Fox, Andrew; Williams, Mathew; Richardson, Andrew D.; **Cameron, David**; Gove, Jeffrey H.; Quaife, Tristan; Ricciuto, Daniel; Reichstein, Markus; Tomelleri, Enrico; Trudinger, Cathy M.; Van Wijk, Mark T.. 2009 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. 1597-1615.

Using model-data fusion to characterise confidence in parameterizations, analyses and forecasts of terrestrial C dynamics

Andrew Fox[1], Mathew Williams[2*], Andrew D. Richardson[3], David Cameron[4], Jeffrey H. Gove[5], Tristan Quaife[6], Daniel Ricciuto[7], Markus Reichstein[8], Enrico Tomelleri[8], Cathy Trudinger[9] and Mark T. Van Wijk[10]

[1] School of Applied Maths, Centre for Terrestrial Carbon Dynamics, University of Sheffield, Sheffield, UK

[2] School of GeoSciences, Centre for Terrestrial Carbon Dynamics, University of Edinburgh, Edinburgh, UK

[3] Complex Systems Research Center, University of New Hampshire, Durham, NH, USA

[4] Centre for Ecology and Hydrology, Bush Estate, Penicuik, Midlothian, UK

[5] USDA Forest Service, Northern Research Station, Durham, NH, USA

[6] Centre for Terrestrial Carbon Dynamics, Department of Geography, UCL, London

[6] Oak Ridge National Laboratory, Oak Ridge, TN, USA

[8] Max Planck Institute for Biogeochemistry, Jena, Germany

[9] CSIRO Marine and Atmospheric Research, Centre for Australian Weather and Climate Research, Aspendale, Victoria, Australia

[10] Wageningen University, Plant Sciences, The Netherlands

Running title: Intercomparison of carbon dynamics model-data fusion techniques

Key words: Data assimilation, metropolis, carbon cycle, ecosystem modelling, Monte Carlo, Kalman Filter, eddy covariance, Reflex project

* Correspondence to Mathew Williams, mat.williams@ed.ac.uk, +44 131 667 8631 (tel) +44 131 662 0478 (fax)

Abstract

We describe a model-data fusion inter-comparison project (REFLEX), aimed at comparing the strengths and weaknesses of various model-data fusion algorithms for estimating parameters, states and fluxes of a simple ecosystem carbon cycle model. Participants were provided with both synthetic net ecosystem exchange (NEE) of CO_2 and leaf area index (LAI) data, generated from a simple C model with added noise, and observed NEE and LAI data from two eddy covariance observations sites within FLUXNET. Participants endeavoured to estimate model parameters and states for all cases over the two years for which data were provided, and generate predictions for one additional year without observations. Nine participants contributed results using Metropolis algorithms, Kalman filters and a genetic algorithm. For the synthetic data case, parameter estimates compared well with the true values. The results of the analyses indicated that parameters linked directly to gross primary production and ecosystem respiration, such as those related to foliage allocation and turnover, or temperature sensitivity of heterotrophic respiration, were best constrained and characterised. Estimates of confidence intervals varied among algorithms, but several algorithms successfully located the true values of annual fluxes from synthetic experiments within relatively narrow 90% confidence intervals, achieving >80% success rate and mean NEE confidence intervals <110 gC m^{-2} yr^{-1} for the synthetic case. For the observed data case, the annual C flux estimates generally agreed with gap-filling approaches using half-hourly data. The estimation of gross fluxes, by partitioning daily NEE data, agreed well with outputs from earlier studies using half-hourly data. The study was revealing in that confidence limits on annual NEE was 88% larger in the prediction year, than in the previous year, when data were available. Confidence intervals on annual NEE also increased by 30% when observed data were used instead of synthetic data, reflecting and quantifying the addition of model error. Finally, our analyses indicated that incorporating additional constraints, using data on

large, slow C pools (wood and soil) would help to reduce uncertainties for model parameters poorly served by eddy covariance data.

Introduction

The carbon cycle is a critical determinant of the Earth‟s climate, but the carbon-climate relationship is complicated by feedbacks between the climate, the terrestrial biosphere and the atmosphere (Heimann and Reichstein 2008). Recent model inter-comparisons have shown that there are significant differences among model predictions of the future C cycle at decadal timescales (Friedlingstein et al. 2006). The causes of these differences among models are not well understood, but are likely to be related to subtle differences in process representation, which can have significant impacts over longer time scales.

Model-data comparison provides an opportunity to highlight areas (in space or time) of poor process representation, and to guide model improvement. Thus, the modelling community is now seeking to test its terrestrial ecosystem models against the growing array of observations (Bonan 2008). One of the critical datasets to be used in evaluating ecosystem models is the FLUXNET database (Baldocchi et al. 2001). FLUXNET is an international network of eddy covariance (EC) flux measurement towers. There are data sets from hundreds of sites worldwide, some with more than a decade of data collection. However, these data are associated with uncertainties and complications. There are gaps in time series that must be filled to obtain integrated (daily to annual) flux sums (Moffat et al. 2007). Also large areas of the globe are poorly sampled, and measurements are affected by systematic and random errors (Lasslop et al. 2008, Richardson et al. 2008), both of which can be large. EC towers measure net ecosystem exchanges (NEE) of CO_2, meaning that the underlying processes of photosynthesis (GPP) and ecosystem respiration (R_e) are not directly measured during daytime (Desai et al. 2008).

A meaningful comparison between models and data is complicated by the need to assess and account for both model and observational errors. Thus, the probability of a model being correct should be assessed by taking into account observational uncertainties. When

comparing a model against multiple datasets, then weighting the confidence one has in the different observations becomes critical (Raupach et al. 2005). Model uncertainty is also an important factor in any comparison with data. Models may be uncertain because of how they represent key processes, how initial conditions are set, or because their parameters are poorly determined. Separating these causes of uncertainty is important for guiding model development.

Model-data fusion approaches, previously used mainly in hydrology and weather forecasting, are now being used more frequently by the terrestrial C community (Raupach et al. 2005). Model-data fusion (MDF) combines models with observations, taking account of model and observational uncertainties. In theory, MDF provides a means to cope with the problems arising from incomplete and noisy observational data, and uncertainty in model processes, initial states and parameters. MDF combines models with observations, and estimates of their uncertainties, to produce estimates of system dynamics with confidence intervals (Williams et al. 2005) and to determine model parameterizations consistent with data. We refer to these outputs of MDF schemes as "analyses" hereafter. MDF can be used as a developmental tool to test hypotheses and then improve model structural representation (Sacks et al. 2007, Stockli et al. 2008, Moore et al. in press). However, in practice, MDF results are conditional both on the model and data used, as well as associated uncertainties and assumptions made about uncertainties.

The capabilities and weaknesses of the various existing MDF approaches remain poorly understood. One recent study, the OptIC experiment, used pseudo-data from a highly simplified test model with 4 parameters to compare parameter estimation methods (Trudinger et al. 2007). OptIC found different methods equally successful, but that the choice of the cost function (quantifying the model-data mismatch) caused the most variation in the estimated parameters. OptIC also demonstrated that the effort expended and experience of the user was

106 a factor in successful solutions. However, OptIC did not use observed data, nor did it test state
107 estimation or model forecast capabilities. With observed data, MDF is complicated by
108 observational and model error and bias.

109 Here we describe the REgional FLux Estimation eXperiment. REFLEX is a model-data
110 fusion inter-comparison project, aimed at comparing the strengths and weaknesses of various
111 MDF algorithms for estimating carbon model parameters and carbon fluxes and states.
112 REFLEX participants fuse an existing C model with observed and synthetic daily NEE data.
113 The key question addressed here is: what are the confidence intervals on model parameters
114 calibrated from eddy covariance (EC) data, and on model analyses and estimates and
115 predictions of net C exchange and carbon stocks over multiple years? The experiment has an
116 explicit focus on how different algorithms and protocols quantify the confidence intervals on
117 parameter estimates and model forecasts, given the same C model and a range of datasets.
118 In generating analyses and predictions of C dynamics with confidence intervals, resulting
119 error can be attributed to a combination of factors (Liu and Gupta 2007). Errors may be
120 related to the particular MDF algorithm employed (for instance, does the algorithm find local
121 or global minima) - the algorithmic error - and the choice of subjective components of the
122 MDF process, including prior assumptions about the probability distributions of parameters
123 and initial conditions – the user error. Error may also be related to the observations, as a
124 function of instrumental precision. And models may contain errors, due to misrepresented or
125 missing fundamental processes. Driver error (i.e. meteorological forcing) is likely to be small
126 in local studies, but is increasingly important at coarser scales due to representativeness, the
127 extent to which a point measurement can represent the surrounding area. The structure of
128 Reflex allowed investigation of several of these components of error.
129 In REFLEX, participants first used synthetic data, generated from the specified C model with
130 noise and gaps added, to explore the capabilities of a range of users and algorithms to retrieve

parameters and states consistent with the C model. This synthetic experiment dealt with observational and algorithmic error, and user error including assumptions related to initial conditions and parameter priors. There was no model error or driver error. REFLEX participants then went on to fuse data from eddy covariance systems and local measurements of leaf area index (LAI) with the C model. This exercise introduced model and, to a lesser extent, driver error, because the model used does not perfectly describe the forest ecosystem, and because meteorological observations may contain small errors. Finally, REFLEX participants used the C model in a prognostic, rather than diagnostic, mode. One year of daily driver data were provided to produce forecasts of C dynamics, using parameters generated in the diagnoses, and the forecasts were tested against withheld data, both synthetic and observed.

What is novel in this study is an explicit focus on comparing how an ensemble of MDF algorithms perform in terms of estimating C model states and parameters, and the uncertainties on these quantities. By using a single common model, and both synthetic and observed data sets, and diagnostic and prognostic tests, we are able to generate insights into current capabilities for assessing and forecasting ecosystem C dynamics using the model-data fusion approach

Methods

Model Description

The requirements for the Reflex C model included simplicity, a C mass balance, a daily time step, and vegetation and soil C pools with time constants covering days to decades. The model outputs had to include daily NEE and LAI. We selected the Data Assimilation Linked Ecosystem Carbon (DALEC) model (Williams et al. 2005), originally designed for evergreen forests, and a modified version (DALEC-D) for deciduous forests (Figure 1). DALEC is a

simple box model of carbon pools connected via fluxes running at a daily time-step. For the evergreen model there are five C pools representing foliage (C_f), woody stems and coarse roots (C_w), and fine roots (C_r) along with fresh leaf and fine root litter (C_{lit}) and soil organic matter and coarse woody debris (C_{som}). In the deciduous model there is an additional labile pool (C_{lab}). The following assumptions were made to determine the fluxes between the C pools:

1. All C fixed during a day is expended either in autotrophic respiration or else allocated to one of the three plant tissue pools, C_f, C_w or C_r.
2. Autotrophic respiration is a constant fraction of the C fixed during a day (Waring et al. 1998).
3. Allocation fractions to vegetation pools are donor-controlled functions which have constant rate parameters.
4. For the deciduous model, the timing of initial leaf-out is controlled by a simple growing degree day accumulation, and leaf-fall by a minimum temperature threshold. The maximum amount of C that can be allocated to leaves is also limited by a parameter (c_{fmax})
5. All C losses are via mineralization (i.e. no dissolved losses).

The aggregated canopy model (ACM) (Williams et al. 1997) is used to calculate daily GPP in DALEC. ACM is a „big leaf", daily time-step model that estimates GPP using a simple aggregated set of equations operating on cumulative or average values of leaf area index (LAI), foliar nitrogen, total daily irradiance, minimum and maximum daily temperature, day length, atmospheric CO_2 concentration, water potential gradient (ψ_d) and total soil-plant hydraulic resistance (r_{tot}). ACM contains 10 parameters which have been calibrated using a fine-scale model (the Soil-Plant-Atmosphere model (SPA), (Williams et al. 1996) across a wide range of driving variables producing a „universal" parameter set which maintains the

essential behaviour of the fine-scale model but at a much reduced complexity. The sole ACM parameter included in the optimisation is the nitrogen use efficiency parameter (a_1), which determines the maximum rate of carboxylation per g foliar N. For the purposes of this experiment the sites were treated as being non-drought stressed. Those variables related to drought effects in ACM, specifically ψ_d and r_{tot}, were given a fixed value in accordance with this assumption.

Datasets

Four datasets (two synthetic and two based on actual measurements) were provided to participants. Each dataset included a variety of information (Table 1), including continuous daily meteorological drivers, intermittent NEE and LAI data, estimates of the initial values of the pools of soil organic matter and wood, and input data on leaf characteristics for the GPP model (Table 2). Initial conditions for foliar, fine root, litter and labile C were not provided.

Synthetic datasets were generated for three years for an evergreen (EV-SYN) and deciduous (DE-SYN) forest, using DALEC and DALEC-D model runs, with nominal parameters, and meteorological driver data selected from European eddy covariance flux tower sites. Gaps were introduced into the synthetic NEE and LAI data time series by thinning the model outputs to match the data availability from the real data. Noise was added to the remaining model outputs to reflect measurement error, by adding Gaussian errors with a variance of 0.5 g C m^{-2} d^{-1} for the NEE and 10% of the truth for the LAI. Though the half hourly measurements may have non-Gaussian errors, once you start aggregating at longer time scales the noise on the sum/mean becomes Gaussian. Participants were provided with the first two years of synthetic observations.

For the observed data, the sites (Loobos, Netherlands and Hesse, France) and site-years (2000-2) were selected on the basis of relatively long, continuous records of fluxes and site meteorology, good quality control, and little or no drought stress. The observed data included

eddy covariance (EC) data, LAI data, and local meteorological data from a deciduous broad-leaf forest (identified as DE-EC) and an evergreen needle-leaf forest (EV-EC). Daily NEE was calculated by summing half-hourly observations, but only if >43 of the possible 48 observations passed quality control. Missing data were filled by the daily mean of remaining data. It is possible that some small bias was introduced by this simple gap-filling, but for the purposes of this study such impacts were deemed insignificant. Typical data coverage was 20-30% of days. LAI data were sparse, usually collected on just a few days. Gap-filled flux data were not used in this experiment, but complete daily meteorological data were required to drive the model, and so gap filled weather data were used. All data were obtained via the FLUXNET site (www.fluxnet.ornl.gov), from relevant, site specific literature and/or from site PIs. Three sequential years of data were assembled, of which the first two years were provided to participants. The source of the EC data was withheld from participants.

Experiments

All participants used DALEC and DALEC-D, the same models used to generate the synthetic data. The use of common reference models allowed direct comparison among MDF algorithms. Upper and lower bounds for the parameters of both deciduous and evergreen versions of the model were provided (Table 3). These bounds were set broad to ensure a high likelihood that reasonable parameters were located in the EC experiments. Participants applied the MDF algorithm of their choice to four experiments (
Table 4). The first two experiments were diagnostic, testing parameter and state estimation using two years of incomplete daily NEE and LAI data, at both an evergreen and deciduous site. These data were either real, collected at a FLUXNET site (experiment 1) or artificial, synthesised from model output with added noise (experiment 2). The final two experiments were prognostic, testing forecast capability, again at the real sites (experiment 3) and the artificial sites (experiment 4). Forecasts of daily C fluxes and pool dynamics were generated

using parameter distributions from the first two experiments, forced by a single extra year of meteorological data. The flux/stock data, both observed and synthetic, for this third year were withheld for later assessment.

Algorithms

A wide range of different MDF algorithms are currently applied (e.g. Raupach et al. 2005). They range from relatively simple Monte Carlo and grid-search approaches in which limited numbers of parameters can be estimated (Van Wijk and Bouten 2002, Williams et al. 2006); local optimization algorithms like the Levenberg–Marquardt algorithm or the Gauss-Newton algorithm (Janssen and Heuberger 1995, Trudinger et al. 2007, Wang et al. 2007, Van Wijk et al. 2008); generic search algorithms that in principle can deal with large numbers of parameters like Genetic Algorithms (Van Wijk and Bouten 2001); an algorithm based on the Metropolis-Hastings algorithm (Metropolis et al. 1953) and using Markov Chain Monte Carlo samplers that recently has become popular (Vrugt et al. 2003, Braswell et al. 2005, Knorr and Kattge 2005, Van Oijen et al. 2005, Ricciuto et al. 2008); and finally algorithms like the Kalman filter that can combine parameter estimation with state updating ((Vrugt et al. 2005, Williams et al. 2005, Quaife et al. 2008, Trudinger et al. 2008). A key element in all of these approaches is the quantification of the uncertainty of the parameters, which requires that uncertainty in the measurements is quantified (Hagen et al. 2006, Richardson et al. 2006, Richardson et al. 2008).

Reflex was an open intercomparison experiment, so the algorithms employed were not selected according to any criteria, rather they were dependent upon the community interest and experience (Table 5). Many of the methods used Monte Carlo approaches based on the Metropolis-Hastings algorithm or variants thereof. There were differences in the implementation, with various cost functions, uncertainty specifications and convergence tests employed. The cost function weights the difference between observations and simulated

quantities, often using observation error estimates, and sometimes model error estimates. There was also a genetic algorithm approach, and an Ensemble Kalman Filter (EnKF). In two cases a Metropolis approach was supplemented by a Kalman filter (one Unscented KF, one EnKF). All the algorithms (bar the free-standing EnKF) used ~10^5 iterations to produce the full set of parameter and state estimates. Most of the algorithms assumed that prior parameter distributions were uniform across the range supplied. The use of a uniform prior suggests that the researcher has a prior belief that all setting of parameters within the range are equally likely. The users made a variety of assumptions about initial conditions for some state variables.

Analyses

To quantify and summarise these different approaches for parameter assessment, we computed for each parameter two (EC) or three (SYN) relative-distance metrics, d_1-d_3. Here, for a given parameter, m_x is algorithm x's best estimate of the parameter; CI_x is the width of the parameter's confidence interval for algorithm x; t is the true value of the parameter; p_{max} and p_{min} are the pre-specified upper and lower limits on the parameter (Table 3); σ is a standard deviation and μ is a mean:

d_1. Consistency among algorithms: $\sigma(m_1,\ldots,m_9)/(p_{max}-p_{min})$

d_2. CI constrained by the data: $\mu(CI_1,\ldots,CI_9)/(p_{max}-p_{min})$

d_3. Consistent with truth (SYN only): $|t-\mu(m_1,\ldots,m_9)|/(p_{max}-p_{min})$

Then the total distance is: $D = \sqrt{d_1^2 + d_2^2 + d_3^2}$ for SYN and $D = \sqrt{d_1^2 + d_2^2}$ for EC datasets; the closer the value D is to zero, the better the parameter is estimated, according to this measure.

We determined two further metrics to aid a comparison among algorithms of parameter estimation capabilities, for the SYN cases only. Mean normalised parameter confidence

interval (d_4) is similar to the d_2 statistic but rates individual algorithm‟s mean 90% confidence intervals across all parameters, normalised by the size of the parameter priors:

d_4: $$\left(\sum_{=} \frac{CI_i}{p_{\max i} - p_{\min i}} \right) n$$

where CI_i is the width of the algorithm‟s 90% confidence interval for parameter x; n is the number of parameters (11 for EV, 17 for DE), $p_{\max i}$ and $p_{\min i}$ are the pre-specified upper and lower limits on each parameter prior.

The metric for consistency with true parameter value (d_5) is similar to the d_3 statistic, but again rates consistency for an individual algorithm across all parameters:

d_5. $$t - \left(\sum_{=} \frac{m_i}{p_{\max i} - p_{\min i}} \right) n \qquad |t - \mu(m_1, \ldots, m_n)| / (p_{\max} - p_{\min})$$

where m_x is parameter x‟s best estimate by the algorithm.

Results

Parameter estimation

Each algorithm produced sets of parameter estimates for each dataset in experiments 1 and 2, describing a multi-dimensional probability density volume. Because of their high dimensionality, these hyper-volumes are not easily described or visualised, so a range of metrics and methods are used. Firstly, we determined the “best” parameter set estimate of each algorithm (Figure 2), based on the minimum of the cost function (e.g. Metropolis algorithm) or the mean value of an ensemble (Ensemble Kalman filter). The best estimates were supplemented by estimates of the 90% confidence intervals on each parameter, determined from the full spread of accepted parameters.

Some parameters were well constrained (Figure 2), so that the analysis resulted in a much reduced spread in the parameter compared to the upper and lower bounds that defined the

prior (Table 3). Conversely, some parameters were poorly constrained, with little reduction in spread from the initial upper and lower bounds. In some cases there was consistency among algorithms in the estimates of the parameter best estimates, but not in others. For the SYN datasets only, it was possible to gauge how effectively the algorithms retrieved the true parameter estimates.

The parameter analysis for both synthetic data (Table 6) and eddy covariance data (Table 7) revealed that, for the both evergreen and deciduous models, turnover rate parameters such as T_s, T_l, and T_f, as well as the temperature parameter E_t, were well estimated overall, across the range of methods. By comparison, the turnover rate parameters T_d and T_w, as well as the allocation parameter F_{nrr}, tended to be poorly estimated overall. In some cases the poor estimates were scattered around the truth (T_w estimate in DE-SYN) while in others there was a clear bias in the algorithms‟ estimates (T_w estimate in EV-SYN). The allocation parameter F_{nf} was well-estimated for EV-SYN but biased in DE-SYN. Of those parameters used only in the deciduous model, F_{ll} and T_{lab} were poorly estimated, whereas F_{lr} was more successfully estimated.

There was a significant correlation of d_1 distances between EC and SYN for EV (r=0.73, P=0.01) but not DE (r=0.31, P=0.24). So the EV parameters that were consistently estimated (across methods) were similar for synthetic and eddy covariance data, while this was not so for DE datasets, perhaps because of the greater number of parameters. There was a significant correlation between EC and SYN d_2 distances for both EV (r=0.87, P=0.0004) and DE (r = 0.84, P<0.0001). Thus parameters that were well constrained (low d_2) by the synthetic data were well constrained by the eddy covariance data.

There was a general trend for those algorithms with large parameter confidence intervals to encompass a large fraction of true parameter values within their 90% confidence intervals (Figure 3). For the DE case, three algorithms (E1, E2, M1) managed to generate relatively

small and reliable confidence intervals. But the for EV case, none of the algorithms managed to balance small confidence intervals with reliability. For the DE case, three algorithms (E1, E2, M1) generated parameters that were most consistent with true values and also had the smallest confidence intervals. For the EV case there was no clear pattern among algorithms; although E2 had the closest agreement with true parameters and the narrowest confidence intervals, it had the smallest fraction of true parameters within the 90% CI, suggesting over-confidence.

Eigenvector analyses of the error covariance matrices were used to supplement the parameter analyses, and these suggested that the best constrained parameter was the turnover rate of SOM, T_s. The next best constrained parameter identified was the temperature rate parameter, E_t. Turnover rate of foliage was well constrained for EV analyses. Allocation to and turnover of roots were poorly constrained for EV analyses. The results for the DE eigenvector analyses were less clear, with differences between DE-EC and DE-SYN. Turnover rate of wood and roots were least well constrained in DE-SYN, while the GDD threshold for leaf out and the turnover rate of labile C were least well constrained in DE-EC. There was some variation in the eigenvectors from the different methods, due to variation in covariance matrices. Some parameters seemed to be well constrained by some methods, but not by others. Comparison of eigenvectors with the distance metric d_2 were largely, but not totally, consistent. Eigenvector analyses did not identify any consistent correlation features, apart from one between fraction of GPP respired (F_g) and the NUE parameter, P_r.

Flux estimates – synthetic data

The seasonal patterns of variation in NEE were generally well reproduced by most algorithms across all three years of each of the different data sets (for example, Figure 4). For the synthetic datasets, where true values were known, daily NEE predictions were generally good. RMSE values ranged from 0.07-0.55 gC m^{-2} d^{-1}, with a mean over all algorithms and

years of 0.20 gC m^{-2} d^{-1}. These error values compared well with the noise added to the truth in order to generate synthetic observations. Partitioning synthetic NEE into GPP and R_e was generally successful, with mean RMSE values over all algorithms of 0.6 gC m^{-2} d^{-1} in both cases. There was no evidence that best-fit or mean predictions of fluxes deteriorated in year 3, the prognostic period when no data were assimilated.

Flux estimates – observed data

For the eddy covariance (EC) datasets, the algorithms‟ predictions were compared to observed NEE. In years 1 and 2, when observations were provided to participants, RMSEs varied from 0.7-1.8 gC m^{-2} d^{-1} (DE) or 0.6-0.9 gC m^{-2} d^{-1} (EV), with a mean value of 1.3 gC m^{-2} d^{-1} for DE datasets and 0.7 gC m^{-2} d^{-1} for EV. In year 3, when observations were not provided to participants, RMSEs varied from 1.1-2.3 gC m^{-2} d^{-1} (DE) or 1.3-1.7 gC m^{-2} d^{-1} (EV), with a mean value of 1.5 gC m^{-2} d^{-1} for both EC and DE datasets (

Table 9). Thus the best NEE estimates of the algorithms tended to agree less well in the prognostic period (year 3) compared to the assimilation period (years 1-2), though this was most striking for the evergreen (EV) case in this study.

Flux confidence intervals – daily data

There was less agreement among algorithms in the assessment of 90% confidence intervals (CI) on daily fluxes (Figure 4) than in the assessment of best estimates. There were differences in confidence interval estimates both in magnitudes and in temporal variability among algorithms. For instance, the mean daily 90% CI varied among algorithms from 0.35 - 1.92 gC m^{-2} d^{-1} in DE-SYN and 0.29 – 2.49 gC m^{-2} d^{-1} in DE-EC. Algorithm confidence intervals typically had large excursions during spring leaf out for DE, but the magnitude of these excursions varied (Figure 4).

We tested whether the 90% CI on daily analyses (years 1 and 2) and predictions (year 3) encompassed the truth from the synthetic datasets for NEE, GPP and R_e for all years, and for observed NEE in year 3 for the EC datasets. The days of each year which passed this test were counted. We expected that 85-95% of the days would pass, roughly consistent with the magnitude of the confidence interval, 90%. For the synthetic experiments (NEE tests are shown in

Table 8) this was rarely the case. In some cases the fraction was 100%, which indicates that the daily CI were likely set too large. In other cases, the fractions were <85% suggesting that the CI were too small or the predictions were biased. For the eddy covariance datasets in year 3, the majority of algorithms‟ confidence intervals on daily NEE were too narrow, with an average of only 40% (DE) or 20% (EV) of the observed year 3 data lying within the 90% confidence interval (

Table 9). This result suggests the algorithms were over-confident in the assessments of daily fluxes.

Flux confidence intervals – annual sums

A comparison of 90% confidence intervals on annual estimates of NEE, GPP and R_e for all years revealed differences of up to an order of magnitude in width (Figure 5, Figure 6, Figure 7). There was no clear relationship between size of CI and algorithm type – for instance, M1 and M2 tended to have small CI compared to M3 and M4, although all used Metropolis algorithms. This result makes clear the importance of the user in determining the confidence interval, rather than the algorithm itself. The mean confidence interval size for NEE (124 gC m^{-2} yr^{-1}) was ~3-fold smaller than those for GPP (389 gC m^{-2} yr^{-1}) and R_e (387 gC m^{-2} yr^{-1}). A comparison of the mean 90% confidence intervals on annual NEE estimates (Table 10) indicated that CI were largest during year 3, the prediction period, and smallest in year 2. Of the 36 cases (4 datasets □ 9 algorithms), 34 had larger confidence intervals on year 1 than

year 2, and 35 had larger CI on year 3 than year 2, so this pattern was general across algorithms and datasets. Averaged over all cases, the 90% CI in the prediction period (year 3) were 88% larger than in the second year of the assimilation period (year 2). Patterns were similar in comparison between outputs from observed and synthetic datasets. However, mean 90% CI across all algorithms were ~31% larger for EC datasets than for SYN datasets. Among algorithms, the increase in 90% CI on EC datasets compared to SYN datasets ranged from 0% (E1) to 100% (E2).

Testing annual flux estimates and confidence intervals

Annual flux outputs estimated and forecast using the synthetic datasets were compared with the synthetic truth. Each algorithm‟s annual output of NEE, GPP and R_e was tested to determine whether the truth lay within the 90% CI for estimates. The fraction of tests that were successful was compared with the mean size of the 90% confidence interval for each specific algorithm (Figure 5). There was evidence of a positive relationship between success rate and confidence interval size, but some algorithms managed to contain the truth within relatively narrow confidence intervals. In the comparison for annual NEE, four algorithms (E1, E2, M1, M3) produced analyses with >80% success rate and mean confidence intervals <110 gC m^{-2} yr^{-1}. In the comparison against component fluxes (GPP and R_e), two algorithms (E2, M2) produced more balanced analyses, with relatively high success rates (>65%) and narrow confidence intervals (<300 gC m^{-2} yr^{-1}). M3 was always 100% successful in containing the truth within its 90% confidence intervals, and this over-confidence was because associated CI were the largest of all algorithms for GPP and R_e. There were successful tests for prognoses in year 3 by several algorithms, indicating that predictions of C fluxes beyond the observational period were successful also (

Table 8).

GPP and R_e estimates

The decomposition of observed NEE data into GPP and R_e revealed major differences among algorithms, with best estimates varying by up to 900 gC m^{-2} yr^{-1} (Figure 6, Figure 7). However there were similar patterns among algorithms across years. For instance, M4 tended to estimate lower magnitudes of these fluxes than other algorithms. In most cases the algorithms ranked the GPP and R_e similarly across years at each site, but not always. For instance, M1 and M5 ranked R_e differently for DE-EC across years (Figure 6). Flux analyses were compared with estimates from other gap-filling and GPP-R_e decomposition algorithms using data from the same sites (Desai et al. 2008). In some cases there was close agreement between estimates, for instance NEE at Loobos in 2000 (Figure 7), but in other, such as Loobos in 2001, there was disagreement.

Stocks

The analyses and predictions of foliar C matched the seasonal cycles and magnitudes of the truth from the synthetic studies adequately (Figure 8). Predictions of year 3 foliar C in the eddy covariance datasets had a mean RMSE among algorithms of 11 gC m^{-2} for DE and 22 gC m^{-2} for EV. However, assessments of confidence intervals were generally poor; most algorithms had 90% CI either too broad or too narrow (

Table 9).

For the synthetic data, the algorithms reproduced the seasonal cycles in fine root biomass, but the magnitude of the cycles and the mean biomass varied among the algorithms by ~+50% (data not shown). This result reflected the choice of initial conditions, or their method of assessment, by the users. We found similar patterns in litter and labile C pools. Seasonal data on the variation in these C pools would be a useful addition to model-data fusion studies.

There were some important differences in the analyses and predictions of the slow turnover C pools in all datasets. C_{som} in most analyses showed slight increases or decreases over time, but

some algorithms showed stocks doubling over three years (Figure 9). Such doublings were unrealistic outcomes, but in these cases the algorithms were able to make these changes consistent with the flux observations. For C_w (Figure 10) most algorithms suggest a small increase in C stocks over time, but the algorithms with increasing C_{som} matched this with decreases in C_w of similar magnitude.

Discussion

There have been previous attempts to parameterise C models using time series of C fluxes (Braswell et al. 2005, Knorr and Kattge 2005, Wang et al. 2007). These studies have tended to focus on calibrating physiological parameters, related to photosynthetic and respiration rates, rather than parameters related to allocation and turnover of C pools. The calibration of parameters interacting on a range of timescales and links to data over several years is thus an important and novel component of REFLEX. The feedbacks between fluxes and stocks (e.g. photosynthesis and foliar C), and between soil organic matter and temperature, are particularly important determinants of NEE in the DALEC model that are investigated in REFLEX.

Parameter estimation

We expected that parameters linked to fast-response processes that mostly determine net ecosystem exchange of CO_2 (NEE) would be well constrained and well characterised, while parameters for slow processes would be poorly characterised. Our analyses largely supported this expectation. The turnover of litter and foliage were well estimated according to our criteria, and these parameters are closely associated with foliage mass and/or gaseous exchanges of C. We had not expected the turnover rate of SOM, a large slow turnover pool, to be so well constrained, but it is an important determinant of heterotrophic respiration nevertheless. Parameters associated with the turnover of wood and allocation to roots were

poorly estimated, and sometimes biased. These parameters were not directly associated with gas exchange or leaf area, and so were only weakly constrained by NEE and LAI data.

Flux and stock estimation

There was weak agreement among algorithms in estimations of 90% CI on NEE and its component fluxes, for all datasets. The differences in CI size were closely related to differences among algorithms in parameter confidence intervals. There were considerable differences in assessments among similar algorithms (e.g. Metropolis), suggesting that the subjective choices of convergence tests versus statistical tests, priors for the parameters, and likelihood function within the method were important determinants of CI. None of the algorithms consistently included ~90% of the synthetic true daily NEE values, or observed EC year 3 daily NEE data, within the 90% confidence interval of the best-fit NEE (Table 8and Table 9). All algorithms at some point over- or underestimated the confidence interval. For annual assessments of NEE, GPP and R_e, there was more success, with some algorithms successful locating the true value from synthetic experiments within relatively narrow 90% confidence intervals (Figure 5).

Assimilation results for annual flux predictions were in overall agreement with previous estimates from gap-filling studies on half-hourly data (Desai et al. 2008). However, in a number of cases the mean 90% CI did not include the gap-filled value (Figure 6 and Figure 7), for instance NEE in 2001 for Loobos. Some differences were to be expected, because the REFLEX database used only a subset of the measured data (when > 90% of half-hourly periods were measured in a day), and the assimilation was based on daily sums rather than half-hourly measurements. The general agreement in the partitioning of NEE into GPP and R_e using daily NEE data by REFLEX and half-hourly data by Desai et al. (2008) is notable. Respiration data can be easily extracted from hourly exchange data, but partitioning using daily data requires an effective GPP model, and sound predictions of foliar C. The

partitioning result suggests that the DALEC GPP and phenology sub-models have worked reasonably at the FLUXNET sites. These results indicate that daily data are effective for model calibration, and that hourly resolution is not necessarily an advantage in generating predictions of annual C exchanges.

Model error

We expected that with EC datasets there would be an increase in parameter uncertainty, and perhaps only 3-4 constrained parameters, because the model would misrepresent key processes affecting the observations. However, we found mean similar values of d_1 and d_2 for both EC and SYN datasets (Table 6 and Table 7). There did not seem to be any improved parameter constraint resulting in the synthetic case, where the model error was zero, as it was known to be valid. However, a comparison of confidence interval size on annual NEE estimates generated from synthetic and observed data did reveal a common pattern, with larger CI for EC datasets. Based on the comparison between CI on SYN and EC datasets, we conclude that the impact of model error was to increase the size of confidence intervals on annual NEE estimates by ~31%.

Prediction error

Prediction error, determined by forcing the model for 12 months beyond the assimilation period, was more complex to determine, because confidence intervals varied strongly between years 1 and 2 of the analysis. The only factor in common to all datasets was the lack of priors for initial conditions of C_f, C_{lit} and C_r. Thus, it is likely that erroneous initial conditions and/or large uncertainties on the initial values caused larger CI in year 1. The initial pools were often out of equilibrium with parameters, and so changed relatively quickly at first. By year two, parameter and state equilibria for these fast C pools reduced uncertainty. For predictions in year 3, lacking constraint of observations, uncertainty increased. CI on predictions (year 3)

were > twice those for year 2 analyses. For the SYN experiments, the year 3 predictions among algorithms were similarly successful to years 1 and 2 – that is, a similar fraction of 90% confidence intervals on annual flux estimates encompassed the truth. This result suggests that the quantification of increasing CI was reasonable.

Algorithm assessment

We examined the different algorithms, to determine if there were distinct winners or losers. All approaches produced broadly similar parameter retrievals (Figure 2) for both synthetic and observed datasets (Table 6, Table 7). All approaches generated effective best estimates and predictions of daily NEE, as shown by the small RMSEs. But the focus of this study was also on the generation of sound confidence intervals to supplement these estimates. At the daily time-step the results were equivocal, with a tendency for algorithms to be over- or under-confident (Table 8). But at the annual timescale, perhaps the most relevant for C studies, we found that most of the algorithms encompassed the truth within 90% CI. A complementary test was to check the mean size of confidence intervals, to identify and weed out those cases where a successful test was obtained by using very broad CI. Thus, the test of annual NEE, GPP and R_e retrievals (year 1 and 2) and predictions (year 3) against the known truth from the synthetic experiments (Figure 5) is perhaps the most useful judgement on the individual algorithms. According to this test, metropolis methods, Kalman filters and genetic algorithms were all capable of correctly identifying a large proportion of true fluxes with relatively small confidence intervals. Thus all approaches were valid, but some implementations were more effective in terms of this test on confidence intervals than others (see appendix for more information on algorithms).

For the Metropolis methods, confidence intervals on fluxes were generated as a function of the set of acceptable parameter sets. These parameters sets were fed into the model to produce a set of possible outcomes, that were then sampled to determine the 90% CI. Differences in

the size of the CI depend on the accept/reject criterion employed by each algorithm in generating acceptable parameter sets (Table 5). The methods employing the Kalman filter employed a further step, once acceptable parameter sets were determined. The state variables of the model, including flux estimates, were updated using sequential assimilation of observations through the times series. This sequential updating, allowing shifts in states through the model run unconnected to parameters, may be connected to the success of the Kalman filter methods (E1, E2, U1) in generating effective, but narrow confidence intervals.

Some algorithms had problems with large changes to C_w and C_{som} pools, which could be made consistent with the flux data, but are not ecologically sound in an undisturbed ecosystem. This seems to be partly related to a steady state assumption being made where pool sizes are first confined to equilibrium which likely leads to a wrong initial system state, potential biases in parameters and inflation of their confidence intervals as shown recently in a specific study by Calvalhais et al. (2008). These symptoms are, for example, also seen in the approach M4, where a spin-up was performed. Hence, a way to estimate the initial state of the system without an *ad-hoc* steady state assumption is crucial to successful MDF and should be explored further. A constraint on the annual changes in these pools based on repeated inventories would help solve this problem. Stem inventories are likely to be easier to undertake with quantifiable error than those on SOM, and so should be the focus for future studies. Nevertheless, if longer time scale are to be addressed there is a need to imposed constraints from soil carbon data, e.g. via chronosequences or profile data. Some algorithms did not explicitly include searching for initial conditions on C_f, C_{lab} and C_{lit}, and this caused some problems for e.g. E2. All algorithms need to assess their estimates of uncertainties and develop new approaches for uncertainty estimates that are consistent with the observations.

This experiment has demonstrated the value of using synthetic datasets in understanding data assimilation problems. It is clear that even with a perfect model, existing model-data fusion

571 approaches find it difficult to analyse parameters using synthetic, noisy and sparse datasets.
572 The information content of data that can be extracted by MDF depends on data quality and
573 coverage. Further synthetic studies will illuminate the relationship between data availability
574 and parameter constraint. It is clear that there is little consensus on how to generate
575 confidence intervals, with very broad ranges among algorithms. Tests using confidence
576 intervals provide a useful first look at assessing the uncertainties quantified by the various
577 algorithms, although representing continuous probability distributions with a confidence
578 interval suffers from using an arbitrary cutoff criteria. Algorithms that are not well
579 constrained by the data, and thus have wide CI's, will be more likely to contain the true value
580 but this suggests they are less able to make use of all the information in the data.
581

582 ## Conclusions

583 A range of model-data fusion algorithms exist that can generate useful estimates of parameter
584 probability density functions and state estimates for C models using a C model and daily net
585 ecosystem exchange data, derived either from observations or synthetically. While there was
586 less agreement among algorithms on the size of confidence intervals on parameter and state
587 estimates, some algorithms were able to make effective estimates of annual fluxes within
588 relatively small CI, when compared to detailed gap-filled estimates or the synthetic „truth".
589 Overall, algorithms generated narrower confidence intervals in analyses using synthetic data
590 compared to observed data. Likewise, confidence intervals were larger by 88% for forecast
591 periods than during data-fusion periods. These results suggest that some algorithms were
592 generally able to make a reasonable quantification of error propagation in prediction periods,
593 and of the likely size of model error, but that differences in estimated confidence intervals
594 suggests further improvements are required. Further studies should explore the importance of
595 assumptions about parameter priors (Gaussian or uniform), and the handling of unknown

initial conditions. Exploring the growth in CI over forecast periods of multiple years also needs to be explored in a further study. Data on slow, large C pools should be included in assimilation experiments, even with large confidence intervals. Such data can help constrain the parameters poorly served by eddy covariance data, which are those related to allocation and turnover of wood and roots.

Acknowledgements

NERC funded much of this work through the CarbonFusion International Opportunities grant. MW, CT, AF and AR were involved in the initial planning and discussions for REFLEX at the CarbonFusion meeting in Edinburgh, May2006. AF undertook the main work on setting up the experiment, collating results, and undertaking analyses. MW initiated and managed the experiment, and wrote the manuscript. AR was involved in defining the analytical process and the parameter estimation assessment. Other authors participated in the experiment and contributed to the paper. We are grateful to A Granier and the Hesse research team, and to EJ Moors and the Loobos research team, for access to their data, and the FLUXNET team for data processing and preparation. We are grateful to Mike Raupach and Damian Barrett for their ideas and input at the start of the project, and in developing the protocol for the experiment outlined here. We also recognise Zhang Li‟s efforts and input to the analysis. Jens Kattge made useful comments on the manuscript. AR acknowledges support from the Office of Science (BER), U.S. Department of Energy, through the Northeastern Regional Center of the National Institute for Climatic Change Research.

References

Baldocchi, D., E. Falge, L. H. Gu, R. Olson, D. Hollinger, S. Running, P. Anthoni, C. Bernhofer, K. Davis, R. Evans, J. Fuentes, A. Goldstein, G. Katul, B. Law, X. H. Lee, Y. Malhi, T. Meyers, W. Munger, W. Oechel, K. T. P. U, K. Pilegaard, H. P. Schmid, R. Valentini, S. Verma, T. Vesala, K. Wilson, and S. Wofsy. 2001. 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**:2415-2434.

Bonan, G. B. 2008. Forests and Climate Change: Forcings, Feedbacks, and the Climate Benefits of Forests. Science **320**:1444-1449.

Braswell, B. H., W. J. Sacks, E. Linder, and D. S. Schimel. 2005. Estimating diurnal to annual ecosystem parameters by synthesis of a carbon flux model with eddy covariance net ecosystem exchange observations. Global Change Biology **11**:335 - 355.

Carvalhais, N., M. Reichstein, S. J., C. G.J., J. Santos Pereira, P. Berbigier, A. Carrara, A. Granier, L. Montagnani, D. Papale, S. Rambal, M. J. Sanz, and R. Valentini. 2008. Implications of Carbon Cycle Steady State Assumptions for Biogeochemical Modeling Performance and Inverse Parameter Retrieval. Global Biogeochemical Cycles **22, GB2007**.

Desai, A. R., A. D. Richardson, A. M. Moffat, J. Kattge, D. Y. Hollinger, A. Barr, E. Falge, A. Noormets, D. Papale, M. Reichstein, and V. J. Stauch. 2008. Cross site evaluation of eddy covariance GPP and RE decomposition techniques. Agricultural and Forest Meteorology **148**:821-838.

Friedlingstein, P., P. Cox, R. Betts, L. Bopp, W. Von Bloh, V. Brovkin, P. Cadule, S. Doney, M. Eby, I. Fung, G. Bala, J. John, C. Jones, F. Joos, T. Kato, M. Kawamiya, W. Knorr, K. Lindsay, H. D. Matthews, T. Raddatz, P. Rayner, C. Reick, E. Roeckner, K. G. Schnitzler, R. Schnur, K. Strassmann, A. J. Weaver, C. Yoshikawa, and N. Zeng. 2006. Climate-carbon cycle feedback analysis: Results from the (CMIP)-M-4 model intercomparison. Journal of Climate **19**:3337-3353.

Gelman, A. 1995. Bayesian data analysis. Chapman & Hall, London.

Gove, J. H., and D. Y. Hollinger. 2006. Application of a dual unscented Kalman filter for simultaneous state and parameter estimation in problems of surface-atmosphere exchange. Journal of Geophysical Research **111**:doi:10.1029/2005JD00621.

Hagen, S. C., B. H. Braswell, E. Linder, S. Frolking, A. D. Richardson, and D. Hollinger. 2006. Statistical uncertainty of eddy flux-based estimates of gross ecosystem carbon exchange at Howland Forest, Maine. Journal of Geophysical Research **111**:D08S03.

Heimann, M., and M. Reichstein. 2008. Terrestrial ecosystem carbon dynamics and climate feedbacks. Nature **451**:289-292.

Janssen, P. H. M., and P. S. C. Heuberger. 1995. Calibration of process-oriented models. Ecological Modelling **83**:55 - 66.

Julier, S. J., and J. K. Uhlmann. 2004. Unscented filtering and nonlinear estimation. Proc. IEEE **92**:410-422.

Kalman, R. E. 1960. A New Approach to Linear Filtering and Prediction Problems. Transactions of the ASME--Journal of Basic Engineering **82**:35-45.

Knorr, W., and J. Kattge. 2005. Inversion of terrestrial ecosystem model parameter values against eddy covariance measurements by Monte Carlo sampling. Global Change Biology **11**:1333-1351.

Lasslop, G., M. Reichstein, J. Kattge, and D. Papale. 2008. Influences of observation errors in eddy flux data on inverse model parameter estimation. Biogeosciences **5**:1311-1324.

Liu, Y., and H. V. Gupta. 2007. Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework. Water Resources Research **43**:W07401.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equation of state calculations by fast computing machines. J. Chem. Phys. **21**:1087–1092.

Moffat, A. M., D. Papale, M. Reichstein, D. Y. Hollinger, A. D. Richardson, A. G. Barr, C. Beckstein, B. H. Braswell, G. Churkina, A. R. Desai, E. Falge, J. H. Gove, M. Heimann, D. Hui, A. J. Jarvis, J. Kattge, A. Noormets, and V. J. Stauch. 2007. Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes. , 147: 209-232. Agricultural and Forest Meteorology **147**:209-232.

Moore, D., J. Hu, W. Sacks, D. Schimel, and R. Monson. in press. Estimating transpiration and the sensitivity of carbon uptake to water availability in a subalpine forest using a simple ecosystem process model informed by measured net CO_2 and H_2O fluxes. Agricultural and Forest Meteorology.

Mosegaard, K., and A. Tarantola. 1995. Monte-Carlo Sampling of Solutions to Inverse Problems. Journal of Geophysical Research-Solid Earth **100**:12431-12447.

Quaife, T., P. Lewis, M. De Kauwe, M. Williams, B. E. Law, M. D. Disney, and P. Bowyer. 2008. Assimilating Canopy Reflectance data into an Ecosystem Model with an Ensemble Kalman Filter. Remote Sensing of the Environment **111**:1347-1364.

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

Ricciuto, D. M., K. J. Davis, and K. Keller. 2008. A Bayesian calibration of a simple carbon cycle model: The role of observations in estimating and reducing uncertainty. Global Biogeochemical Cycles **22**:GB2030.

Richardson, A. D., D. Y. Hollinger, G. G. Burba, K. J. Davis, L. B. Flanagan, G. G. Katul, J. W. Munger, D. M. Ricciuto, P. C. Stoy, A. E. Suyker, S. B. Verma, and S. C. Wofsy. 2006. A multi-site analysis of random error in tower-based measurements of carbon and energy fluxes. Agricultural and Forest Meteorology **136**:1-18.

Richardson, A. D., M. D. Mahecha, E. Falge, J. Kattge, A. M. Moffat, D. Papale, M. Reichstein, V. J. Stauch, B. H. Braswell, G. Churkina, B. Kruijt, and D. Y. Hollinger. 2008. Statistical properties of random CO_2 flux measurement uncertainty inferred from model residuals. Agricultural and Forest Meteorology **148**:38-50.

Sacks, W. J., D. S. Schimel, and R. K. Monson. 2007. Coupling between carbon cycling and climate in a high-elevation, subalpine forest: a model-data fusion analysis. Oecologia **151**:54-68.

Stockli, R., D. M. Lawrence, G. Y. Niu, K. W. Oleson, P. E. Thornton, Z. L. Yang, G. B. Bonan, A. S. Denning, and S. W. Running. 2008. Use of FLUXNET in the community land model development. Journal of Geophysical Research-Biogeosciences **113**.

Trudinger, C. M., M. R. Raupach, P. J. Rayner, and I. G. Enting. 2008. Using the Kalman filter for parameter estimation in biogeochemical models. Environmetrics **DOI: 10.1002/env.910**.

Trudinger, C. M., M. R. Raupach, P. J. Rayner, J. Kattge, Q. Liu, B. Pak, M. Reichstein, L. Renzullo, A. D. Richardson, S. H. Roxburgh, J. Styles, Y. P. Wang, P. Briggs, D. Barrett, and S. Nikolova. 2007. OptIC project: An intercomparison of optimization techniques for parameter estimation in terrestrial biogeochemical models. Journal of Geophysical Research-Biogeosciences **112**.

Van Oijen, M., J. Rougier, and R. Smith. 2005. Bayesian calibration of process-based forest models: bridging the gap between models and data. Tree Physiology **25**:915-927.

Van Wijk, M. T., and W. Bouten. 2001. Towards understanding tree root profiles: simulating hydrologically optimal strategies for root distribution. Hydrology and Earth System Sciences **5(4)**:629 - 644.

Van Wijk, M. T., and W. Bouten. 2002. Simulating daily and half-hourly fluxes of forest carbon dioxide and water vapor exchange with a simple model of light and water use. Ecosystems **5**:597-610.

Van Wijk, M. T., B. Van Putten, D. Y. Hollinger, and A. D. Richardson. 2008. Comparison of different objective functions for parameterization of simple respiration models. Journal of Geophysical Research **113**:G03008.

Vrugt, J. A., C. G. H. Diks, H. V. Gupta, W. Bouten, and J. M. Verstraten. 2005. Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation. Water Resources Research **41**:W01017.

Vrugt, J. A., H. V. Gupta, W. Bouten, and S. Sorooshian. 2003. A shuffled complex evolution metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. Water Resources Research **39**:Art. no. 1201.

Wang, Y. P., D. Baldocchi, R. Leuning, E. Falge, and T. Vesala. 2007. Estimating parameters in a land-surface model by applying nonlinear inversion to eddy covariance flux measurements from eight FLUXNET sites. Global Change Biology **13**:652 - 670.

Waring, R. H., J. J. Landsberg, and M. Williams. 1998. Net primary production of forests: a constant fraction of gross primary production? Tree Physiology **18**:129-134.

Williams, M., E. B. Rastetter, D. N. Fernandes, M. L. Goulden, G. R. Shaver, and L. C. Johnson. 1997. Predicting gross primary productivity in terrestrial ecosystems. Ecological Applications **7**:882-894.

Williams, M., E. B. Rastetter, D. N. Fernandes, M. L. Goulden, S. C. Wofsy, G. R. Shaver, J. M. Melillo, J. W. Munger, S.-M. Fan, and K. J. Nadelhoffer. 1996. Modelling the soil-plant-atmosphere continuum in a *Quercus-Acer* stand at Harvard Forest: the regulation of stomatal conductance by light, nitrogen and soil/plant hydraulic properties. Plant, Cell and Environment **19**:911-927.

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

Williams, M., L. Street, M. T. V. Wijk, and G. R. Shaver. 2006. Identifying differences in carbon exchange among arctic ecosystem types. Ecosystems **9**:288-304.

Appendix: Details of algorithms

E1: Stage 1. Parameter estimation. Parameters were initially estimated using a simple Metropolis MCMC-type approach. (e.g. (Mosegaard and Tarantola 1995, Knorr and Kattge 2005)). Initial prior distributions were assumed to be uniform and encompass the entire possible suggested range and so a single stage accept/reject criterion was used based on comparison of model output with data alone. Initial values for Cr, Clit and Clab were estimated in the same manner as parameters, initial values for C_f were based on first available observation (EV case) or set to zero (DE case). The model was initialized from a random location, and step size was constant and determined as 0.001 of log-transformed parameter range. This was determined through „tuning" initial runs to ensure an acceptance probability of between 0.2 and 0.8 at each step. The number of steps required to sufficiently sample the parameter space was assessed using the Gelman criteria (Gelman 1995)) to test convergence between chains.
E1: Stage 2. State estimation. 8000 parameter sets were randomly sampled from the accepted parameters from Stage 1. These were then used in an 8000 member Ensemble Kalman Filter (EnKF, Evensen, 1994; Williams et al., 2005). A unique parameter set was assigned to each ensemble member with the intention this would cause divergence between ensemble members representing model error and cause a growth in state uncertainty equivalent to that inherent from parameter uncertainty alone. This was done instead of adding a stochastic forcing term at each time step. This is possibly correct in the SYN cases when model „structural" error is known to be zero, but will probably underestimate model error in the EC cases and overly restrict growth in state uncertainty. Nonetheless, assimilation of observations did alter the state variables in the resulting analysis and reduce uncertainties in state estimates even though these same observation data had already been used to generate the parameter sets in Stage 1 so offered little additional information to the EnKF.

E2: Ensemble Kalman filter. This method was set up for joint estimation of states and parameters, so parameters were included within the state vector for assimilation. Model parameter errors were set within bounds - small enough to avoid tracking daily noise in observations, and large enough to shift over weekly-seasonal timescales in response to process signals. Errors on model states were set smaller than for parameters, so that assimilation was focused on updating parameters rather than states. Initial values for all parameters and initial conditions for $C_{f,}$ C_{lab} and C_r were estimated. After an initial assimilation of observations, these initial parameter estimates were updated with the final estimates from the assimilation. We assumed that $C_{f,}$ C_{lab} and C_r would be in approximate steady state over annual cycles, and adjusted initial values accordingly. A further EnKF assimilation was then applied, using these updated initial parameters and initial conditions, to generate final analyses.

U1: Unscented Kalman filter. The UKF was used to provide state estimates for each of the experiments. The UKF (Julier and Uhlmann 2004) is a nonlinear version of the traditional linear Kalman filter (Kalman 1960), that uses a deterministic sampling of so-called sigma points in order to capture the mean and covariance of the state. Similar to other Kalman type filters it employs a two-step `predictor-corrector' scheme where model predictions are corrected by measurements as they arrive sequentially in time. At time periods where measurements are missing, only the prediction step is used. To employ the UKF, the general nonlinear state space model was assumed, with the variants of the model taking the form of the state evolution equations. A linear measurement model was used in all runs. Both the state

and measurement equations assume zero mean random noise processes with associated full-dimensional covariance matrices (Gove and Hollinger 2006). The later were estimated from the information provided. The parameter estimates used in the filter runs were arrived at via simulated annealing method M3. Parameters for the unscented transformation were set to α=1, β=2 and κ=1 for all experiments (see Gove and Hollinger 2006 for an explanation).

G1: Genetic algorithm: The implementation was from Haupt and Haupt (2004). The population size was 100, and was run for 1000 iterations (generations). Initial stores Cr, Clit, Cf and Clab were estimated by the GA as additional parameters. To estimate uncertainties, the roughly 1600 (unique) parameter sets with cost function values closest to the final best cost function value were saved, and used to estimate the covariance matrix and 90% CI.

M1: This method sought to make as few approximations as possible to Bayes Theorem, choosing the simplest algorithm to generate a representative sample from the posterior. We chose the beta distribution for our prior. The Metropolis algorithm (Metropolis et al. 1953) generates a chain that sequentially "walks through parameter space" in such a way that the chain of visited points is the sought-after sample from the posterior. Each new point in the chain is found by randomly generating a multivariate normal step away from the current vector. In this case a simple diagonal variance matrix defined this multivariate normal "proposal distribution". Whether a proposed candidate vector was accepted or not depended on the *Metropolis ratio*, which is the ratio of two products: likelihood times prior for the candidate and likelihood times prior for the current point. If the Metropolis ratio was larger than 1 (i.e. the candidate point has a higher posterior probability then the current point), it was always accepted. If the Metropolis ratio was less than 1 (i.e. the candidate was "less probable" than the current vector), the candidate could still be accepted but only with probability equal to the Metropolis ratio. The chain was stopped when it "converged", i.e. it had explored the parameter space adequately. Convergence was confirmed visually using the trace plots of the different parameters, i.e. plots that show how the chain moves through parameter space for each individual parameter. If one or more of the trace plots was still showing drift towards unexplored parts or parameter space, the chain was deemed not to have converged.

M2: A combined optimization approach estimated model parameters and state variables. A genetic algorithm, Stochastic Evolutionary Ranking Strategy (SRES) was used to find the global optimum (Runarrsson and Yao, 2000). Markov chain monte carlo (MCMC) using the Metropolis-Hastings algorithm was then used to explore the parameter space around the optimum to estimate the full joint distribution of parameters and to estimate predictive uncertainty. Two chains were run for each experiment; convergence was determined by visually comparing the parameter PDFs from both chains. The ranges given for p1-17 were used as uniform distributions; no additional information was used. The initial values of pools C_r, C_{lit} and C_{lab} were also estimated as model parameters, using the prior range 20-200 gC m^{-2} as recommended. All observations are assumed to drawn from independent distributions. Both NEE and LAI errors were assumed normally distributed.

M3: Optimization of parameters and initial values of C pools took place in three stages. First, the parameter and initial state space was randomly explored for 50,000 iterations, at which point the parameter set and initial conditions with the lowest cost function was used as the starting point for the Metropolis algorithm. Second, the Metropolis algorithm was implemented to ensure progressive down-slope movement while at the same time avoiding local minima. The cost function was a weighted-sum-of-squares of both NEE and LAI deviations. 200,000 steps were taken in this manner. Third, reverting to the best parameter set

obtained, the parameter space was explored again until 1,000 parameter sets have been accepted as "almost as good as" the optimal parameter set, using a χ^2 test to determine the threshold contour (90% confidence interval) (assuming $n - 1$ degrees of freedom for LAI and $n - p - 1$ degrees of freedom for NEE. These parameter sets were used to define the uncertainty estimates on both parameters and model predictions.

M4: Markov Chain Monte Carlo Metropolis. The algorithm adopted a global search method with an uniform walk in the model parameter space. The method is based on a Bayesian approach where the comparison between model output and data is used to update our prior knowledge of the parameter distribution. The prior distributions were considered to be uniform. The Metropolis rules prevented the algorithm from being trapped in local minima, allowing for changes in the searching direction. Spin-up was used to initialize the C pools (C_r, C_{lit} and C_{lab}); we sampled the parameters and we ran the model replicating the meteorological data until the total difference between one year and the other was less than 1g of C. The other C pools were initialized as from the experiment description.

M5. The SCEM-UA algorithm (Vrugt et al., 2003) is a modified version of the original SCE-UA global optimization algorithm (Duan et al., 1992). The algorithm is Bayesian in nature and operates by merging the strengths of the Metropolis algorithm, controlled random search, competitive evolution, and complex shuffling to continuously update the proposal distribution and evolve the sampler to the posterior target distribution. The SCEM-UA algorithm uses the Metropolis-Hastings (Metropolis et al., 1953) search strategy to generate a sequence of parameter sets (θ_1, θ_2,..., θ_n) that adapts to the target posterior distribution. It starts with an initial population of points (parameter sets) randomly distributed throughout the feasible parameter space defined by the prior parameter distributions. The population is partitioned into q complexes, and in each complex k (k = 1, 2, ..., q) a parallel sequence is launched from the point that exhibits the highest posterior density. A new candidate point in each sequence k is generated using a multivariate normal distribution either centred around the current draw of the sequence k, or the mean of the points in complex k, augmented with the covariance structure induced between the points in complex k. The Metropolis-annealing criterion is used to test whether the candidate point should be added to the current sequence. Subsequently the new candidate point randomly replaces an existing member of the complex. Finally, after a certain number of iterations new complexes are formed through a process of shuffling the old complexes. The objective function used in this study is a combination of the model errors (expressed as SSE, Sum of Squared Errors) of describing the CO_2 fluxes and the Leaf Area Index, weighted by the error variance of each variable.

866 ## Tables

Observation	Units	Interval	Source
Global Radiation	MJ m^{-2} d^{-1}	Daily	Fluxnet data portal
Min Temperature	°C	Daily	Fluxnet data portal
Max temperature	°C	Daily	Fluxnet data portal
CO_2 conc.	μmol mol^{-1}	Daily	Fluxnet data portal
NEE	g C m^{-2} d^{-1}	Daily	Fluxnet data portal
LAI	m^2 m^{-2}	When available	References/site PI
Foliar N *	gN m^{-2} leaf area	Constant	References/site PI
Aboveground C mass*	kg C m^{-2}	Initial condition	References/site PI
SOM C mass*	kg C m^{-2}	Initial condition	References/site PI
Leaf mass per area*	g C m^{-2} leaf area	Constant	References/site PI

867 **Table 1. Time series data available for use in the experiments. Data with a "constant" interval are fixed**

868 **values throughout model runs. *Ancillary data contained in Table 2 for all experiments.**

869

870

Site	Latitude (°N)	Soil organic matter C (g C m^{-2})	Above-ground biomass (g C m^{-2})	Leaf mass per area (g C m^{-2} leaf area)	Foliar N (g N m^{-2} leaf area)
EV-EC (Loobos)	52	11000	9200	110	4.0
EV-SYN	50	9700	12400	110	3.8
DE-EC (Hesse)	48	7100	8800	22	1.0
DE-SYN	51	9900	8900	22	1.1

871 **Table 2. Site details, including latitude, initial conditions for large C pools, and foliage parameters.**

872

873

	Description	Code	Range (low/high)
p1	Decomposition rate (per day)	T_d	1 x 10^{-6}/0.01
p2	Fraction of GPP respired	F_g	0.2/0.7
p3	Fraction of NPP allocated to foliage	F_{nf}	0.01/0.5
p4	Fraction of NPP2 allocated to roots	F_{nrr}	0.01/0.5
p5	Turnover rate of foliage (per day)	T_f	1 x 10^{-4}/0.1
p6	Turnover rate of wood (per day)	T_w	1 x 10^{-6}/0.01
p7	Turnover rate of roots (per day)	T_r	1 x 10^{-4}/0.01
p8	Mineralisation rate of litter (per day)	T_l	1 x 10^{-5}/0.1
p9	Mineralisation rate of SOM/CWD (per day)	T_s	1 x 10^{-6}/0.01
p10	Parameter in exponential term of temperature dependent rate parameter	E_t	0.05/0.2
p11	Nitrogen use efficiency parameter (a1) in ACM	P_r	5/20
p12 *	*GDD* value causing leaf out	L_{out}	200/400
p13 *	Minimum daily temperature causing leaf fall	L_{fall}	8/15
p14 *	Fraction of leaf loss transferred to litter	F_{ll}	0.2/0.7
p15 *	Turnover rate of labile carbon (per day)	T_{lab}	1 x 10^{-4}/0.1
p16 *	Fraction of labile transfers respired	F_{lr}	0.01/0.5
p17 *	Maximum C_f value (gC m^{-2})	C_{fmax}	100/500

874 **Table 3. Model parameters requiring calibration. NPP_2 is NPP remaining after allocation to foliage.**

875 *** parameters p12-17 are used in DALEC-deciduous only.**

876

877

878

Experiment	Data	Drivers	Sites	Parameters	States
1	FLUXNET NEE and LAI data, 2000-1	Observed, 2000-1	DE-EC, EV-EC	Generated by MDF, with 90% CI	Generated by MDF with 90% CI
2	Artificial /synthetic	Artificial	DE-SYN, EV-SYN	Generated by MDF, with 90% CI	Generated by MDF with 90% CI
3	None	Observed, 2002	DE-EC, EV-EC	From Expt 1	Generated by MDF with 90% CI
4	None	Artificial	DE-SYN, EV-SYN	From Expt 2	Generated by MDF with 90% CI

879

880 **Table 4. Experimental summary for Reflex. The table shows for each experiment the input data, the**
881 **source of the meteorological drivers, and the site codes. The first two experiments generated parameter**
882 **estimates and estimates of model states (fluxes and pools of C), while the final two experiments were**
883 **forecasts of model states only. Acronyms: DE - deciduous vegetation, EV - evergreen vegetation, SYN –**
884 **synthetic data, EC- observed eddy covariance and LAI data.**

Participant	Name – type of methodology	Code	Prior	Cost/objective function	Initial pools	Convergence tests	Number of parameter sets produced	Number of model iterations	Programming language
E1 (stage 1)	MCMC Metropolis, then EnKF		Uniform	$J=\frac{1}{2}\sum_{i=1}^{N} f\frac{x_i,p-OBS_i^2}{\sigma_{obs,i}^2}$	Parameters to be estimated	Gelman and Rubin (1992)	~400000	~1000000	Fortran
E1 (stage 2)		Evensen (2003)	PDFs from stage 1	Kalman gain	PDFs from stage 1	n/a	State only	8000	Fortran
E2	Ensemble Kalman filter	Evensen (2003)	Gaussian	Kalman gain	Cr=Cfmax, Clit=0.5Cfmax, Clab=0.5Cfmax + EnKF 2 times	n/a	~2000	800	Fortran
U1	Unscented Kalman filter	Gove & Hollinger (2006)	Gaussian	Minimize posterior error covariance via the Kalman gain.	As estimated by M3	n/a	State only	n/a	R
G1	Genetic algorithm	Based on Haupt and Haupt (2004)	uniform	$J=\sum_{i=1}^{N}\frac{f[\cdot OBS_i^2}{\sigma_{obs,i}^2}+qx\left[\frac{c[365-c[0}{c[0}\right]\left[\frac{c[730-c[0}{c[0}\right]$	Tuned with parameters	n/a	~100000		Fortran
M1	MCMC – Metropolis			Gaussian likelihood	Included in calibration	visual		300000	Fortran
M2	MCMC – Metropolis	MCMC1	uniform	$J=\frac{1}{2}\sum_{i=1}^{N} f\frac{x_i,p-OBS_i^2}{\sigma_{obs,i}^2}$	Parameters to be estimated	Visual comparison of parameter PDFs from 2 chains	1000000	1000000	Fortran
M3	Simulated annealing-Metropolis	SAM	uniform	$J=2\frac{f[x,p-OBS]^2}{\sigma_{obs}^2}\cdot\frac{1}{N}$	Parameters to be estimated	n/a	1000	~250000	Fortran
M4	MCMC – Metropolis	MCMC3	uniform	$J=\frac{1}{2}\sum_{i=1}^{N} f\frac{x_i,p-OBS_i^2}{\sigma_{obs,i}^2}$	Spinup to equilibrium of total C	Heidelberger and Welch (1983)	80000	~300000	R
M5	Multiple complex MCMC – Metropolis	SCEM	uniform	$J=\frac{SSE_{OBS}}{\sigma_{OBS}^2}$	Parameters to be estimated	Gelman and Rubin (1992)	~500000	150000	Matlab

Table 5. A summary of the algorithms used in the experiment. Methods using Metropolis algorithm alone are labelled Mx. U1 and E1 used a Kalman filter after an initial Metropolis algorithm search for parameters. G1 and E2 are the only methods not using the Metropolis algorithm.

	Evergreen: EV-SYN					Deciduous: DE-SYN				
Param	d_1	d_2	d_3	D	Rank	d_1	d_2	d_3	D	Rank
T_d	0.26	0.36	0.75	0.87	11	0.26	0.42	0.72	0.87	17
F_g	0.30	0.41	0.02	0.51	6	0.11	0.42	0.09	0.45	8
F_{nf}	0.07	0.49	0.00	0.50	5	0.26	0.53	0.37	0.70	16
F_{nrr}	0.24	0.65	0.31	0.76	9	0.19	0.60	0.07	0.64	15
T_f	0.06	0.20	0.03	0.21	1	0.05	0.16	0.01	0.17	3
T_w	0.22	0.40	0.69	0.83	10	0.27	0.37	0.22	0.51	12
T_r	0.27	0.52	0.03	0.59	8	0.04	0.28	0.02	0.28	5
T_l	0.07	0.22	0.03	0.23	2	0.03	0.15	0.03	0.15	2
T_s	0.05	0.16	0.21	0.27	4	0.04	0.08	0.01	0.09	1
E_t	0.04	0.24	0.00	0.24	3	0.05	0.17	0.04	0.18	4
P_r	0.21	0.47	0.15	0.54	7	0.14	0.46	0.06	0.49	10
L_{out}						0.22	0.40	0.19	0.49	11
L_{fall}						0.14	0.25	0.10	0.30	6
F_{ll}						0.13	0.52	0.24	0.59	14
T_{lab}						0.19	0.54	0.01	0.57	13
F_{lr}						0.18	0.33	0.00	0.38	7
C_{fmax}						0.22	0.36	0.17	0.46	9
Mean	0.16	0.38	0.20	0.51		0.15	0.36	0.14	0.43	

Table 6. Parameter estimation metrics using 9 different algorithms based on synthetic data for evergreen (left) and deciduous (right) forest. Metric d_1 quantifies consistency among methods; d_2 quantifies the data constraint on the confidence intervals; and d_3 quantifies the consistency with the truth. D is the sum of the $d_{1\text{-}3}$. The rank column identifies the rank of D for each parameter, with lower values of D, and lower ranks, indicating better estimation.

Evergreen EV-EC					Deciduous DE-EC			
	d_1	d_2	D	Rank	d_1	d_2	D	Rank
T_d	0.28	0.42	0.5	9	0.29	0.36	0.47	14
F_g	0.11	0.36	0.37	6	0.08	0.3	0.31	7
F_{nf}	0.16	0.31	0.35	5	0.2	0.55	0.58	17
F_{nrr}	0.29	0.6	0.66	11	0.15	0.53	0.55	16
T_f	0.08	0.19	0.2	3	0.12	0.25	0.28	6
T_w	0.24	0.35	0.42	7	0.21	0.35	0.4	12
T_r	0.29	0.35	0.45	8	0.32	0.2	0.38	9
T_l	0.09	0.23	0.25	4	0.08	0.18	0.2	1
T_s	0.08	0.1	0.13	1	0.05	0.2	0.21	2
E_t	0.02	0.2	0.2	2	0.09	0.19	0.21	3
P_r	0.14	0.52	0.53	10	0.17	0.35	0.39	11
L_{out}					0.21	0.37	0.43	13
L_{fall}					0.2	0.32	0.38	10
F_{ll}					0.16	0.32	0.36	8
T_{lab}					0.1	0.49	0.5	15
F_{lr}					0.12	0.23	0.25	5
C_{fmax}					0.03	0.25	0.25	4
Mean	0.16	0.33	0.37		0.15	0.32	0.36	

Table 7 Parameter estimation metrics using 9 different algorithms based on observed data for evergreen (left) and deciduous (right) forest. Metric d_1 quantifies consistency among methods; d_2 quantifies the data constraint on the confidence intervals. D is the sum of the $d_{1\text{-}2}$. The rank column identifies the rank of D for each parameter, with lower values of D, and lower ranks, indicating better estimation.

	DE-Syn			EV-Syn		
	Year 1	Year 2	Year 3	Year 1	Year 2	Year 3
Algorithm						
M1	**0.95**	0.97	0.99	0.81	**0.89**	1.00
M2	0.73	0.65	0.81	**0.95**	0.61	0.51
M3	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*
M4	**0.95**	0.97	0.96	0.80	**0.86**	**0.85**
M5	0.66	0.37	0.36	0.39	0.25	0.35
E1	**0.90**	0.83	**0.95**	0.93	0.77	0.69
E2	0.85	0.99	*1.00*	0.44	0.61	0.60
U1	0.99	0.99	*1.00*	0.99	0.98	*1.00*
G1	*1.00*	*1.00*	0.98	*1.00*	0.99	*1.00*

Table 8. Fraction of days in each year where 90% confidence interval encompassed the synthetic "true" value of NEE. Fractions are shown for each of the 3 individual years for DE-SYN and EV-SYN datasets. Values between 0.85-0.95 are in bold and are consistent with the 90% CI. Values of 1.0 are indicated by italics.

	Foliar C mass (C_f)				Daily NEE			
	RSME (gC m^{-2})		CI frac		RMSE (gC m^{-2} d^{-1})		CI frac	
Algorithm	DE-EC	EV-EC	DE-EC	EV-EC	DE-EC	EV-EC	DE-EC	EV-EC
M1	12.3	29.9	1	0.5	1.42	1.50	0.14	0.14
M2	7.6	16.6	0	0.83	1.21	1.34	0.11	0.06
M3	6.9	19.4	1	**0.94**	1.35	1.42	0.56	0.16
M4	16.9	18.9	1	1	1.57	1.73	0.39	0.19
M5	10.6	17.8	0	0.17	2.25	1.37	0.2	0.16
E1	6.1	20.3	1	0.33	1.10	1.49	0.14	0.08
E2	30.2	37.8	1	1	1.70	1.45	**0.86**	0.16
U1	4.1	15.5	1	1	1.34	1.37	0.84	0.61
G1	4.2	22.6	1	0.83	1.24	1.54	0.43	0.16
n	1	18	1	18	218	171	218	171

Table 9. Assessment of year 3 best-fit predictions and 90% confidence intervals (CI) for the EC datasets. Comparisons with both foliar carbon mass (C_f) and daily net ecosystem exchange (NEE) are shown. Assessment of best fit predictions is through root mean square error (RMSE) on observations for year 3 for deciduous (DE) and evergreen (EV) forests. Assessment of confidence intervals is through quantifying the fraction of days in year 3 where the 90% confidence interval encompassed the observed NEE. Values between 0.85-0.95 are in bold and are deemed consistent with the 90% CI. Algorithms are identified by codes. *n* is number of observations in year 3, which were withheld from the experimental team.

Dataset	Year 1	Year 2	Year 3
DE-EC	181.0	96.6	186.2
EV-EC	119.3	92.6	169.4
DE-SYN	139.3	83.8	148.9
EV-SYN	95.4	58.1	117.9

Table 10. Mean size of 90% confidence interval on annual NEE for 3 years. Assessments were made with outputs from the nine algorithms, and compared for different years and datasets. The outputs for the first two years were analyses, based on model-data fusion. The output for the final year was generated from model predictions using estimated parameters and meteorological forcing, and no data. Units are gC m^{-2} yr^{-1}.

Figures

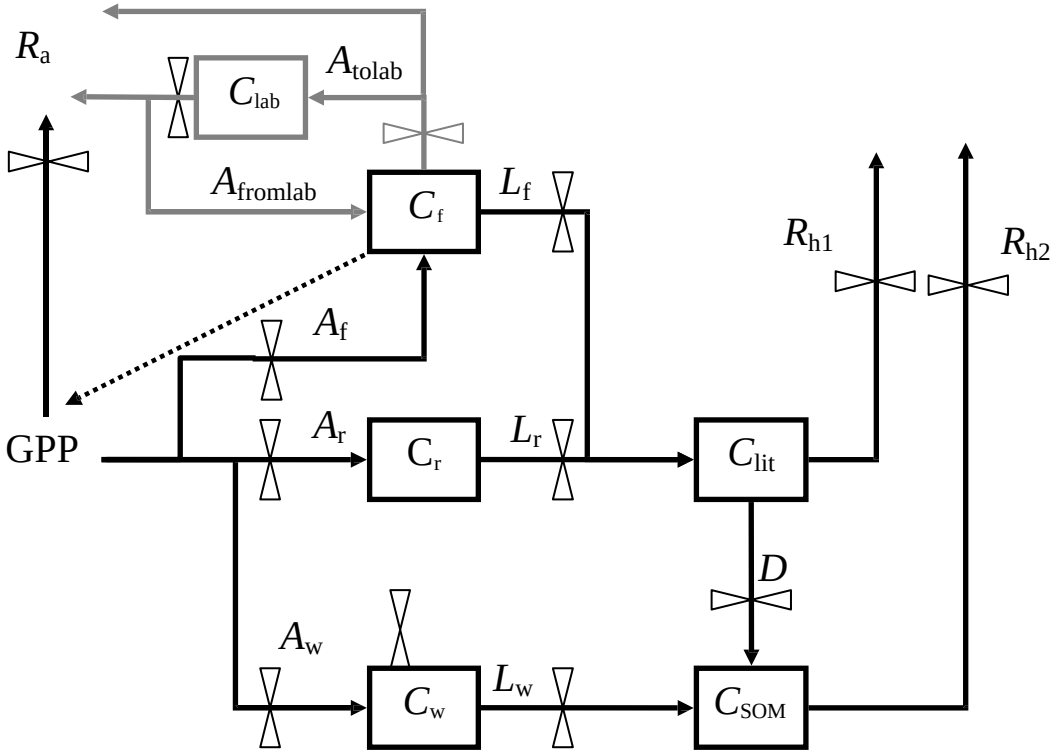

Figure 1 A schematic of the DALEC (black) and DALEC-deciduous (black and grey) models. The figures show pools (boxes) and fluxes (arrows) of C. Feed back between DALEC and the photosynthesis model is indicated by dotted line. Allocation fluxes are *A*, litter-fall fluxes by *L*, and respiration by *R*, split between autotrophic (a) and heterotrophic (h). *D* is decomposition and GPP is gross primary productivity.

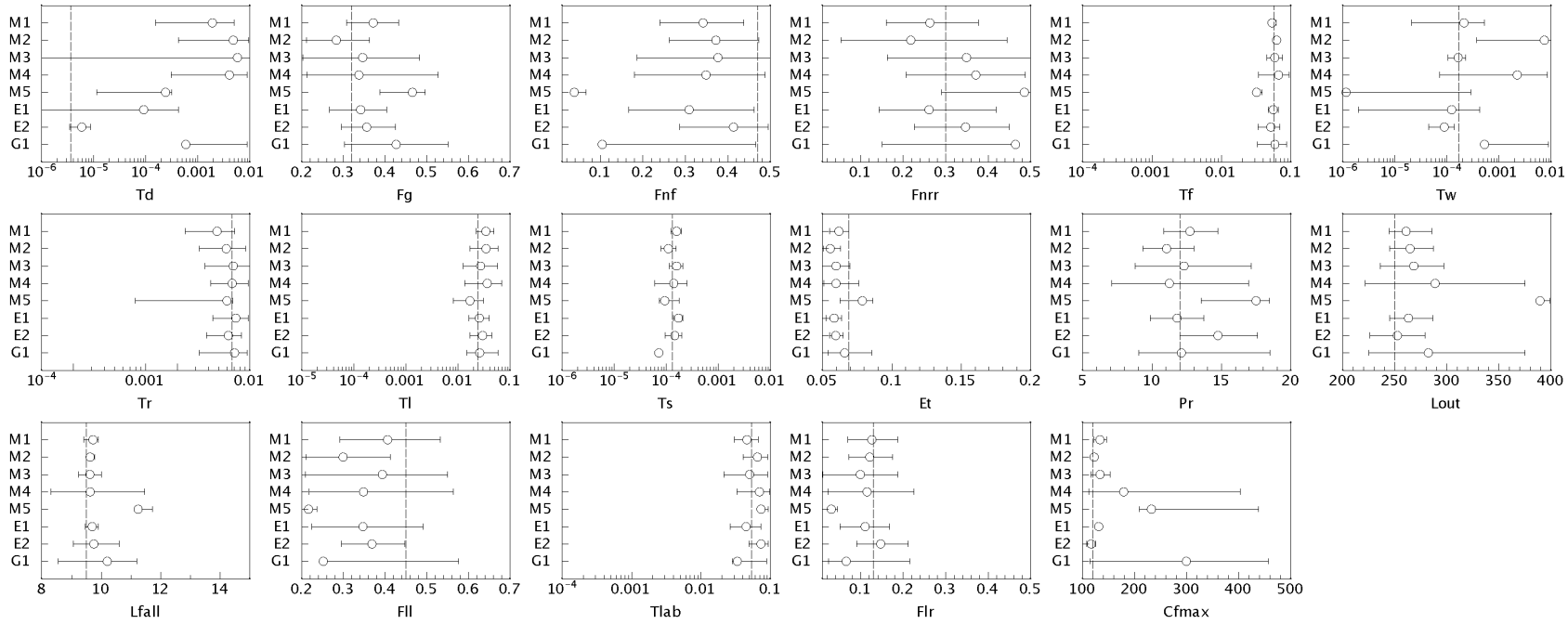

Figure 2. Parameter estimation for deciduous synthetic (DE-SYN) data. The panels shows each of the algorithms‟ best estimate of each parameter, and the magnitude of each 90% confidence intervals. The „true‟ value of the parameter used in generating the synthetic data is indicated by the d vertical line. The upper and lower bounds of each parameter, as provided to the experimenters, is indicated by the range of each x-axis. X-axes are log scaled for turnover rates (all parameters beginning *T*). For an explanation of parameter symbols see Table 5.

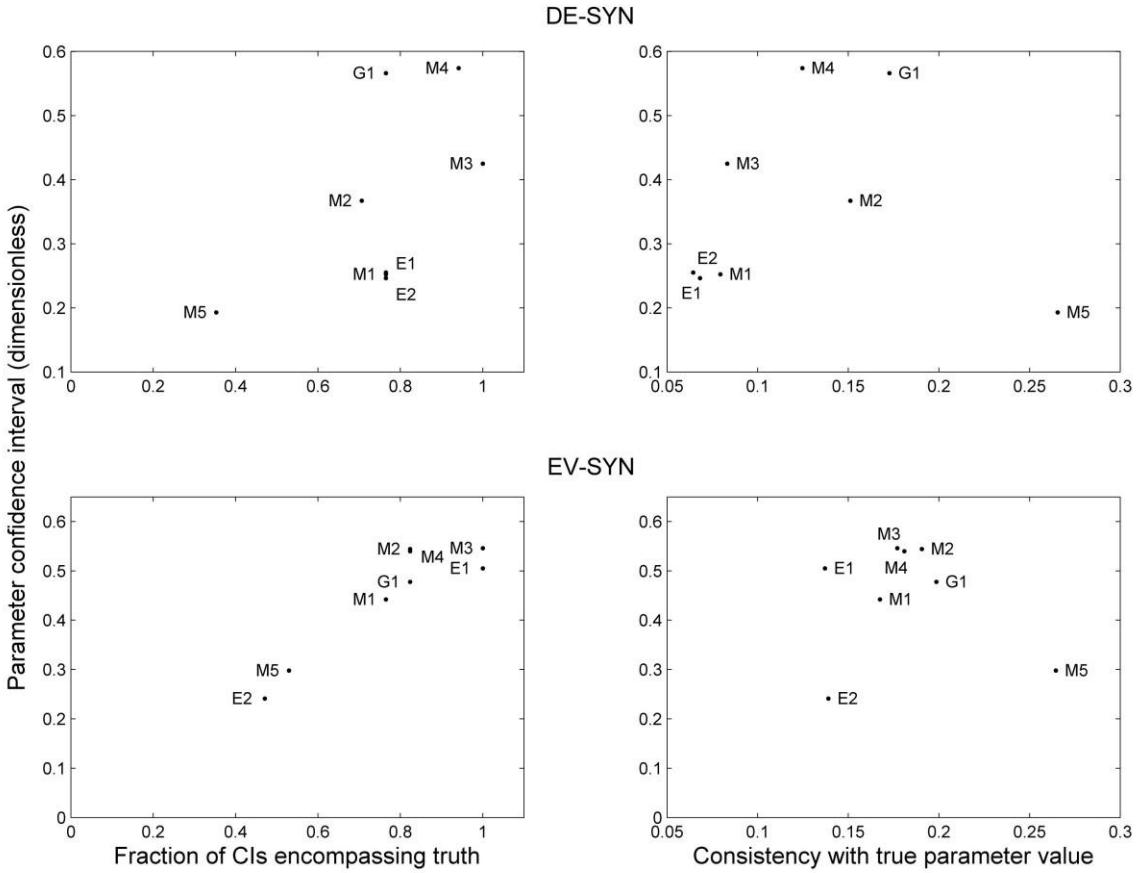

Figure 3. A comparison of two metrics of parameter calibration success against mean parameter 90% confidence intervals of each algorithm (d_4, see text). Parameter calibration success is judged in two ways: (1) by the fraction of 90% confidence intervals encompassing the true parameter values obtained by each algorithm, see left panels – high values are better; (2) by the mean normalized difference between best estimate and true parameter values obtained by each algorithm (*d5*, see text), see right panels – low values are better. Individual algorithms are identified by alphanumerics (Table 5). The top two panels are generated from the deciduous synthetic data, the bottom two from evergreen synthetic data. Data for the synthetic experiments are shown, where true values of the parameters are known.

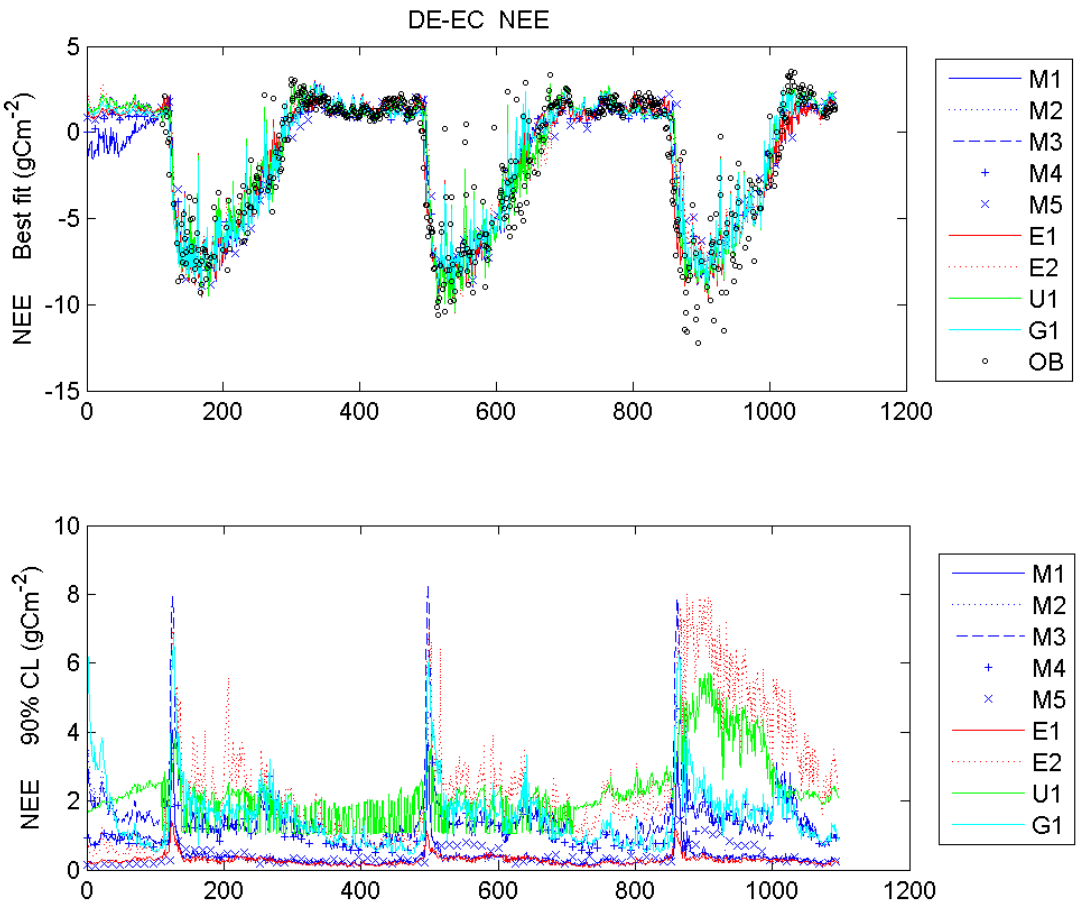

Figure 4. Estimated time series of net ecosystem exchange of CO_2 (NEE) over 3 years from each algorithm using observations from the DE-EC dataset over the first 2 years (top panel) and 90% confidence intervals on the estimates (lower panel). The eddy covariance data is shown as open symbols.

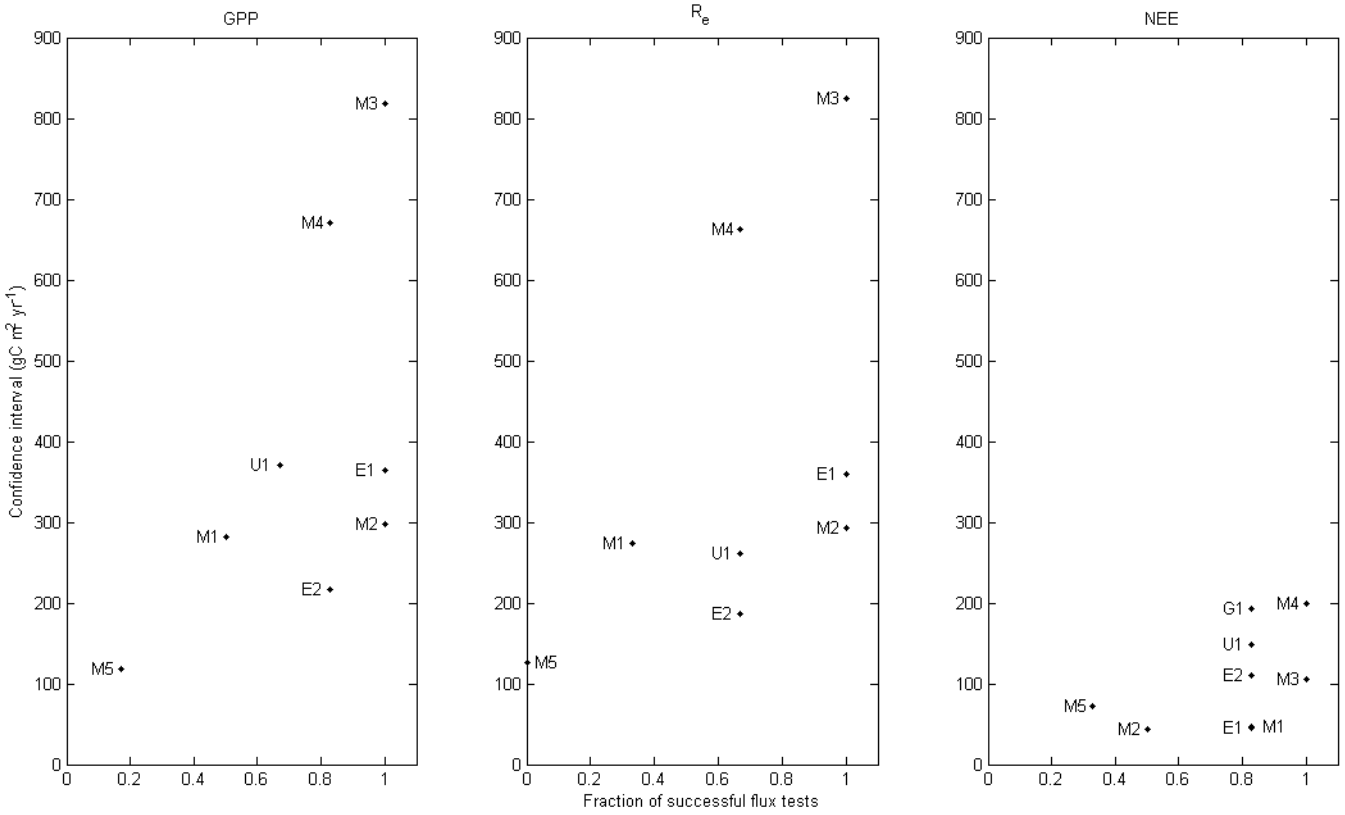

Figure 5. A comparison between the summary success rate of annual estimates of GPP (left), R_e (centre) and NEE (right) for each algorithm plotted against the mean size of the 90% confidence interval used in the tests. The tests were for DE-SYN and EV-SYN, the synthetic datasets. Success was judged on whether each "true" annual flux was within the 90% confidence interval of the estimate. There were 6 tests (3 years x 2 datasets) for each flux. On the right panel the results for E1 and M1 were very similar. All panels have the same scale.

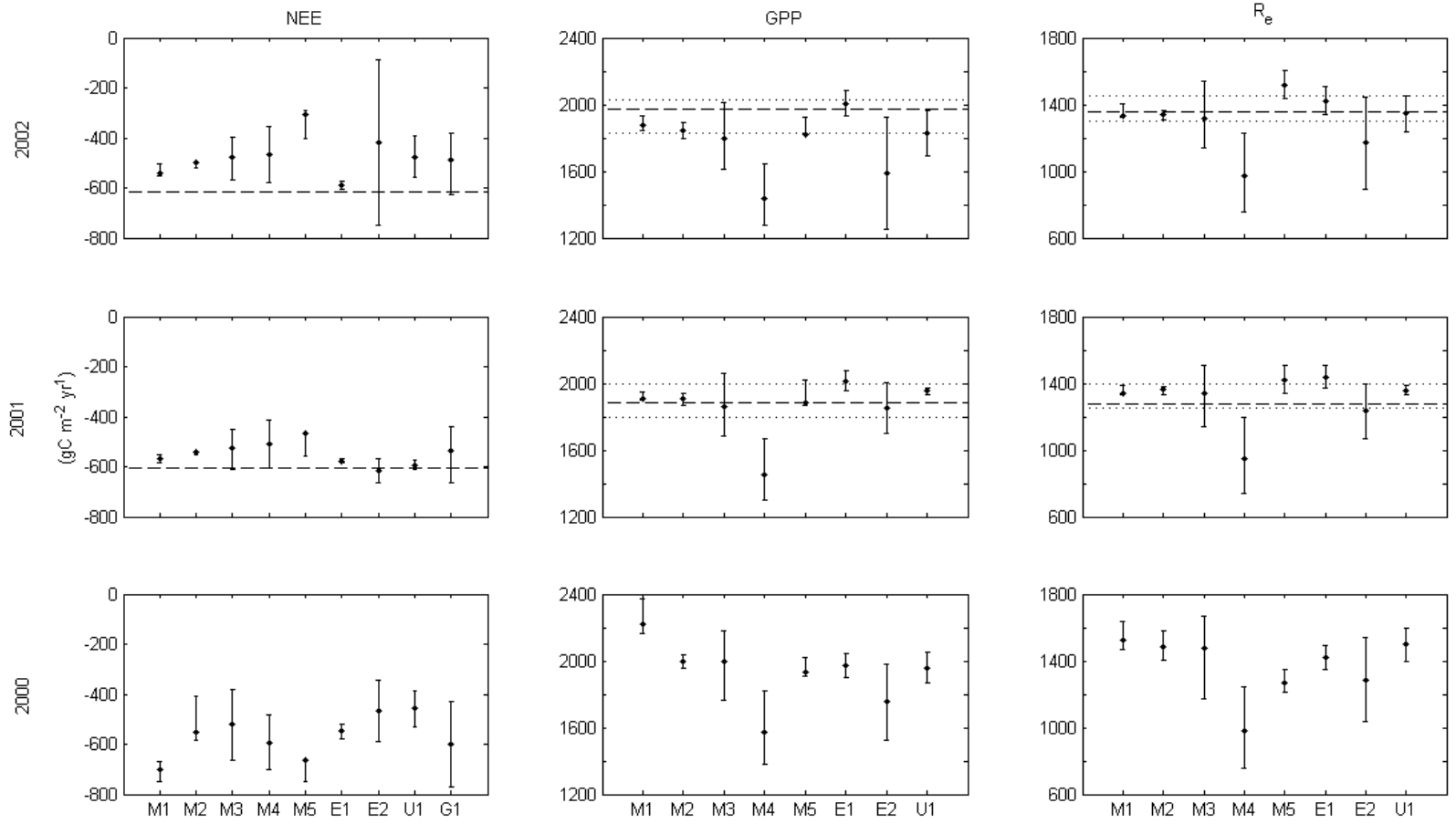

Figure 6. Annual analyses of NEE, GPP and Re for 2000, 2001 and prognoses for 2002 generated with the DE-EC dataset from Hesse, France. Results are shown for each algorithm for NEE and for 8 algorithms for GPP and R_e, with 90% confidence intervals indicated. The dashed lines show the best estimates from gap-filling routines using hourly NEE data, while the dotted lines show interquartile range among the estimates from the array of gap-filling routines for 2001 and 2002. (Desai et al. 2008).

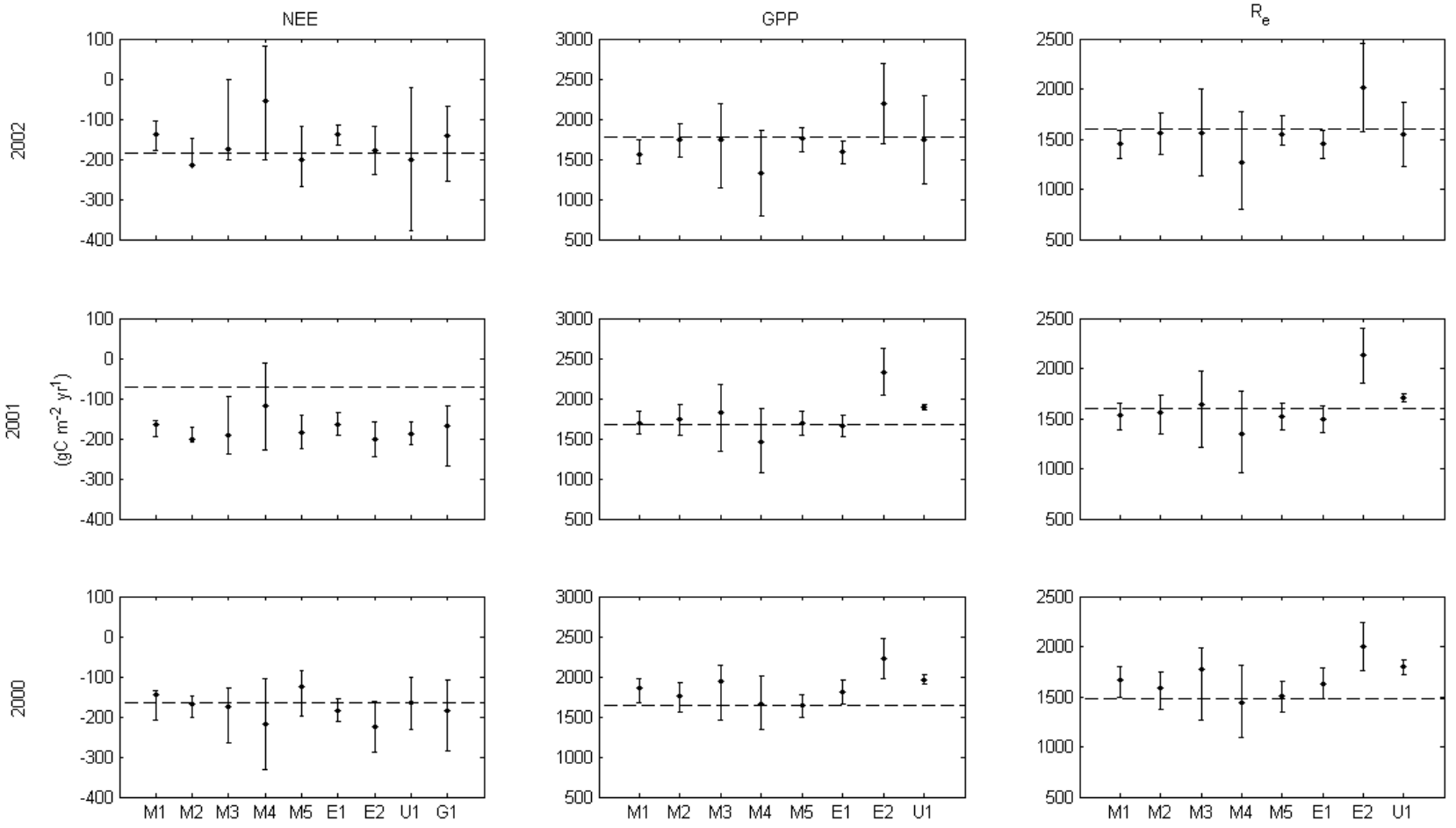

Figure 7 Annual analyses of NEE, GPP and Re for 2000, 2001 and prognoses for 2002 generated with the EV-EC dataset from Loobos, Netherlands. Results are shown for each algorithm for NEE and for 8 algorithms for GPP and R_e, with 90% confidence intervals indicated. The dashed lines show the best estimates from gap-filling routines using hourly NEE data (Desai et al. 2008).

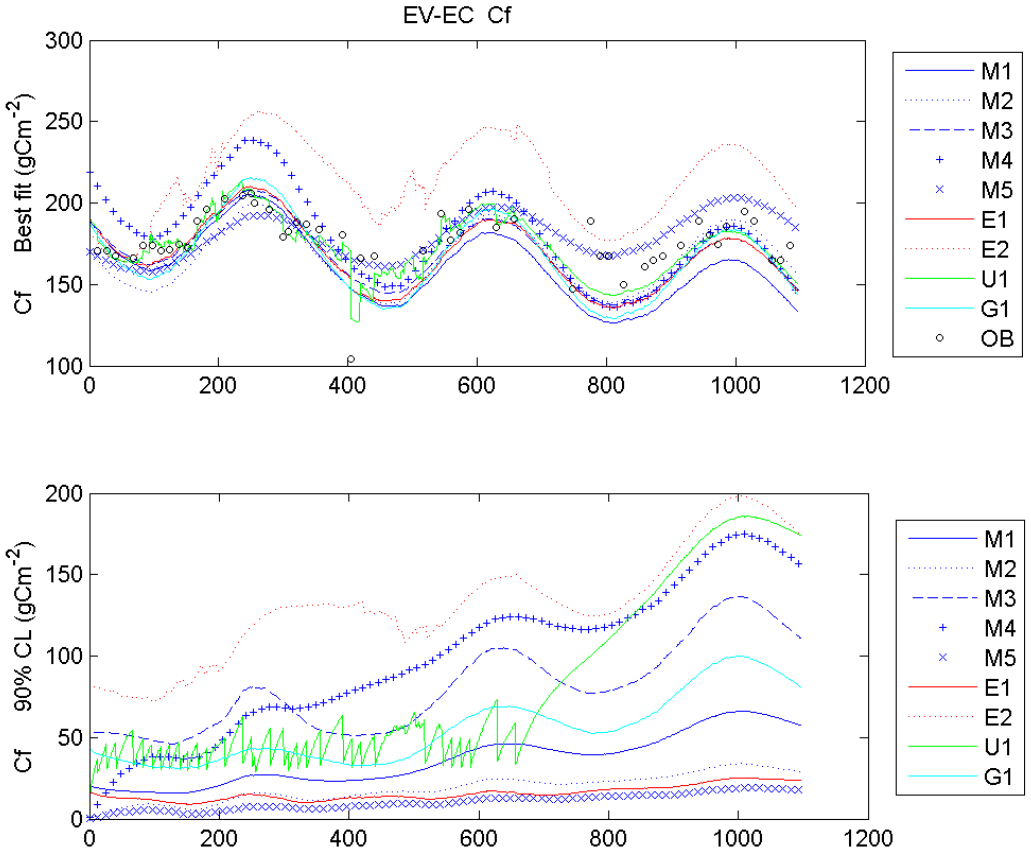

Figure 8 Retrieved estimates of foliar C stocks over three years for the EV-EC deciduous site with observations of NEE fluxes and LAI assimilated. The upper panel shows best fit or mean for C_f, with observations marked, and the lower panel shows the width of the 90% confidence interval.

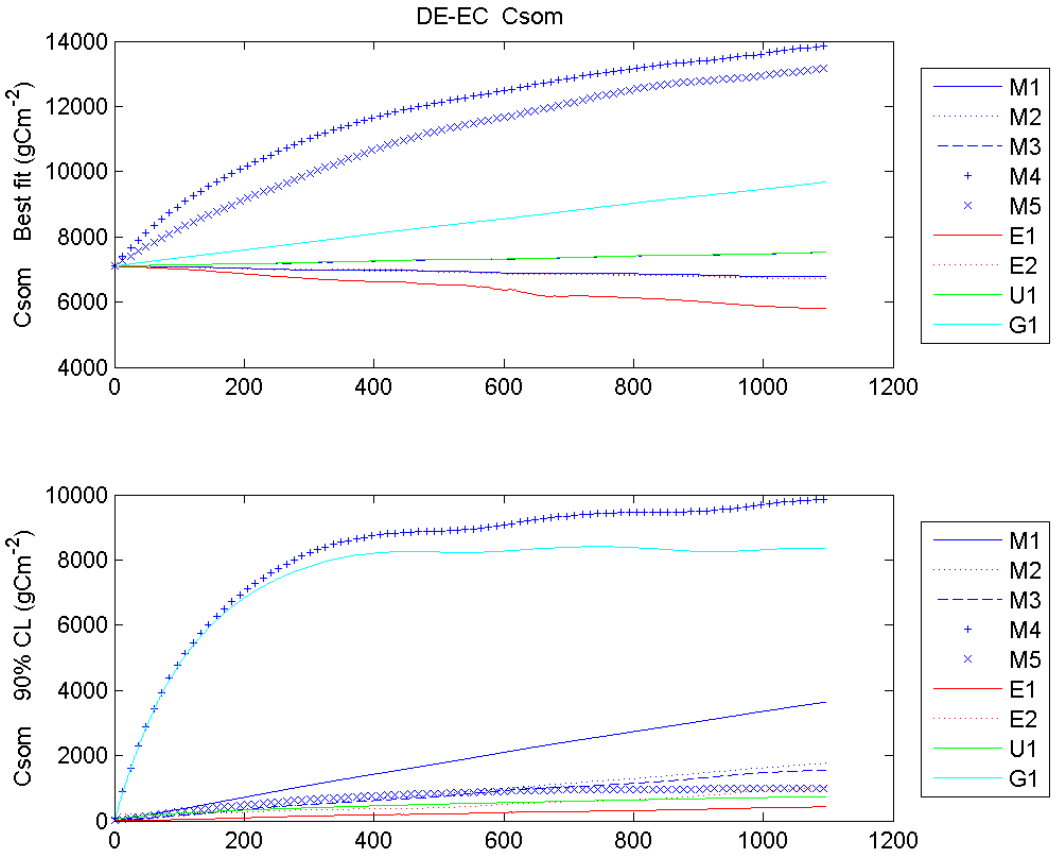

Figure 9. Retrieved estimates of soil organic matter/coarse woody debris C stocks over three years for the DE-EC deciduous site with observations of NEE fluxes and LAI assimilated. The upper panel shows best fit or mean for C_{som}, and the lower panel shows the width of the 90% confidence interval.

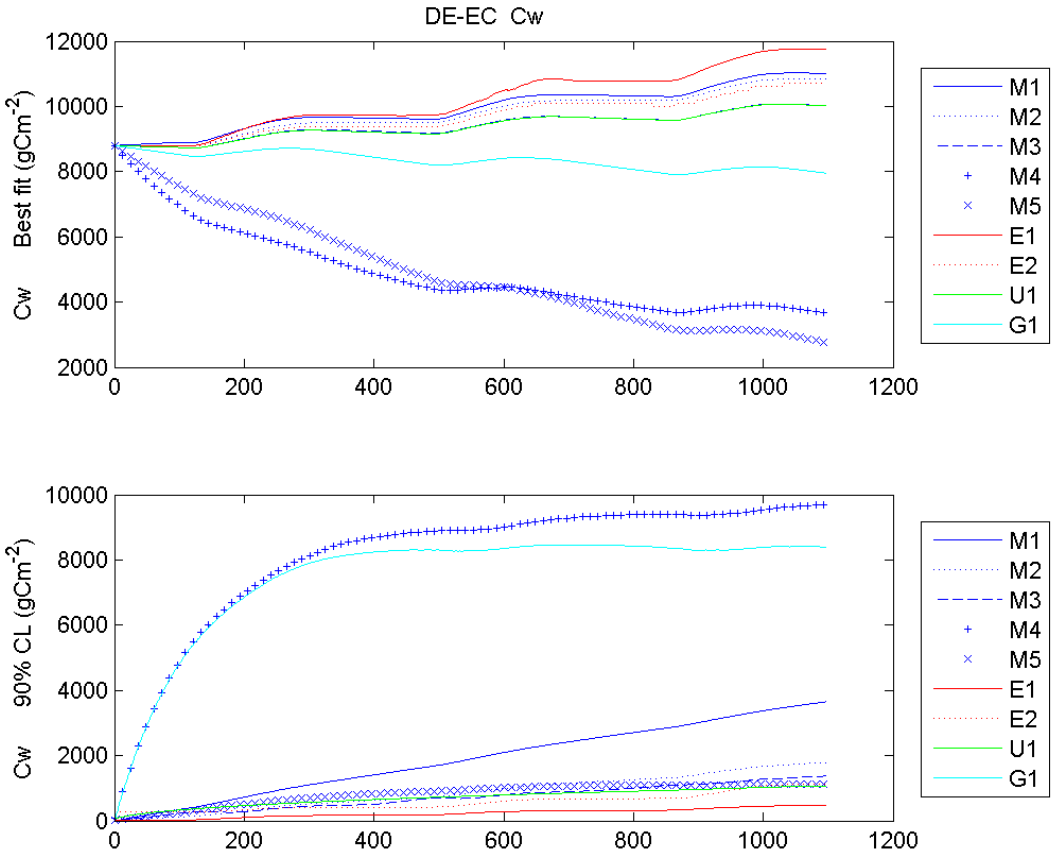

Figure 10. Retrieved estimates of woody C stocks over three years for the DE-EC deciduous site with observations of NEE fluxes and LAI assimilated. The upper panel shows best fit or mean for C_w, and the lower panel shows the width of the 90% confidence interval.