Each ensemble member evolves in time according to

$$Y_{j-}^{k+1} = M(Y_{j+}^k, U_j^k), \quad j = 1, \cdots, N \qquad (6)$$

where N denotes the number of model state ensemble members, Y_{j-}^{k+1} is the component of the jth ensemble member forecast at time k+1 and Y_{j+}^k is the jth updated ensemble member at time k. The noise term is not explicitly represented because the EnKF represents multiplicative model errors through forcing data perturbation [Evensen,

1997]. The forcing data perturbations are made by adding white noise (subject to Gaussian distribution with zero mean and covariance Q_u^k) to forcing data at each time step:

$$U_j^k = U^k + \eta_j^k,\ \eta_j^k \sim N(0, Q_u^k) \qquad (7)$$

Now we discuss how to build an evolution of the parameters similar to that of the state variables. The conventional artificial parameter evolution, which adds small random perturbation at each time step, results in over-dispersion of parameter samples and loss of continuity between two consecutive points in time. We used the kernel smoothing of parameter samples to remedy the problem according to [West, 1993]

$$\begin{cases} \theta_{j-}^{k+1} = a\theta_{j+}^k + (1-a)\bar{\theta}_+^k + h\tau_j^k \\ \tau_j^k \sim N(0, V_+^k),\ \bar{\theta}_+^k = \text{mean}(\theta_{j+}^k), \\ V_+^k = \text{var}(\theta_{j+}^k),\ a^2 + h^2 = 1 \end{cases} \qquad (8)$$

where a is the shrinkage factor in (0,1) of the kernel location, which is typically around 0.45–0.49, h is the smoothing or variance reduction parameter, θ_{j-}^{k+1} is the component of the jth ensemble member forecast at time k+1, and θ_{j+}^k is the component of the jth updated ensemble member at time k.

Similarly, observation data are treated as random variables by generating an ensemble of observations from a distribution with the mean equal to the measurement value and a covariance equal to the estimated measurement error [Williams et al., 2005].

$$Z_j^{k+1} = Z^{k+1} + \delta_j^{k+1}, \delta \sim N(0, R^{k+1}) \qquad (9)$$

Because the true state is generally unknown, we calculate an ensemble covariance matrix to substitute definitions of the error covariance matrix in the Standard Kalman filter. The forecasted ensemble error covariance is calculated according to

$$P_-^{k+1} = \frac{1}{N-1}[M_Y^{k+1} - \bar{M}_Y^{K+1}][M_Y^{k+1} - \bar{M}_Y^{k+1}]^T \qquad (10)$$

where $M_Y^k = [Y_{1-}^k - \bar{Y}_{1-}^k, \cdots, Y_{N-}^k - \bar{Y}_{N-}^k]$ and $\bar{M}_Y^k = \frac{1}{N}\sum_{j=1}^N Y_{j-}^k$.

The updated scheme of the EnKF is as follows:

$$Y_{j+}^{k+1} = Y_{j-}^{k+1} + K^{k+1}(Z_j^{k+1} - H(Y_{j-}^{k+1}, U_j^{k+1})) \qquad (11)$$

where K^{k+1} is Kalman gain

$$K^{k+1} = P_-^{k+1}H^T(HP_-^{k+1}H^T + R^{k+1})^{-1} \qquad (12)$$

2.2. Application of SEnKF to C Modelling

2.2.1. Flux Partition Model

We use a "flux partition model" as our test dynamic model for the SEnKF method. Our selection is based on two considerations. First, it is an important model for constructing bottom-up estimates of continental carbon balance components. Second, it is appropriate for testing robustness of the SEnKF method because it is nonlinear, there are sufficient observations of state varaiables, and it has multiple unknown parameters. The model partitions net ecosystem exchange (NEE) into gross primary production (GPP) and total ecosystem respiration (RESP) as follows:

$$\begin{cases} NEE_t = GPP_t - RESP_t \\ GPP_t = LUE_t \cdot PAR_t \cdot NDVI_t \cdot D_{temp} \cdot D_{VPD} \\ RESP_t = R_{ref,t} \exp[E_0(\frac{1}{T_{ref,t} - T_0} - \frac{1}{T_{air,t} - T_0})] \end{cases} \qquad (13a–c)$$

where subscript t denotes time-dependent, LUE_t is light use efficiency, PAR_t is photosysnthetic active radiation, $NDVI_t$ is the normalized difference vegetation index, $R_{ref,t}$ is respiration when air temperature ($T_{air,t}$) equals reference temperature ($T_{ref,t,}$ usually specified as 10 ^{o}C), E_0 is temperature sensitivity, and T_0 is a datum of temperature to avoid a denominator of zero in the model (13c), kept constant at –46.02 ^{o}C as in Reichstein et al., 2005a. D_{temp} determines the effect of temperature on photosynthesis, and D_{VPD} expresses the decrease in leaf exchange from both photosynthesis and transpiration due to vapour pressure deficit (VPD), according to

$$D_{temp} = \max\left[\frac{(T_{\max} - Tair)(Tair - T_{\min})}{(T_{\max} - Tair)(Tair - T_{\min}) + (T_{opt} - Tair)^2}, 0\right] \qquad (14)$$

$$D_{vpd} = 0.5\left[1 + \frac{1}{1 + v_0 \exp(v_1 VPD)}\right] \qquad (15)$$

where T_{min}, T_{opt}, and T_{max} denote minimum, optimal, and maximum temperatures for photosynthesis, respectively, VPD is vapour pressure deficiency, and v_0 and v_1 are two unknown coefficients.

If we define state and driving force vectors as $Y_t = (NEE_t, GPP_t, RESP_t, LUE_t,$

$$T_{\min}, T_{opt}, T_{\max}, \nu_0, \nu_1, R_{ref}, E_0)$$

and $U_t = (T_t, PAR_t, VPD_t, NDVI_t)$, then the model can be expressed in the form of (6).

2.2.2. Flux Data

Eddy flux estimates of net ecosystem exchange (NEE) are based on the covariance of high frequency fluctuations in vertical wind velocity and CO_2 concentration [Baldocchi et al., 1988]. We applied the SEnKF approach for the AmeriFlux station in Howland, Maine, USA. The period was from 2000 to 2004 because there were sufficient hourly and daily data at the station and NDVI data from the Moderate Resolution Imaging Spectroradiometer (MODIS) were not available before 2000. Field observations included hourly observations of NEE, humidity, photosynthetically active radiation (PAR), air temperature, air pressure, wind speed, and daily precipitation data. In our analysis, we used daily estimates of three state variables (NEE, GPP, and RESP) and four driving force variables (air temperature, PAR, VPD, and NDVI). NEE was directly downloaded from the AmeriFlux Web station. RESP was calculated from the temperature dependence curve of ecosystem respiration derived from nighttime NEE observations. GPP was calculated as a total of NEE and RESP (13a). Gaps in carbon exchange and meteorological data were filled using multivariable nonlinear regression. Daily NDVI was calculated using linear interpolation of the MODIS 16-day composites. We assumed that data errors were subject to a Gaussian distribution with a zero mean and a variance of 20% of the average data based on uncertainty analysis in eddy covariance measurements [Hollinger and Richardson, 2005]. The transition operator (*H*) in (4) was taken as a 3 x N linear matrix with elements of 1 at diagonal nodes and 0 at other nodes.

To evaluate the performance of the SEnKF method, we created a "base" model run by obtaining an optimal set of parameters for the flux partition model using the conventional nonlinear inversion procedures in the statistical analytical software (SAS). However, this set of parameters was generated from a mixture of 15 AmeriFlux stations, covering various ecosystem types. Therefore, these parameter values were not necessarily optimal for the particular station used in this study (i.e., Howland). We refer to this model as the base model in this paper.

To test the predictive power of the SEnKF, we held 80% of the data for model validation. Data assimilation was performed on only 1/5 of the observations.

3. RESULTS

Figure 1 shows the results of SEnKF data assimilation and the comparison with the base

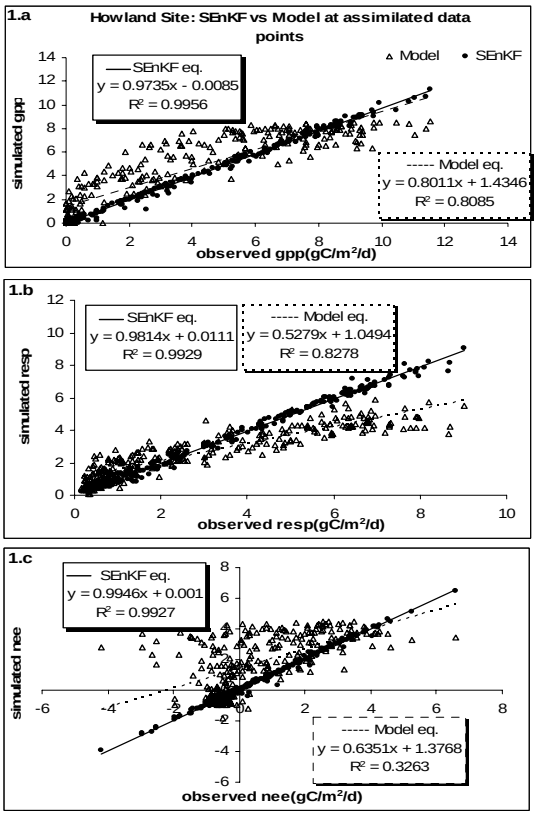

Figure 1. Comparison of estimates of GPP, RESP, and NEE generated by the SEnKF and the base model (20% of the data assimilated).

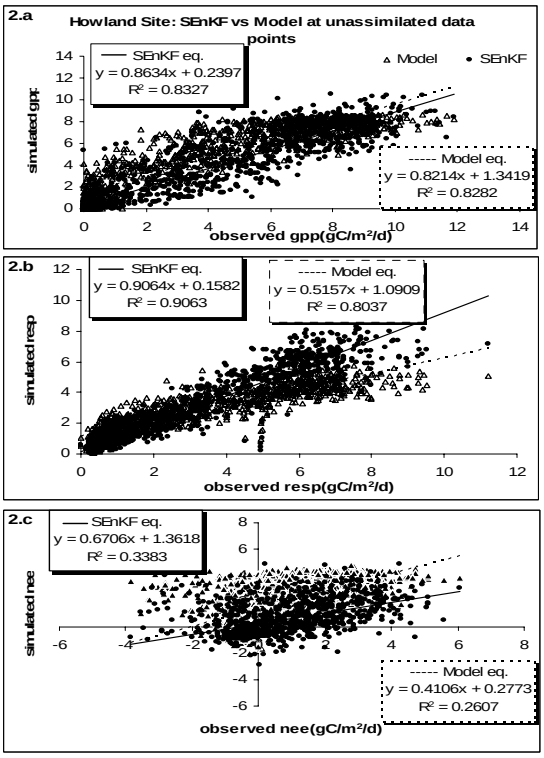

Figure 2. Forecasted values of the model modified by the SEnKF and the base model against unassimilated data.

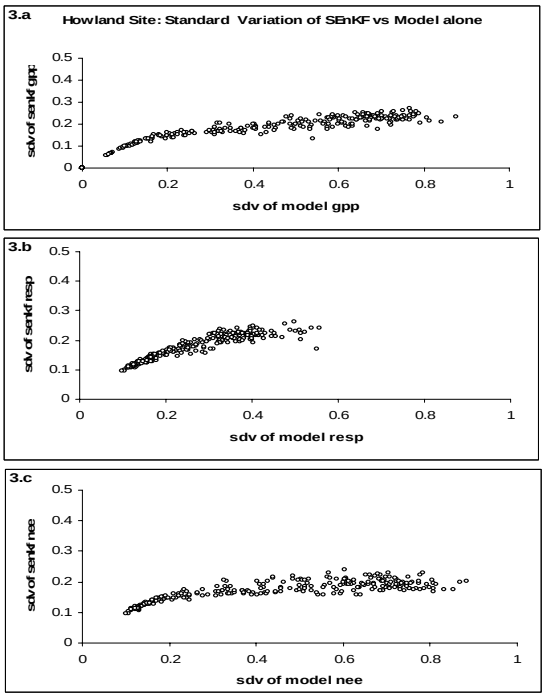

Figure 3. Comparison of ensemble variances of GPP, RESP, and NEE generated by the SEnKF and the base model. The results show the SEnKF can more dramatically reduce variances of state variables than the ensemble based only on Monte Carlo technique.

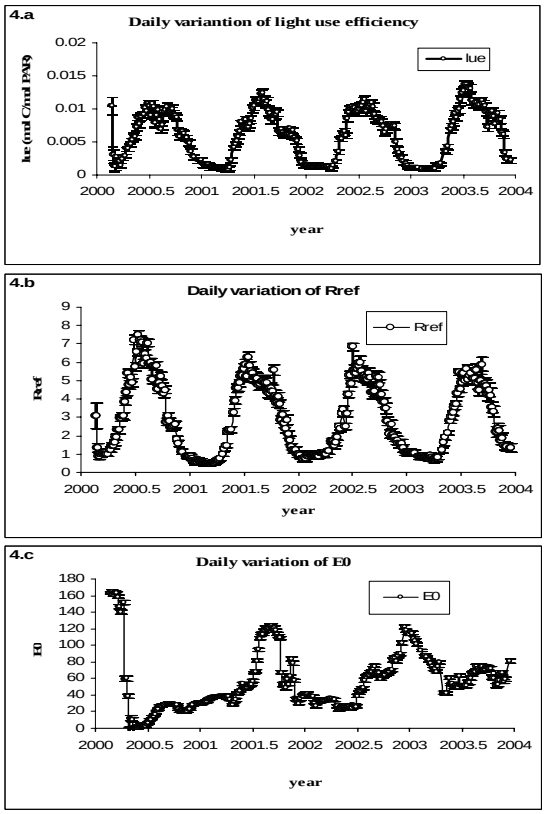

Figure 4. Temporal variations of three key parameters in the "flux partition model."

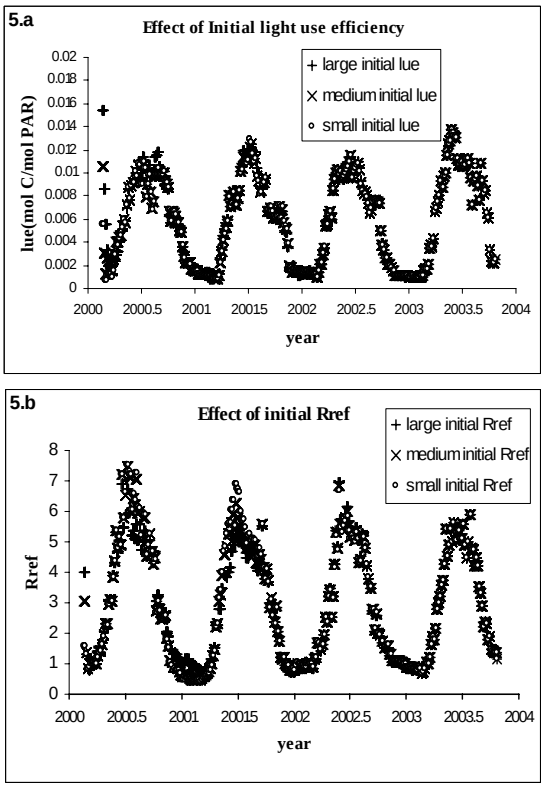

Figure 5. Stabilization of parameters by the SEnKF. Note the difference in the initial parameter values and the speedy convergence.

model (20% of total observations). It can be seen that SEnKF dramatically improved the estimation of ecosystem states compared with the base model. Of course, we have to see if the new parameter values derived from the SEnKF can be used to improve the prediction of system conditions. Based on the performance of the SEnKF on the data that were held for validation (i.e., 80% of the data), we see that the estimates of the three state variables using the SEnKF matched more closely with observations than those using the base model did. The SEnKF dramatically reduced the uncertainty stemming from parameters and driving forces, especially when uncertainty was high (Fig. 3). The SEnKF also revealed that the parameter values demonstrated a strong seasonality or temporal variability (Fig. 4). The temporal change of parameter values was relatively smooth because a smoothing procedure was implemented in the SEnKF to control the over-dispersion of parameter sampling.

The SEnKF can quickly stabilize the parameter values regardless of the initial values of the parameters (Fig. 5). This demonstrates the robustness of the SEnKF. In future work, we plan to analyse the collinearity of parameter values so that we can evaluate the confounding effects of parameters in the data assimilation process.

4. CONCLUSIONS

The SEnKF method greatly improves state estimates of the flux partition model and dramatically reduces uncertainties stemming from parameters and driving forces. Simultaneous parameter estimation can use near real-time observations to improve the prediction capability of dynamic models. The model based on the SEnKF can be used to fill data gaps in observations. This research demonstrates that the SEnKF is a robust and effective algorithm for evaluating and developing ecosystem models and improving understanding and quantification of the carbon cycle parameters and processes.

5. ACKNOWLEDGEMENTS

Work performed by M. Chen and S. Liu was performed under USGS contract 03CRCN0001. Wenping Yuan, from the Chinese Academy of Sciences, participated in the analysis of the flux data while he was visiting the U.S. Geological Survey (USGS) Center for Earth Resources Observation and Science (EROS) in 2005. The research was funded by the Earth Surface Dynamics Program and the Geographic Analysis and Monitoring Program of the USGS.

6. REFERENCES

Baldocchi, D.D., B.B. Hicks, and T.P. Meyers, Measuring biosphere-atmosphere exchanges of biologically related gases with micrometeorological methods, *Ecology, 69*(5), 1331–1340, 1988.

Bond-Lamberty, B., C. Wang, and S.T. Gower, Spatiotemporal measurement and modeling of stand-level boreal forest soil temperatures, *Agricultural and Forest Meteorology, 131*(1–2), 27–40, 2005.

Braswell, B.H., W.J. Sacks, E. Linder, and D.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.

Evensen, G., Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte-Carlo methods to forecast error statistics, *Journal of Geophysical Research-Oceans, 99*(C5), 10143–10162, 1994.

Evensen, G., Advanced data assimilation for strongly nonlinear dynamics, *Monthly Weather Review, 125*(6), 1342–1354, 1997.

Evensen, G., and P.J. Van Leeuwen, An ensemble Kalman smoother for nonlinear dynamics, *Monthly Weather Review, 128*(6), 1852–1867, 2000.

Evensen, G., The ensemble Kalman filter: theoretical formulation and practical implementation, *Ocean Dynamics, 53*(4), 343–367, 2003.

Hollinger, D.Y., and A.D. Richardson, Uncertainty in eddy covariance measurements and its application to physiological models, *Tree Physiology, 25*(7), 873–885, 2005.

Nichol, C.J., J. Lloyd, O. Shibistova, A. Arneth, C. Roser, A. Knohl, S. Matsubara, J. and Grace, Remote sensing of photosynthetic-light-use efficiency of a Siberian boreal forest, *Tellus: Series B, 54*(5), 677, 2002.

Raupach M. R., P.J. Rayner, D.J. Barrett, R.S. DeFries, M. Heimann, D.S. Ojima, S. Quegan, and C.C. Schmullius, Model-data synthesis in terrestrial carbon observation: methods, data requirements and data uncertainty specifications, *Global Change Biology, 11*(3), 378–397(20), 2005.

Reichstein, M., E. Falge, D. Baldocchi, D. Papale, M. Aubinet, P. Berbigier, C. Bernhofer, N. Buchmann, T. Gilmanov, A. Granier, T. Grünwald, K. Havránková, H. Ilvesniemi, D. Janous, A. Knohl, T. Laurila, A. Lohila, D. Loustau, G. Matteucci, T. Meyers, F.

Miglietta, J.M. Ourcival, J. Pumpanen, S. Rambal, E. Rotenberg, M. Sanz, J. Tenhunen, G. Seufert, F. Vaccari, T. Vesala, D. Yakir, and R. Valentini, On the separation of net ecosystem exchange into assimilation and ecosystem respiration: review and improved algorithm, *Global Change Biology*, *11*(9), 1424–1439, 2005a.

Reichstein, M., E. Falge, D. Baldocchi, D. Papale, M. Aubinet, P. Berbigier, C. Bernhofer, N. Buchmann, T. Gilmanov, A. Granier, T. Grünwald, K. Havránková, H. Ilvesniemi, D. Janous, A. Knohl, T. Laurila, A. Lohila, D. Loustau, G. Matteucci, and T. Meyers, On the separation of net ecosystem exchange into assimilation and ecosystem respiration: review and improved algorithm, *Global Change Biology*, *11*(9), 1424–1439, 2005b.

West, M., Mixture models, Monte Carlo, Bayesian updating and dynamic models, *Computing Science and Statistics*, *24*, 325–333, 1993.

Williams, M., P.A. Schwarz, B.E. Law, J. Irvine, and M.R. Kurpius, An improved analysis of forest carbon dynamics using data assimilation, *Global Change Biology*, *11*(1), 89–105, 2005.