

*The REFLEX project: Comparing different algorithms and implementations for the inversion of a terrestrial ecosystem model against eddy covariance data*

Article

Accepted Version

Fox, A., Willaims, M., Richardson, A. D., Cameron, D., Gove, J. H., Quaife, T. ORCID: <https://orcid.org/0000-0001-6896-4613>, Ricciuto, D., Reichstein, M., Tomelleri, E., Trudinger, C. M. and Van Wijk, M. 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 (10). pp. 1597-1615. ISSN 0168-1923 doi: 10.1016/j.agrformet.2009.05.002 Available at <https://centaur.reading.ac.uk/28464/>

It is advisable to refer to the publisher's version if you intend to cite from the work. See [Guidance on citing](#).

To link to this article DOI: <http://dx.doi.org/10.1016/j.agrformet.2009.05.002>

Publisher: Elsevier

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the [End User Agreement](#).

[www.reading.ac.uk/centaur](http://www.reading.ac.uk/centaur)

## **CentAUR**

Central Archive at the University of Reading

Reading's research outputs online

## Article (refereed)

---

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.

Copyright © 2009 Elsevier B.V.

This version available <http://nora.nerc.ac.uk/9438/>

NERC has developed NORA to enable users to access research outputs wholly or partially funded by NERC. Copyright and other rights for material on this site are retained by the authors and/or other rights owners. Users should read the terms and conditions of use of this material at <http://nora.nerc.ac.uk/policies.html#access>

**This document is the author's final manuscript version of the journal article, incorporating any revisions agreed during the peer review process. Some differences between this and the publisher's version remain. You are advised to consult the publisher's version if you wish to cite from this article.**

[www.elsevier.com/](http://www.elsevier.com/)

Contact CEH NORA team at  
[noraceh@ceh.ac.uk](mailto:noraceh@ceh.ac.uk)

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

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

7  
8 <sup>1</sup> School of Applied Maths, Centre for Terrestrial Carbon Dynamics, University of Sheffield,  
9 Sheffield, UK

10 <sup>2</sup> School of GeoSciences, Centre for Terrestrial Carbon Dynamics, University of Edinburgh,  
11 Edinburgh, UK

12 <sup>3</sup> Complex Systems Research Center, University of New Hampshire, Durham, NH, USA

13 <sup>4</sup> Centre for Ecology and Hydrology, Bush Estate, Penicuik, Midlothian, UK

14 <sup>5</sup> USDA Forest Service, Northern Research Station, Durham, NH, USA

15 <sup>6</sup> Centre for Terrestrial Carbon Dynamics, Department of Geography, UCL, London

16 <sup>6</sup> Oak Ridge National Laboratory, Oak Ridge, TN, USA

17 <sup>8</sup> Max Planck Institute for Biogeochemistry, Jena, Germany

18 <sup>9</sup> CSIRO Marine and Atmospheric Research, Centre for Australian Weather and Climate  
19 Research, Aspendale, Victoria, Australia

20 <sup>10</sup> Wageningen University, Plant Sciences, The Netherlands

21

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

23 Key words: Data assimilation, metropolis, carbon cycle, ecosystem modelling, Monte Carlo,

24 Kalman Filter, eddy covariance, Reflex project

Fox, Williams et al.

25 \* Correspondence to Mathew Williams, [mat.williams@ed.ac.uk](mailto:mat.williams@ed.ac.uk), +44 131 667 8631 (tel) +44

26 131 662 0478 (fax)

27

28 **Abstract**

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

Fox, Williams et al.

53 large, slow C pools (wood and soil) would help to reduce uncertainties for model parameters

54 poorly served by eddy covariance data.

55

## 56 **Introduction**

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

64       Model-data comparison provides an opportunity to highlight areas (in space or time) of  
65 poor process representation, and to guide model improvement. Thus, the modelling  
66 community is now seeking to test its terrestrial ecosystem models against the growing array of  
67 observations (Bonan 2008). One of the critical datasets to be used in evaluating ecosystem  
68 models is the FLUXNET database (Baldocchi et al. 2001). FLUXNET is an international  
69 network of eddy covariance (EC) flux measurement towers. There are data sets from hundreds  
70 of sites worldwide, some with more than a decade of data collection. However, these data are  
71 associated with uncertainties and complications. There are gaps in time series that must be  
72 filled to obtain integrated (daily to annual) flux sums (Moffat et al. 2007). Also large areas of  
73 the globe are poorly sampled, and measurements are affected by systematic and random errors  
74 (Lasslop et al. 2008, Richardson et al. 2008), both of which can be large. EC towers measure  
75 net ecosystem exchanges (NEE) of CO<sub>2</sub>, meaning that the underlying processes of  
76 photosynthesis (GPP) and ecosystem respiration ( $R_e$ ) are not directly measured during  
77 daytime (Desai et al. 2008).

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



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

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

100 The capabilities and weaknesses of the various existing MDF approaches remain poorly  
101 understood. One recent study, the OptIC experiment, used pseudo-data from a highly  
102 simplified test model with 4 parameters to compare parameter estimation methods (Trudinger  
103 et al. 2007). OptIC found different methods equally successful, but that the choice of the cost  
104 function (quantifying the model-data mismatch) caused the most variation in the estimated  
105 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

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

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

## 148 **Methods**

### 149 **Model Description**

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

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

- 161 1. All C fixed during a day is expended either in autotrophic respiration or else allocated to  
162 one of the three plant tissue pools,  $C_f$ ,  $C_w$  or  $C_r$ .
- 163 2. Autotrophic respiration is a constant fraction of the C fixed during a day (Waring et al.  
164 1998).
- 165 3. Allocation fractions to vegetation pools are donor-controlled functions which have  
166 constant rate parameters.
- 167 4. For the deciduous model, the timing of initial leaf-out is controlled by a simple growing  
168 degree day accumulation, and leaf-fall by a minimum temperature threshold. The  
169 maximum amount of C that can be allocated to leaves is also limited by a parameter  
170 ( $c_{fmax}$ )
- 171 5. All C losses are via mineralization (i.e. no dissolved losses).

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

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

## 186 **Datasets**

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

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

202 For the observed data, the sites (Loobos, Netherlands and Hesse, France) and site-years  
203 (2000-2) were selected on the basis of relatively long, continuous records of fluxes and site  
204 meteorology, good quality control, and little or no drought stress. The observed data included

205 eddy covariance (EC) data, LAI data, and local meteorological data from a deciduous broad-  
206 leaf forest (identified as DE-EC) and an evergreen needle-leaf forest (EV-EC). Daily NEE  
207 was calculated by summing half-hourly observations, but only if >43 of the possible 48  
208 observations passed quality control. Missing data were filled by the daily mean of remaining  
209 data. It is possible that some small bias was introduced by this simple gap-filling, but for the  
210 purposes of this study such impacts were deemed insignificant. Typical data coverage was 20-  
211 30% of days. LAI data were sparse, usually collected on just a few days. Gap-filled flux data  
212 were not used in this experiment, but complete daily meteorological data were required to  
213 drive the model, and so gap filled weather data were used. All data were obtained via the  
214 FLUXNET site ([www.fluxnet.ornl.gov](http://www.fluxnet.ornl.gov)), from relevant, site specific literature and/or from site  
215 PIs. Three sequential years of data were assembled, of which the first two years were  
216 provided to participants. The source of the EC data was withheld from participants.

## 217 **Experiments**

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

Fox, Williams et al.

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

### 233 **Algorithms**

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

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

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

## 264 **Analyses**

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

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

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

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

274 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

275 closer the value  $D$  is to zero, the better the parameter is estimated, according to this measure.

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



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

$$280 \quad d_4: \quad \left( \sum_{= p_{\max i} - \gamma_{\min i}} \frac{CI_i}{=} \right) n$$

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

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

$$286 \quad d_5: \quad t - \left( \sum_{= p_{\max i} - \gamma_{\min i}} \frac{m_i}{=} \right) n \quad |t - \mu(m_1, \dots, m_n)| / (p_{\max} - p_{\min})$$

287 where  $m_x$  is parameter  $x$ 's best estimate by the algorithm.

288

## 289 **Results**

### 290 **Parameter estimation**

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

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

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

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

316 There was a significant correlation of  $d_1$  distances between EC and SYN for EV ( $r=0.73$ ,  
317  $P=0.01$ ) but not DE ( $r=0.31$ ,  $P=0.24$ ). So the EV parameters that were consistently estimated  
318 (across methods) were similar for synthetic and eddy covariance data, while this was not so  
319 for DE datasets, perhaps because of the greater number of parameters. There was a significant  
320 correlation between EC and SYN  $d_2$  distances for both EV ( $r=0.87$ ,  $P=0.0004$ ) and DE ( $r =$   
321  $0.84$ ,  $P<0.0001$ ). Thus parameters that were well constrained (low  $d_2$ ) by the synthetic data  
322 were well constrained by the eddy covariance data.

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

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

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

#### 346 **Flux estimates – synthetic data**

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

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

#### 356 **Flux estimates – observed data**

357 For the eddy covariance (EC) datasets, the algorithms' predictions were compared to  
358 observed NEE. In years 1 and 2, when observations were provided to participants, RMSEs  
359 varied from  $0.7\text{-}1.8 \text{ gC m}^{-2} \text{ d}^{-1}$  (DE) or  $0.6\text{-}0.9 \text{ gC m}^{-2} \text{ d}^{-1}$  (EV), with a mean value of  $1.3 \text{ gC}$   
360  $\text{m}^{-2} \text{ d}^{-1}$  for DE datasets and  $0.7 \text{ gC m}^{-2} \text{ d}^{-1}$  for EV. In year 3, when observations were not  
361 provided to participants, RMSEs varied from  $1.1\text{-}2.3 \text{ gC m}^{-2} \text{ d}^{-1}$  (DE) or  $1.3\text{-}1.7 \text{ gC m}^{-2} \text{ d}^{-1}$   
362 (EV), with a mean value of  $1.5 \text{ gC m}^{-2} \text{ d}^{-1}$  for both EC and DE datasets (  
363 Table 9). Thus the best NEE estimates of the algorithms tended to agree less well in the  
364 prognostic period (year 3) compared to the assimilation period (years 1-2), though this was  
365 most striking for the evergreen (EV) case in this study.

#### 366 **Flux confidence intervals – daily data**

367 There was less agreement among algorithms in the assessment of 90% confidence intervals  
368 (CI) on daily fluxes (Figure 4) than in the assessment of best estimates. There were  
369 differences in confidence interval estimates both in magnitudes and in temporal variability  
370 among algorithms. For instance, the mean daily 90% CI varied among algorithms from  $0.35 -$   
371  $1.92 \text{ gC m}^{-2} \text{ d}^{-1}$  in DE-SYN and  $0.29 - 2.49 \text{ gC m}^{-2} \text{ d}^{-1}$  in DE-EC. Algorithm confidence  
372 intervals typically had large excursions during spring leaf out for DE, but the magnitude of  
373 these excursions varied (Figure 4).

374 We tested whether the 90% CI on daily analyses (years 1 and 2) and predictions (year 3)  
375 encompassed the truth from the synthetic datasets for NEE, GPP and  $R_e$  for all years, and for  
376 observed NEE in year 3 for the EC datasets. The days of each year which passed this test  
377 were counted. We expected that 85-95% of the days would pass, roughly consistent with the  
378 magnitude of the confidence interval, 90%. For the synthetic experiments (NEE tests are  
379 shown in  
380 Table 8) this was rarely the case. In some cases the fraction was 100%, which indicates that  
381 the daily CI were likely set too large. In other cases, the fractions were <85% suggesting that  
382 the CI were too small or the predictions were biased. For the eddy covariance datasets in year  
383 3, the majority of algorithms' confidence intervals on daily NEE were too narrow, with an  
384 average of only 40% (DE) or 20% (EV) of the observed year 3 data lying within the 90%  
385 confidence interval (  
386 Table 9). This result suggests the algorithms were over-confident in the assessments of daily  
387 fluxes.

### 388 **Flux confidence intervals – annual sums**

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

399 year 2, and 35 had larger CI on year 3 than year 2, so this pattern was general across  
400 algorithms and datasets. Averaged over all cases, the 90% CI in the prediction period (year 3)  
401 were 88% larger than in the second year of the assimilation period (year 2). Patterns were  
402 similar in comparison between outputs from observed and synthetic datasets. However, mean  
403 90% CI across all algorithms were ~31% larger for EC datasets than for SYN datasets.  
404 Among algorithms, the increase in 90% CI on EC datasets compared to SYN datasets ranged  
405 from 0% (E1) to 100% (E2).

#### 406 **Testing annual flux estimates and confidence intervals**

407 Annual flux outputs estimated and forecast using the synthetic datasets were compared with  
408 the synthetic truth. Each algorithm's annual output of NEE, GPP and  $R_e$  was tested to  
409 determine whether the truth lay within the 90% CI for estimates. The fraction of tests that  
410 were successful was compared with the mean size of the 90% confidence interval for each  
411 specific algorithm (Figure 5). There was evidence of a positive relationship between success  
412 rate and confidence interval size, but some algorithms managed to contain the truth within  
413 relatively narrow confidence intervals. In the comparison for annual NEE, four algorithms  
414 (E1, E2, M1, M3) produced analyses with >80% success rate and mean confidence intervals  
415 <110 gC m<sup>-2</sup> yr<sup>-1</sup>. In the comparison against component fluxes (GPP and  $R_e$ ), two algorithms  
416 (E2, M2) produced more balanced analyses, with relatively high success rates (>65%) and  
417 narrow confidence intervals (<300 gC m<sup>-2</sup> yr<sup>-1</sup>). M3 was always 100% successful in  
418 containing the truth within its 90% confidence intervals, and this over-confidence was  
419 because associated CI were the largest of all algorithms for GPP and  $R_e$ . There were  
420 successful tests for prognoses in year 3 by several algorithms, indicating that predictions of C  
421 fluxes beyond the observational period were successful also (  
422 Table 8).

423 **GPP and  $R_e$  estimates**

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

434 **Stocks**

435 The analyses and predictions of foliar C matched the seasonal cycles and magnitudes of the  
436 truth from the synthetic studies adequately (Figure 8). Predictions of year 3 foliar C in the  
437 eddy covariance datasets had a mean RMSE among algorithms of  $11 \text{ gC m}^{-2}$  for DE and  $22$   
438  $\text{gC m}^{-2}$  for EV. However, assessments of confidence intervals were generally poor; most  
439 algorithms had 90% CI either too broad or too narrow (  
440 Table 9).

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

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

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

## 453 **Discussion**

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

## 463 **Parameter estimation**

464 We expected that parameters linked to fast-response processes that mostly determine net  
465 ecosystem exchange of CO<sub>2</sub> (NEE) would be well constrained and well characterised, while  
466 parameters for slow processes would be poorly characterised. Our analyses largely supported  
467 this expectation. The turnover of litter and foliage were well estimated according to our  
468 criteria, and these parameters are closely associated with foliage mass and/or gaseous  
469 exchanges of C. We had not expected the turnover rate of SOM, a large slow turnover pool, to  
470 be so well constrained, but it is an important determinant of heterotrophic respiration  
471 nevertheless. Parameters associated with the turnover of wood and allocation to roots were



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

#### 474 **Flux and stock estimation**

475 There was weak agreement among algorithms in estimations of 90% CI on NEE and its  
476 component fluxes, for all datasets. The differences in CI size were closely related to  
477 differences among algorithms in parameter confidence intervals. There were considerable  
478 differences in assessments among similar algorithms (e.g. Metropolis), suggesting that the  
479 subjective choices of convergence tests versus statistical tests, priors for the parameters, and  
480 likelihood function within the method were important determinants of CI. None of the  
481 algorithms consistently included ~90% of the synthetic true daily NEE values, or observed  
482 EC year 3 daily NEE data, within the 90% confidence interval of the best-fit NEE (Table  
483 8 and Table 9). All algorithms at some point over- or underestimated the confidence interval.  
484 For annual assessments of NEE, GPP and  $R_e$ , there was more success, with some algorithms  
485 successful locating the true value from synthetic experiments within relatively narrow 90%  
486 confidence intervals (Figure 5).

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

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

#### 501 **Model error**

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

#### 512 **Prediction error**

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

521 were > twice those for year 2 analyses. For the SYN experiments, the year 3 predictions  
522 among algorithms were similarly successful to years 1 and 2 – that is, a similar fraction of  
523 90% confidence intervals on annual flux estimates encompassed the truth. This result suggests  
524 that the quantification of increasing CI was reasonable.

### 525 **Algorithm assessment**

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

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

546 the size of the CI depend on the accept/reject criterion employed by each algorithm in  
547 generating acceptable parameter sets (Table 5). The methods employing the Kalman filter  
548 employed a further step, once acceptable parameter sets were determined. The state variables  
549 of the model, including flux estimates, were updated using sequential assimilation of  
550 observations through the times series. This sequential updating, allowing shifts in states  
551 through the model run unconnected to parameters, may be connected to the success of the  
552 Kalman filter methods (E1, E2, U1) in generating effective, but narrow confidence intervals.  
553 Some algorithms had problems with large changes to  $C_w$  and  $C_{som}$  pools, which could be  
554 made consistent with the flux data, but are not ecologically sound in an undisturbed  
555 ecosystem. This seems to be partly related to a steady state assumption being made where  
556 pool sizes are first confined to equilibrium which likely leads to a wrong initial system state,  
557 potential biases in parameters and inflation of their confidence intervals as shown recently in  
558 a specific study by Calvalhais et al. (2008). These symptoms are, for example, also seen in the  
559 approach M4, where a spin-up was performed. Hence, a way to estimate the initial state of the  
560 system without an *ad-hoc* steady state assumption is crucial to successful MDF and should be  
561 explored further. A constraint on the annual changes in these pools based on repeated  
562 inventories would help solve this problem. Stem inventories are likely to be easier to  
563 undertake with quantifiable error than those on SOM, and so should be the focus for future  
564 studies. Nevertheless, if longer time scale are to be addressed there is a need to imposed  
565 constraints from soil carbon data, e.g. via chronosequences or profile data. Some algorithms  
566 did not explicitly include searching for initial conditions on  $C_f$ ,  $C_{lab}$  and  $C_{lit}$ , and this caused  
567 some problems for e.g. E2. All algorithms need to assess their estimates of uncertainties and  
568 develop new approaches for uncertainty estimates that are consistent with the observations.  
569 This experiment has demonstrated the value of using synthetic datasets in understanding data  
570 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

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

## 601 **Acknowledgements**

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

## 616 **References**

617 Baldocchi, D., E. Falge, L. H. Gu, R. Olson, D. Hollinger, S. Running, P. Anthoni, C. Bernhofer, K. Davis, R.  
618 Evans, J. Fuentes, A. Goldstein, G. Katul, B. Law, X. H. Lee, Y. Malhi, T. Meyers, W. Munger, W.  
619 Oechel, K. T. P. U, K. Pilegaard, H. P. Schmid, R. Valentini, S. Verma, T. Vesala, K. Wilson, and S.  
620 Wofsy. 2001. FLUXNET: A new tool to study the temporal and spatial variability of ecosystem-scale  
621 carbon dioxide, water vapor, and energy flux densities. *Bulletin of the American Meteorological*  
622 *Society* **82**:2415-2434.

- 623 Bonan, G. B. 2008. Forests and Climate Change: Forcings, Feedbacks, and the Climate Benefits of Forests.  
624 *Science* **320**:1444-1449.
- 625 Braswell, B. H., W. J. Sacks, E. Linder, and D. S. Schimel. 2005. Estimating diurnal to annual ecosystem  
626 parameters by synthesis of a carbon flux model with eddy covariance net ecosystem exchange  
627 observations. *Global Change Biology* **11**:335 - 355.
- 628 Carvalhais, N., M. Reichstein, S. J., C. G.J., J. Santos Pereira, P. Berbigier, A. Carrara, A. Granier, L.  
629 Montagnani, D. Papale, S. Rambal, M. J. Sanz, and R. Valentini. 2008. Implications of Carbon Cycle  
630 Steady State Assumptions for Biogeochemical Modeling Performance and Inverse Parameter Retrieval.  
631 *Global Biogeochemical Cycles* **22**, GB2007.
- 632 Desai, A. R., A. D. Richardson, A. M. Moffat, J. Kattge, D. Y. Hollinger, A. Barr, E. Falge, A. Noormets, D.  
633 Papale, M. Reichstein, and V. J. Stauch. 2008. Cross site evaluation of eddy covariance GPP and RE  
634 decomposition techniques. *Agricultural and Forest Meteorology* **148**:821-838.
- 635 Friedlingstein, P., P. Cox, R. Betts, L. Bopp, W. Von Bloh, V. Brovkin, P. Cadule, S. Doney, M. Eby, I. Fung,  
636 G. Bala, J. John, C. Jones, F. Joos, T. Kato, M. Kawamiya, W. Knorr, K. Lindsay, H. D. Matthews, T.  
637 Raddatz, P. Rayner, C. Reick, E. Roeckner, K. G. Schnitzler, R. Schnur, K. Strassmann, A. J. Weaver,  
638 C. Yoshikawa, and N. Zeng. 2006. Climate-carbon cycle feedback analysis: Results from the (CMIP)-  
639 M-4 model intercomparison. *Journal of Climate* **19**:3337-3353.
- 640 Gelman, A. 1995. Bayesian data analysis. Chapman & Hall, London.
- 641 Gove, J. H., and D. Y. Hollinger. 2006. Application of a dual unscented Kalman filter for simultaneous state and  
642 parameter estimation in problems of surface-atmosphere exchange. *Journal of Geophysical Research*  
643 **111**:doi:10.1029/2005JD00621.
- 644 Hagen, S. C., B. H. Braswell, E. Linder, S. Frohling, A. D. Richardson, and D. Hollinger. 2006. Statistical  
645 uncertainty of eddy flux-based estimates of gross ecosystem carbon exchange at Howland Forest,  
646 Maine. *Journal of Geophysical Research* **111**:D08S03.
- 647 Heimann, M., and M. Reichstein. 2008. Terrestrial ecosystem carbon dynamics and climate feedbacks. *Nature*  
648 **451**:289-292.
- 649 Janssen, P. H. M., and P. S. C. Heuberger. 1995. Calibration of process-oriented models. *Ecological Modelling*  
650 **83**:55 - 66.
- 651 Julier, S. J., and J. K. Uhlmann. 2004. Unscented filtering and nonlinear estimation. *Proc. IEEE* **92**:410-422.
- 652 Kalman, R. E. 1960. A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME--*  
653 *Journal of Basic Engineering* **82**:35-45.
- 654 Knorr, W., and J. Kattge. 2005. Inversion of terrestrial ecosystem model parameter values against eddy  
655 covariance measurements by Monte Carlo sampling. *Global Change Biology* **11**:1333-1351.
- 656 Lasslop, G., M. Reichstein, J. Kattge, and D. Papale. 2008. Influences of observation errors in eddy flux data on  
657 inverse model parameter estimation. *Biogeosciences* **5**:1311-1324.
- 658 Liu, Y., and H. V. Gupta. 2007. Uncertainty in hydrologic modeling: Toward an integrated data assimilation  
659 framework. *Water Resources Research* **43**:W07401.
- 660 Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equation of state  
661 calculations by fast computing machines. *J. Chem. Phys.* **21**:1087-1092.
- 662 Moffat, A. M., D. Papale, M. Reichstein, D. Y. Hollinger, A. D. Richardson, A. G. Barr, C. Beckstein, B. H.  
663 Braswell, G. Churkina, A. R. Desai, E. Falge, J. H. Gove, M. Heimann, D. Hui, A. J. Jarvis, J. Kattge,  
664 A. Noormets, and V. J. Stauch. 2007. Comprehensive comparison of gap-filling techniques for eddy  
665 covariance net carbon fluxes. , 147: 209-232. *Agricultural and Forest Meteorology* **147**:209-232.
- 666 Moore, D., J. Hu, W. Sacks, D. Schimel, and R. Monson. in press. Estimating transpiration and the sensitivity of  
667 carbon uptake to water availability in a subalpine forest using a simple ecosystem process model  
668 informed by measured net CO<sub>2</sub> and H<sub>2</sub>O fluxes. *Agricultural and Forest Meteorology*.
- 669 Mosegaard, K., and A. Tarantola. 1995. Monte-Carlo Sampling of Solutions to Inverse Problems. *Journal of*  
670 *Geophysical Research-Solid Earth* **100**:12431-12447.
- 671 Quaife, T., P. Lewis, M. De Kauwe, M. Williams, B. E. Law, M. D. Disney, and P. Bowyer. 2008. Assimilating  
672 Canopy Reflectance data into an Ecosystem Model with an Ensemble Kalman Filter. *Remote Sensing*  
673 *of the Environment* **111**:1347-1364.
- 674 Raupach, M. R., P. J. Rayner, D. J. Barrett, R. S. DeFries, M. Heimann, D. S. Ojima, S. Quegan, and C. C.  
675 Schimullius. 2005. Model-data synthesis in terrestrial carbon observation: methods, data requirements  
676 and data uncertainty specifications. *Global Change Biology* **11**:378-397.
- 677 Ricciuto, D. M., K. J. Davis, and K. Keller. 2008. A Bayesian calibration of a simple carbon cycle model: The  
678 role of observations in estimating and reducing uncertainty. *Global Biogeochemical Cycles* **22**:GB2030.
- 679 Richardson, A. D., D. Y. Hollinger, G. G. Burba, K. J. Davis, L. B. Flanagan, G. G. Katul, J. W. Munger, D. M.  
680 Ricciuto, P. C. Stoy, A. E. Suyker, S. B. Verma, and S. C. Wofsy. 2006. A multi-site analysis of  
681 random error in tower-based measurements of carbon and energy fluxes. *Agricultural and Forest*  
682 *Meteorology* **136**:1-18.

- 683 Richardson, A. D., M. D. Mahecha, E. Falge, J. Kattge, A. M. Moffat, D. Papale, M. Reichstein, V. J. Stauch, B.  
684 H. Braswell, G. Churkina, B. Kruijt, and D. Y. Hollinger. 2008. Statistical properties of random CO<sub>2</sub>  
685 flux measurement uncertainty inferred from model residuals. *Agricultural and Forest Meteorology*  
686 **148**:38-50.
- 687 Sacks, W. J., D. S. Schimel, and R. K. Monson. 2007. Coupling between carbon cycling and climate in a high-  
688 elevation, subalpine forest: a model-data fusion analysis. *Oecologia* **151**:54-68.
- 689 Stockli, R., D. M. Lawrence, G. Y. Niu, K. W. Oleson, P. E. Thornton, Z. L. Yang, G. B. Bonan, A. S. Denning,  
690 and S. W. Running. 2008. Use of FLUXNET in the community land model development. *Journal of*  
691 *Geophysical Research-Biogeosciences* **113**.
- 692 Trudinger, C. M., M. R. Raupach, P. J. Rayner, and I. G. Enting. 2008. Using the Kalman filter for parameter  
693 estimation in biogeochemical models. *Environmetrics* DOI: **10.1002/env.910**.
- 694 Trudinger, C. M., M. R. Raupach, P. J. Rayner, J. Kattge, Q. Liu, B. Pak, M. Reichstein, L. Renzullo, A. D.  
695 Richardson, S. H. Roxburgh, J. Styles, Y. P. Wang, P. Briggs, D. Barrett, and S. Nikolova. 2007. OptIC  
696 project: An intercomparison of optimization techniques for parameter estimation in terrestrial  
697 biogeochemical models. *Journal of Geophysical Research-Biogeosciences* **112**.
- 698 Van Oijen, M., J. Rougier, and R. Smith. 2005. Bayesian calibration of process-based forest models: bridging the  
699 gap between models and data. *Tree Physiology* **25**:915-927.
- 700 Van Wijk, M. T., and W. Bouten. 2001. Towards understanding tree root profiles: simulating hydrologically  
701 optimal strategies for root distribution. *Hydrology and Earth System Sciences* **5(4)**:629 - 644.
- 702 Van Wijk, M. T., and W. Bouten. 2002. Simulating daily and half-hourly fluxes of forest carbon dioxide and  
703 water vapor exchange with a simple model of light and water use. *Ecosystems* **5**:597-610.
- 704 Van Wijk, M. T., B. Van Putten, D. Y. Hollinger, and A. D. Richardson. 2008. Comparison of different  
705 objective functions for parameterization of simple respiration models. *Journal of Geophysical Research*  
706 **113**:G03008.
- 707 Vrugt, J. A., C. G. H. Diks, H. V. Gupta, W. Bouten, and J. M. Verstraten. 2005. Improved treatment of  
708 uncertainty in hydrologic modeling: Combining the strengths of global optimization and data  
709 assimilation. *Water Resources Research* **41**:W01017.
- 710 Vrugt, J. A., H. V. Gupta, W. Bouten, and S. Sorooshian. 2003. A shuffled complex evolution metropolis  
711 algorithm for optimization and uncertainty assessment of hydrologic model parameters. *Water*  
712 *Resources Research* **39**:Art. no. 1201.
- 713 Wang, Y. P., D. Baldocchi, R. Leuning, E. Falge, and T. Vesala. 2007. Estimating parameters in a land-surface  
714 model by applying nonlinear inversion to eddy covariance flux measurements from eight FLUXNET  
715 sites. *Global Change Biology* **13**:652 - 670.
- 716 Waring, R. H., J. J. Landsberg, and M. Williams. 1998. Net primary production of forests: a constant fraction of  
717 gross primary production? *Tree Physiology* **18**:129-134.
- 718 Williams, M., E. B. Rastetter, D. N. Fernandes, M. L. Goulden, G. R. Shaver, and L. C. Johnson. 1997.  
719 Predicting gross primary productivity in terrestrial ecosystems. *Ecological Applications* **7**:882-894.
- 720 Williams, M., E. B. Rastetter, D. N. Fernandes, M. L. Goulden, S. C. Wofsy, G. R. Shaver, J. M. Melillo, J. W.  
721 Munger, S.-M. Fan, and K. J. Nadelhoffer. 1996. Modelling the soil-plant-atmosphere continuum in a  
722 *Quercus-Acer* stand at Harvard Forest: the regulation of stomatal conductance by light, nitrogen and  
723 soil/plant hydraulic properties. *Plant, Cell and Environment* **19**:911-927.
- 724 Williams, M., P. Schwarz, B. E. Law, J. Irvine, and M. R. Kurpius. 2005. An improved analysis of forest carbon  
725 dynamics using data assimilation. *Global Change Biology* **11**:89-105.
- 726 Williams, M., L. Street, M. T. V. Wijk, and G. R. Shaver. 2006. Identifying differences in carbon exchange  
727 among arctic ecosystem types. *Ecosystems* **9**:288-304.
- 728  
729  
730



731 **Appendix: Details of algorithms**

732 E1: Stage 1. Parameter estimation. Parameters were initially estimated using a simple  
733 Metropolis MCMC-type approach. (e.g. (Mosegaard and Tarantola 1995, Knorr and Kattge  
734 2005)). Initial prior distributions were assumed to be uniform and encompass the entire  
735 possible suggested range and so a single stage accept/reject criterion was used based on  
736 comparison of model output with data alone. Initial values for  $C_r$ ,  $C_{lit}$  and  $C_{lab}$  were  
737 estimated in the same manner as parameters, initial values for  $C_f$  were based on first available  
738 observation (EV case) or set to zero (DE case). The model was initialized from a random  
739 location, and step size was constant and determined as 0.001 of log-transformed parameter  
740 range. This was determined through ‘tuning’ initial runs to ensure an acceptance probability  
741 of between 0.2 and 0.8 at each step. The number of steps required to sufficiently sample the  
742 parameter space was assessed using the Gelman criteria (Gelman 1995)) to test convergence  
743 between chains.

744 E1: Stage 2. State estimation. 8000 parameter sets were randomly sampled from the accepted  
745 parameters from Stage 1. These were then used in an 8000 member Ensemble Kalman Filter  
746 (EnKF, Evensen, 1994; Williams et al., 2005). A unique parameter set was assigned to each  
747 ensemble member with the intention this would cause divergence between ensemble members  
748 representing model error and cause a growth in state uncertainty equivalent to that inherent  
749 from parameter uncertainty alone. This was done instead of adding a stochastic forcing term  
750 at each time step. This is possibly correct in the SYN cases when model ‘structural’ error is  
751 known to be zero, but will probably underestimate model error in the EC cases and overly  
752 restrict growth in state uncertainty. Nonetheless, assimilation of observations did alter the  
753 state variables in the resulting analysis and reduce uncertainties in state estimates even though  
754 these same observation data had already been used to generate the parameter sets in Stage 1  
755 so offered little additional information to the EnKF.

756  
757 E2: Ensemble Kalman filter. This method was set up for joint estimation of states and  
758 parameters, so parameters were included within the state vector for assimilation. Model  
759 parameter errors were set within bounds - small enough to avoid tracking daily noise in  
760 observations, and large enough to shift over weekly-seasonal timescales in response to  
761 process signals. Errors on model states were set smaller than for parameters, so that  
762 assimilation was focused on updating parameters rather than states. Initial values for all  
763 parameters and initial conditions for  $C_f$ ,  $C_{lab}$  and  $C_r$  were estimated. After an initial  
764 assimilation of observations, these initial parameter estimates were updated with the final  
765 estimates from the assimilation. We assumed that  $C_f$ ,  $C_{lab}$  and  $C_r$  would be in approximate  
766 steady state over annual cycles, and adjusted initial values accordingly. A further EnKF  
767 assimilation was then applied, using these updated initial parameters and initial conditions, to  
768 generate final analyses.

769  
770 U1: Unscented Kalman filter. The UKF was used to provide state estimates for each of the  
771 experiments. The UKF (Julier and Uhlmann 2004) is a nonlinear version of the traditional  
772 linear Kalman filter (Kalman 1960), that uses a deterministic sampling of so-called sigma  
773 points in order to capture the mean and covariance of the state. Similar to other Kalman type  
774 filters it employs a two-step ‘predictor-corrector’ scheme where model predictions are  
775 corrected by measurements as they arrive sequentially in time. At time periods where  
776 measurements are missing, only the prediction step is used. To employ the UKF, the general  
777 nonlinear state space model was assumed, with the variants of the model taking the form of  
778 the state evolution equations. A linear measurement model was used in all runs. Both the state

779 and measurement equations assume zero mean random noise processes with associated full-  
 780 dimensional covariance matrices (Gove and Hollinger 2006). The later were estimated from  
 781 the information provided. The parameter estimates used in the filter runs were arrived at via  
 782 simulated annealing method M3. Parameters for the unscented transformation were set to  
 783  $\alpha=1$ ,  $\beta=2$  and  $\kappa=1$  for all experiments (see Gove and Hollinger 2006 for an explanation).  
 784

785 G1: Genetic algorithm: The implementation was from Haupt and Haupt (2004). The  
 786 population size was 100, and was run for 1000 iterations (generations). Initial stores Cr, Clit,  
 787 Cf and Clab were estimated by the GA as additional parameters. To estimate uncertainties, the  
 788 roughly 1600 (unique) parameter sets with cost function values closest to the final best cost  
 789 function value were saved, and used to estimate the covariance matrix and 90% CI.  
 790

791 M1: This method sought to make as few approximations as possible to Bayes Theorem,  
 792 choosing the simplest algorithm to generate a representative sample from the posterior. We  
 793 chose the beta distribution for our prior. The Metropolis algorithm (Metropolis et al. 1953)  
 794 generates a chain that sequentially “walks through parameter space” in such a way that the  
 795 chain of visited points is the sought-after sample from the posterior. Each new point in the  
 796 chain is found by randomly generating a multivariate normal step away from the current  
 797 vector. In this case a simple diagonal variance matrix defined this multivariate normal  
 798 “proposal distribution”. Whether a proposed candidate vector was accepted or not depended  
 799 on the *Metropolis ratio*, which is the ratio of two products: likelihood times prior for the  
 800 candidate and likelihood times prior for the current point. If the Metropolis ratio was larger  
 801 than 1 (i.e. the candidate point has a higher posterior probability then the current point), it was  
 802 always accepted. If the Metropolis ratio was less than 1 (i.e. the candidate was “less probable”  
 803 than the current vector), the candidate could still be accepted but only with probability equal  
 804 to the Metropolis ratio. The chain was stopped when it “converged”, i.e. it had explored the  
 805 parameter space adequately. Convergence was confirmed visually using the trace plots of the  
 806 different parameters, i.e. plots that show how the chain moves through parameter space for  
 807 each individual parameter. If one or more of the trace plots was still showing drift towards  
 808 unexplored parts or parameter space, the chain was deemed not to have converged.  
 809

810 M2: A combined optimization approach estimated model parameters and state variables. A  
 811 genetic algorithm, Stochastic Evolutionary Ranking Strategy (SRES) was used to find the  
 812 global optimum (Runarsson and Yao, 2000). Markov chain monte carlo (MCMC) using the  
 813 Metropolis-Hastings algorithm was then used to explore the parameter space around the  
 814 optimum to estimate the full joint distribution of parameters and to estimate predictive  
 815 uncertainty. Two chains were run for each experiment; convergence was determined by  
 816 visually comparing the parameter PDFs from both chains. The ranges given for p1-17 were  
 817 used as uniform distributions; no additional information was used. The initial values of pools  
 818  $C_r$ ,  $C_{lit}$  and  $C_{lab}$  were also estimated as model parameters, using the prior range 20-200  $gC\ m^{-2}$   
 819 as recommended. All observations are assumed to drawn from independent distributions.  
 820 Both NEE and LAI errors were assumed normally distributed.  
 821

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

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

834  
835 M4: Markov Chain Monte Carlo Metropolis. The algorithm adopted a global search method  
836 with an uniform walk in the model parameter space. The method is based on a Bayesian  
837 approach where the comparison between model output and data is used to update our prior  
838 knowledge of the parameter distribution. The prior distributions were considered to be  
839 uniform. The Metropolis rules prevented the algorithm from being trapped in local minima,  
840 allowing for changes in the searching direction. Spin-up was used to initialize the C pools ( $C_r$ ,  
841  $C_{lit}$  and  $C_{lab}$ ); we sampled the parameters and we ran the model replicating the meteorological  
842 data until the total difference between one year and the other was less than 1g of C. The other  
843 C pools were initialized as from the experiment description.

844  
845 M5. The SCEM-UA algorithm (Vrugt et al., 2003) is a modified version of the original SCE-  
846 UA global optimization algorithm (Duan et al., 1992). The algorithm is Bayesian in nature  
847 and operates by merging the strengths of the Metropolis algorithm, controlled random search,  
848 competitive evolution, and complex shuffling to continuously update the proposal distribution  
849 and evolve the sampler to the posterior target distribution. The SCEM-UA algorithm uses the  
850 Metropolis-Hastings (Metropolis et al., 1953) search strategy to generate a sequence of  
851 parameter sets ( $\theta_1, \theta_2, \dots, \theta_n$ ) that adapts to the target posterior distribution. It starts with an  
852 initial population of points (parameter sets) randomly distributed throughout the feasible  
853 parameter space defined by the prior parameter distributions. The population is partitioned  
854 into  $q$  complexes, and in each complex  $k$  ( $k = 1, 2, \dots, q$ ) a parallel sequence is launched from  
855 the point that exhibits the highest posterior density. A new candidate point in each sequence  $k$   
856 is generated using a multivariate normal distribution either centred around the current draw of  
857 the sequence  $k$ , or the mean of the points in complex  $k$ , augmented with the covariance  
858 structure induced between the points in complex  $k$ . The Metropolis-annealing criterion is used  
859 to test whether the candidate point should be added to the current sequence. Subsequently the  
860 new candidate point randomly replaces an existing member of the complex. Finally, after a  
861 certain number of iterations new complexes are formed through a process of shuffling the old  
862 complexes. The objective function used in this study is a combination of the model errors  
863 (expressed as SSE, Sum of Squared Errors) of describing the CO<sub>2</sub> fluxes and the Leaf Area  
864 Index, weighted by the error variance of each variable.

865

866 **Tables**

Observation	Units	Interval	Source
Global Radiation	MJ m <sup>-2</sup> d <sup>-1</sup>	Daily	Fluxnet data portal
Min Temperature	°C	Daily	Fluxnet data portal
Max temperature	°C	Daily	Fluxnet data portal
CO <sub>2</sub> conc.	μmol mol <sup>-1</sup>	Daily	Fluxnet data portal
NEE	g C m <sup>-2</sup> d <sup>-1</sup>	Daily	Fluxnet data portal
LAI	m <sup>2</sup> m <sup>-2</sup>	When available	References/site PI
Foliar N *	gN m <sup>-2</sup> leaf area	Constant	References/site PI
Aboveground C mass*	kg C m <sup>-2</sup>	Initial condition	References/site PI
SOM C mass*	kg C m <sup>-2</sup>	Initial condition	References/site PI
Leaf mass per area*	g C m <sup>-2</sup> 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 <sup>-2</sup> )	Above-ground biomass (g C m <sup>-2</sup> )	Leaf mass per area (g C m <sup>-2</sup> leaf area)	Foliar N (g N m <sup>-2</sup> 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)	$T_d$	$1 \times 10^{-6}/0.01$
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)	$T_f$	$1 \times 10^{-4}/0.1$
p6	Turnover rate of wood (per day)	$T_w$	$1 \times 10^{-6}/0.01$
p7	Turnover rate of roots (per day)	$T_r$	$1 \times 10^{-4}/0.01$
p8	Mineralisation rate of litter (per day)	$T_l$	$1 \times 10^{-5}/0.1$
p9	Mineralisation rate of SOM/CWD (per day)	$T_s$	$1 \times 10^{-6}/0.01$
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)	$T_{lab}$	$1 \times 10^{-4}/0.1$
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<sub>2</sub> 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	$J = \frac{1}{2} \sum_{i=1}^N f \frac{(x_i, p_i - OBS_i)^2}{\sigma_{obs,i}^2}$	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	$J = \sum_{i=1}^N \frac{(f_i - OBS_i)^2}{\sigma_{obs,i}^2} + a \times \left[ \frac{(c[365] - c[0])^2}{c[0]} \right] + \left[ \frac{(c[730] - c[0])^2}{c[0]} \right]$	Tuned parameters with	n/a	~100000		Fortran
M1	MCMC – Metropolis			Gaussian likelihood	Included calibration in	visual		300000	Fortran
M2	MCMC – Metropolis	MCMC1	uniform	$J = \frac{1}{2} \sum_{i=1}^N f \frac{(x_i, p_i - OBS_i)^2}{\sigma_{obs,i}^2}$	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	$J = \frac{1}{2} \sum_{i=1}^N f \frac{(x_i, p_i - OBS_i)^2}{\sigma_{obs,i}^2}$	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  $d_{1-3}$ . 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  $d_{1-2}$ . The rank column identifies the rank of  $D$  for each parameter, with lower values of  $D$ , and lower ranks, indicating better estimation.

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

**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.**

Algorithm	Foliar C mass ( $C_f$ )				Daily NEE			
	RSME ( $\text{gC m}^{-2}$ )		CI frac		RMSE ( $\text{gC m}^{-2} \text{d}^{-1}$ )		CI frac	
	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	<b>0.94</b>	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	<b>0.86</b>	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
<i>n</i>	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  $\text{gC m}^{-2} \text{yr}^{-1}$ .**

Figures

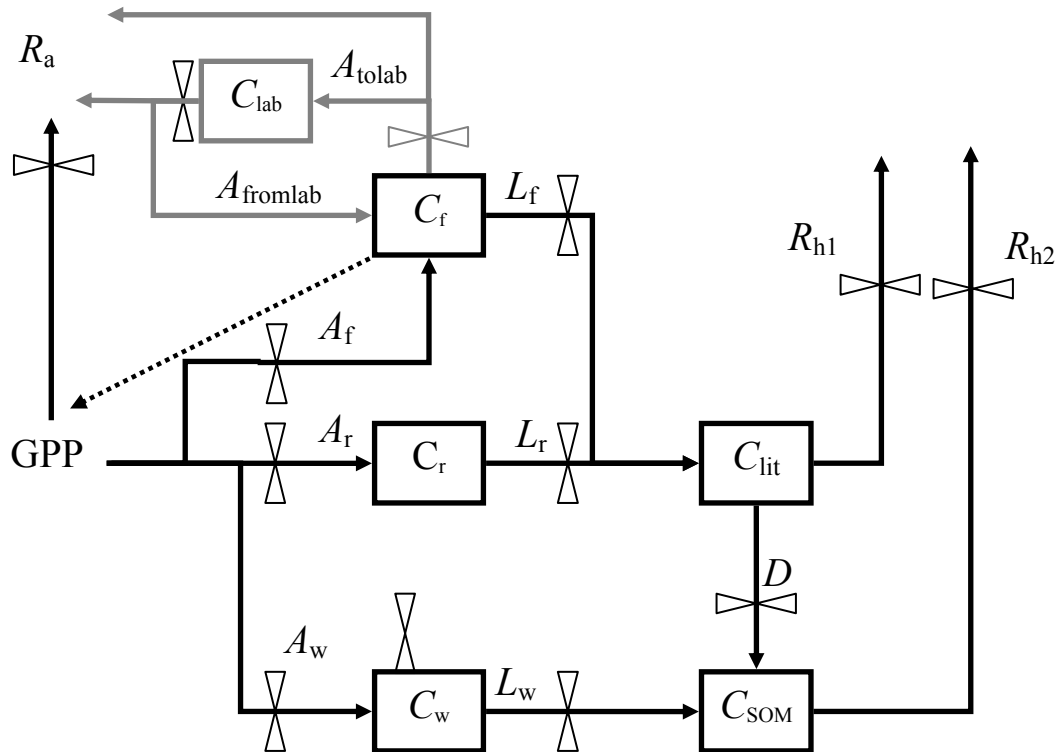
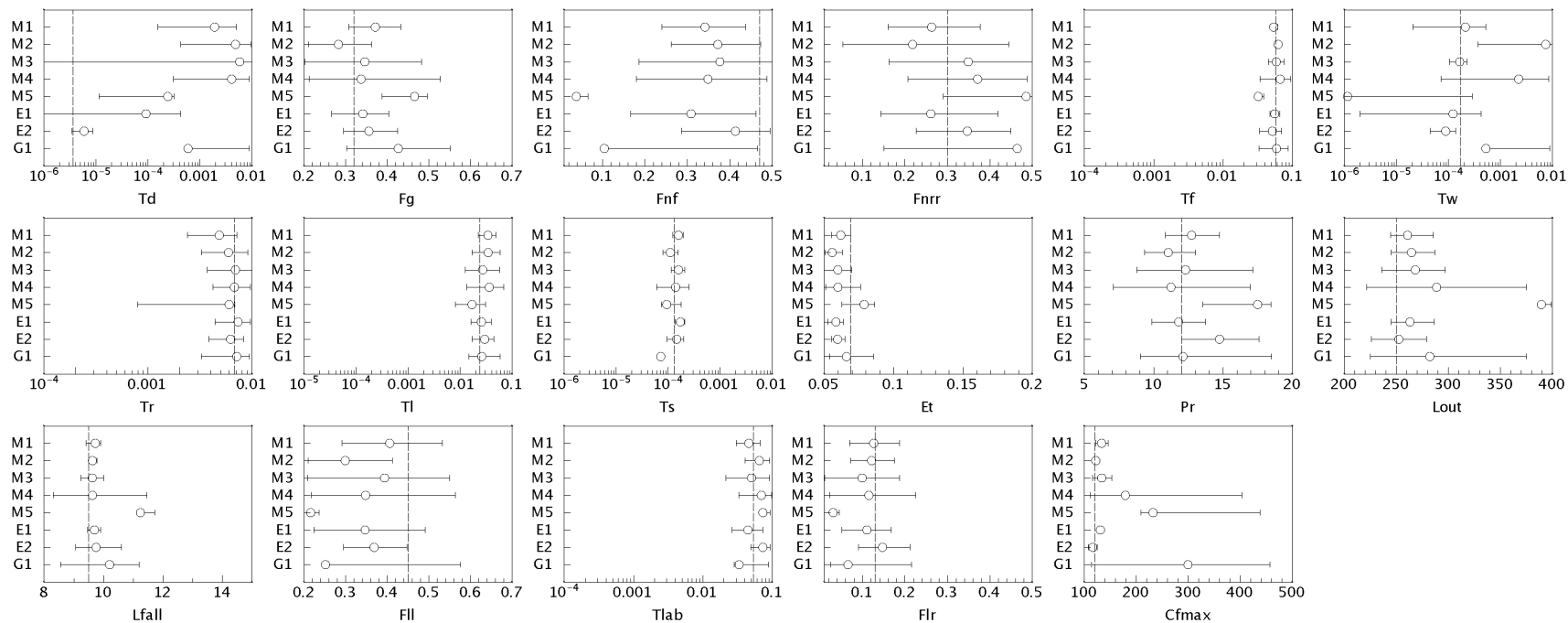
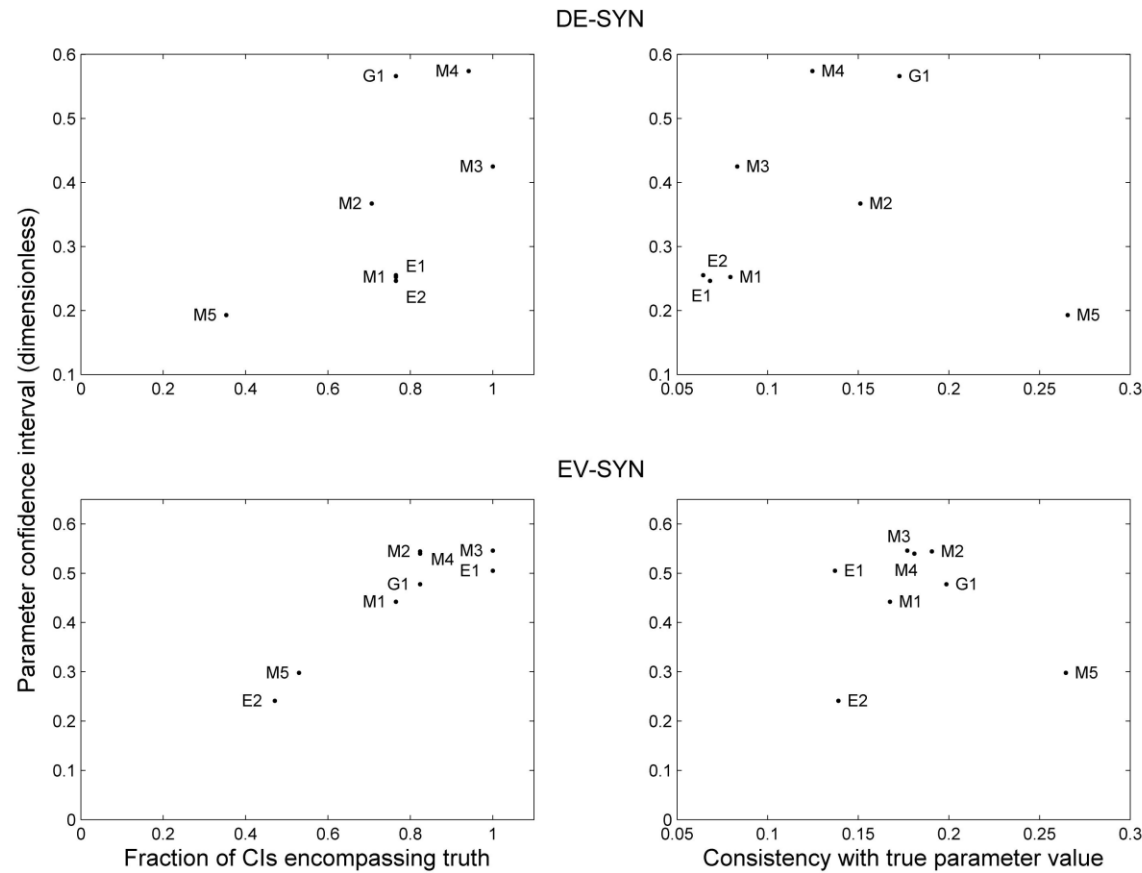


Figure 1 A schematic of the DALEC (black) and DALEC-deciduous (black and grey) models. The figures show pools (boxes) and fluxes (arrows) of C. Feed back between DALEC and the photosynthesis model is indicated by dotted line. Allocation fluxes are  $A$ , litter-fall fluxes by  $L$ , and respiration by  $R$ , 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 ( $d_5$ , see text), see right panels – low values are better. Individual algorithms are identified by alphanumeric (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.**



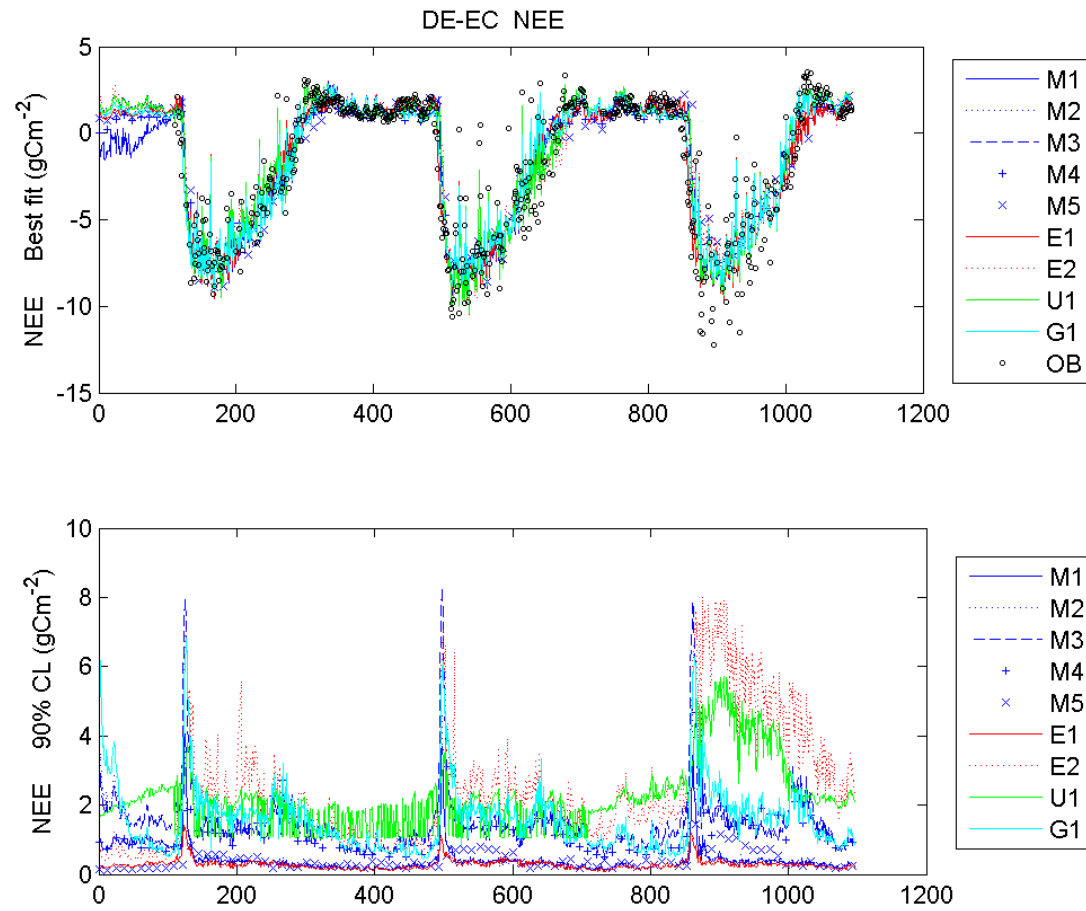
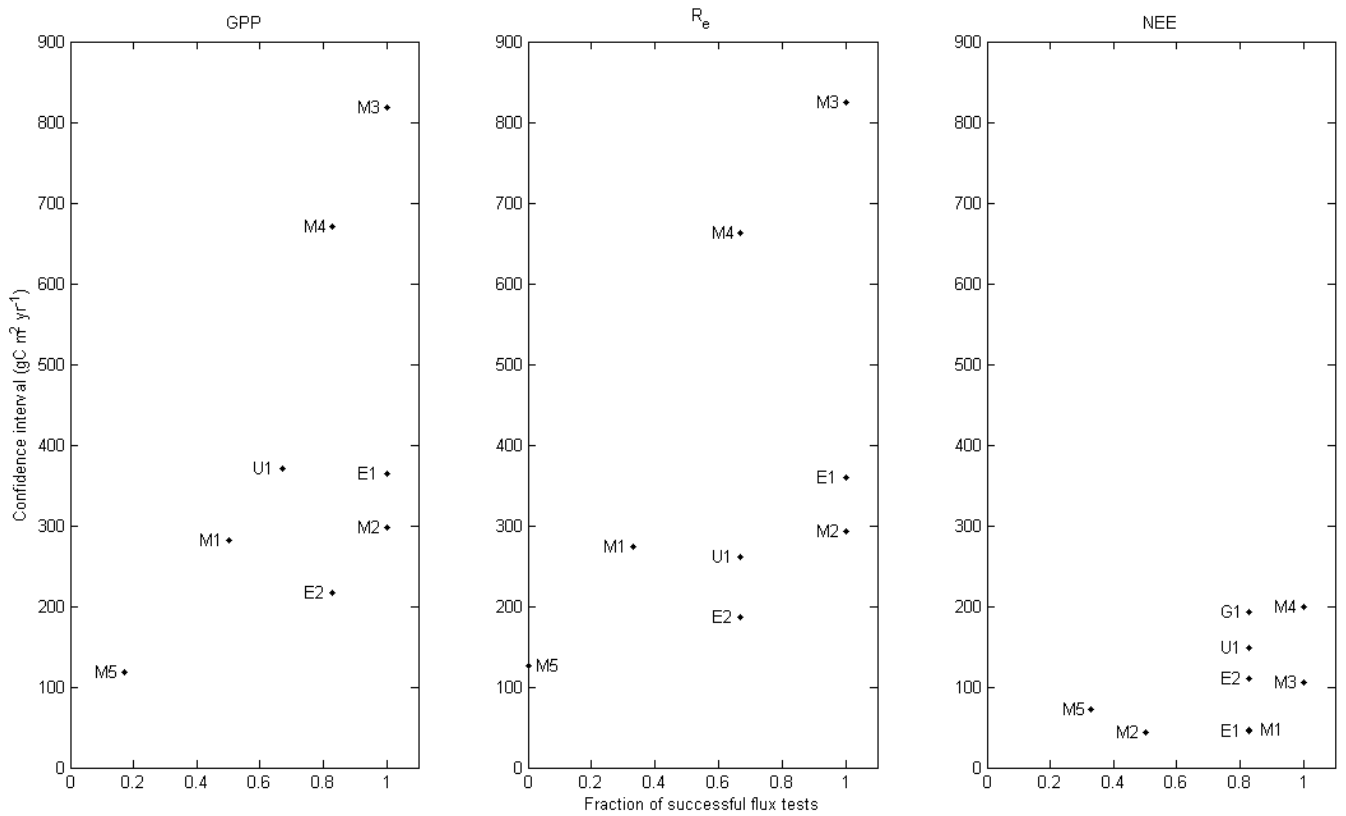
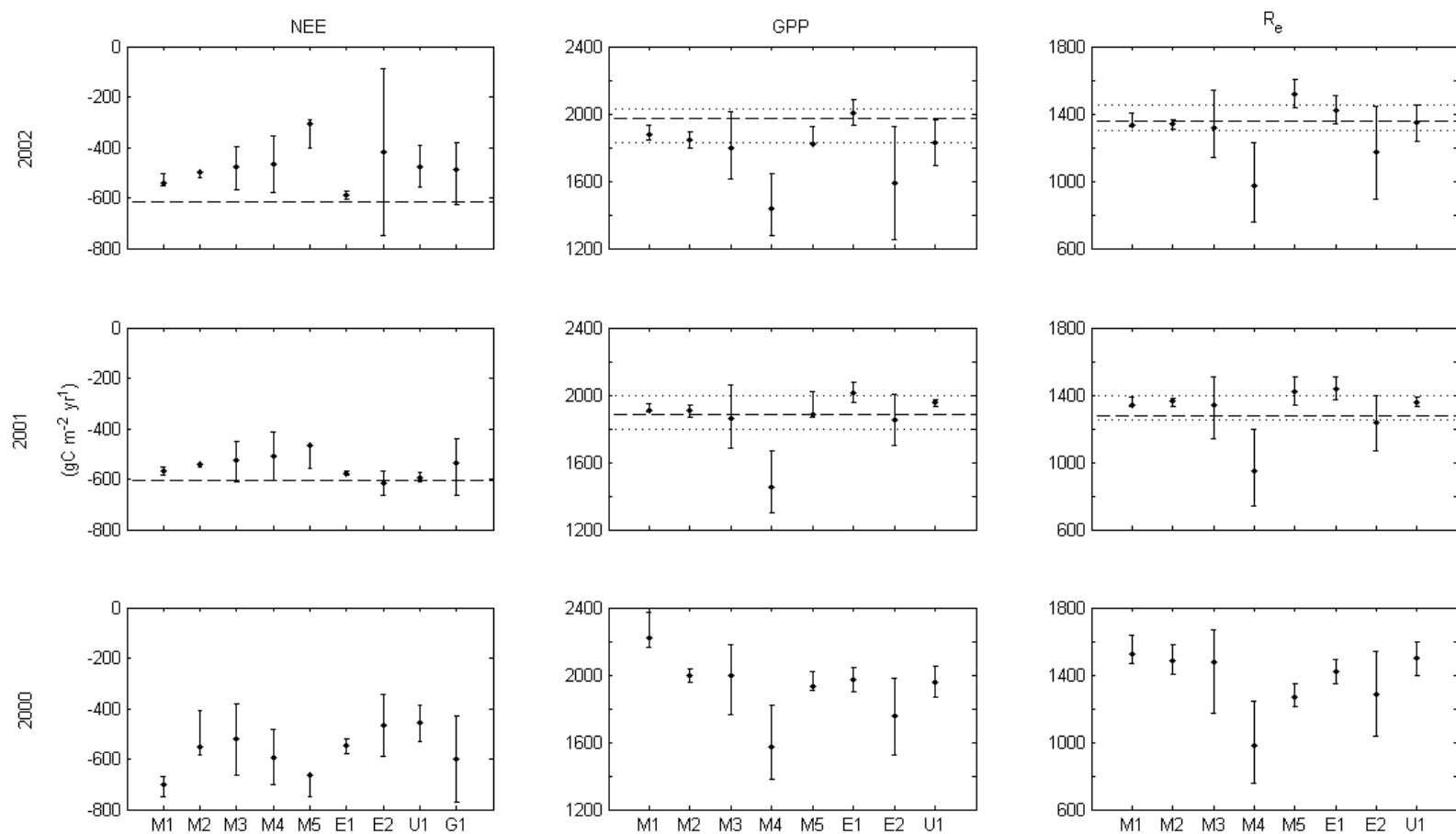


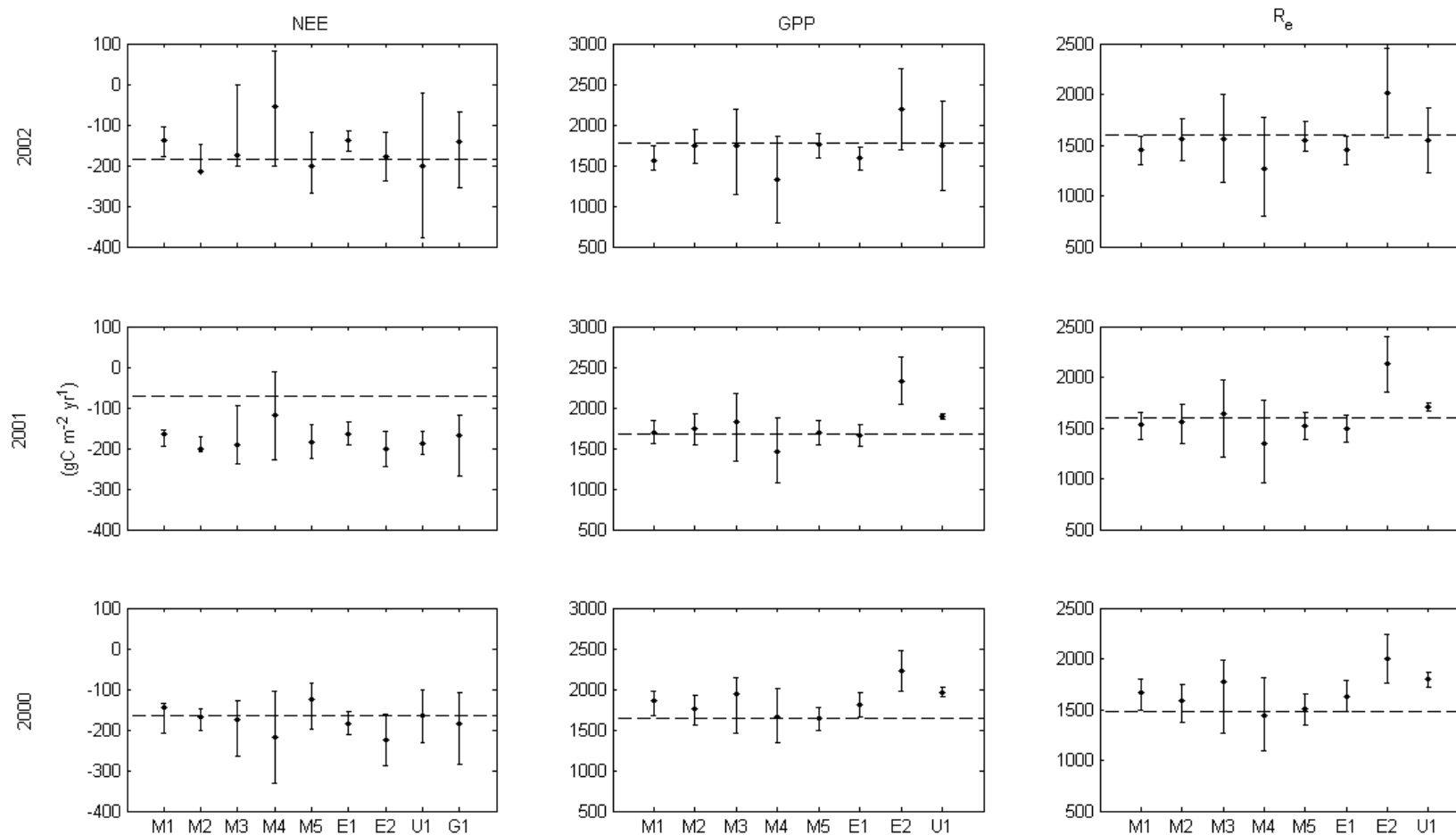
Figure 4. Estimated time series of net ecosystem exchange of CO<sub>2</sub> (NEE) over 3 years from each algorithm using observations from the DE-EC dataset over the first 2 years (top panel) and 90% confidence intervals on the estimates (lower panel). The eddy covariance data is shown as open symbols.



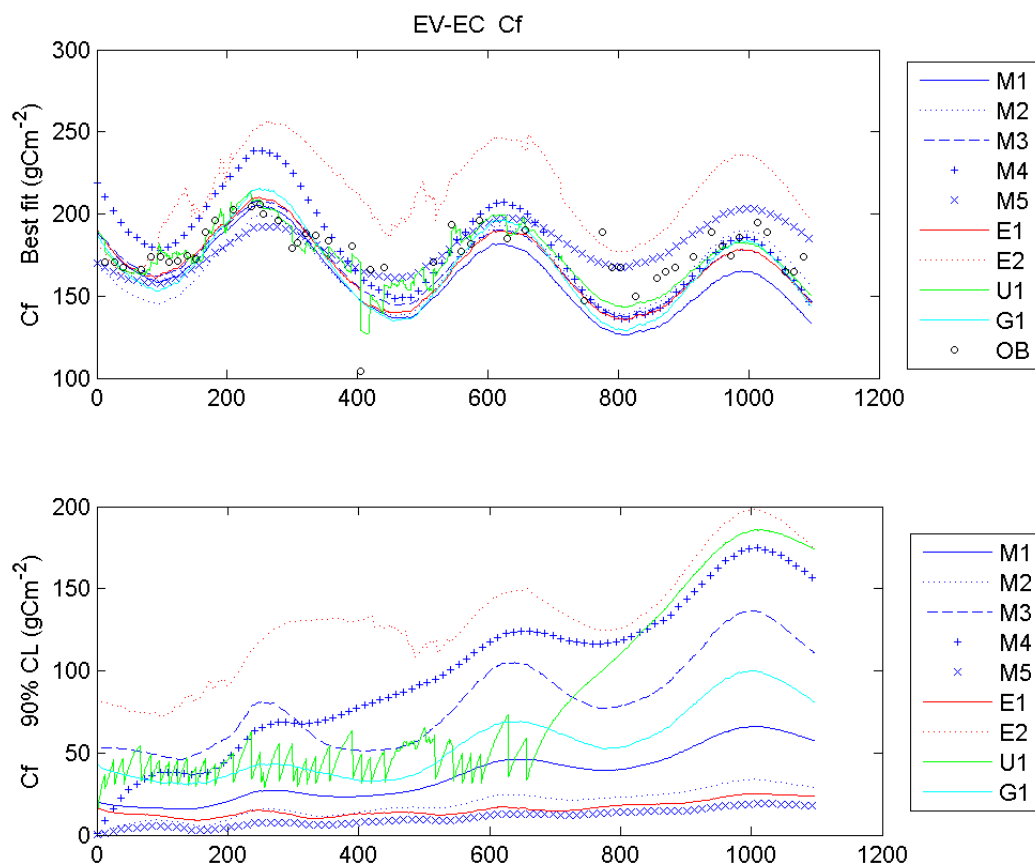
**Figure 5. A comparison between the summary success rate of annual estimates of GPP (left),  $R_e$  (centre) and NEE (right) for each algorithm plotted against the mean size of the 90% confidence interval used in the tests. The tests were for DE-SYN and EV-SYN, the synthetic datasets. Success was judged 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.**



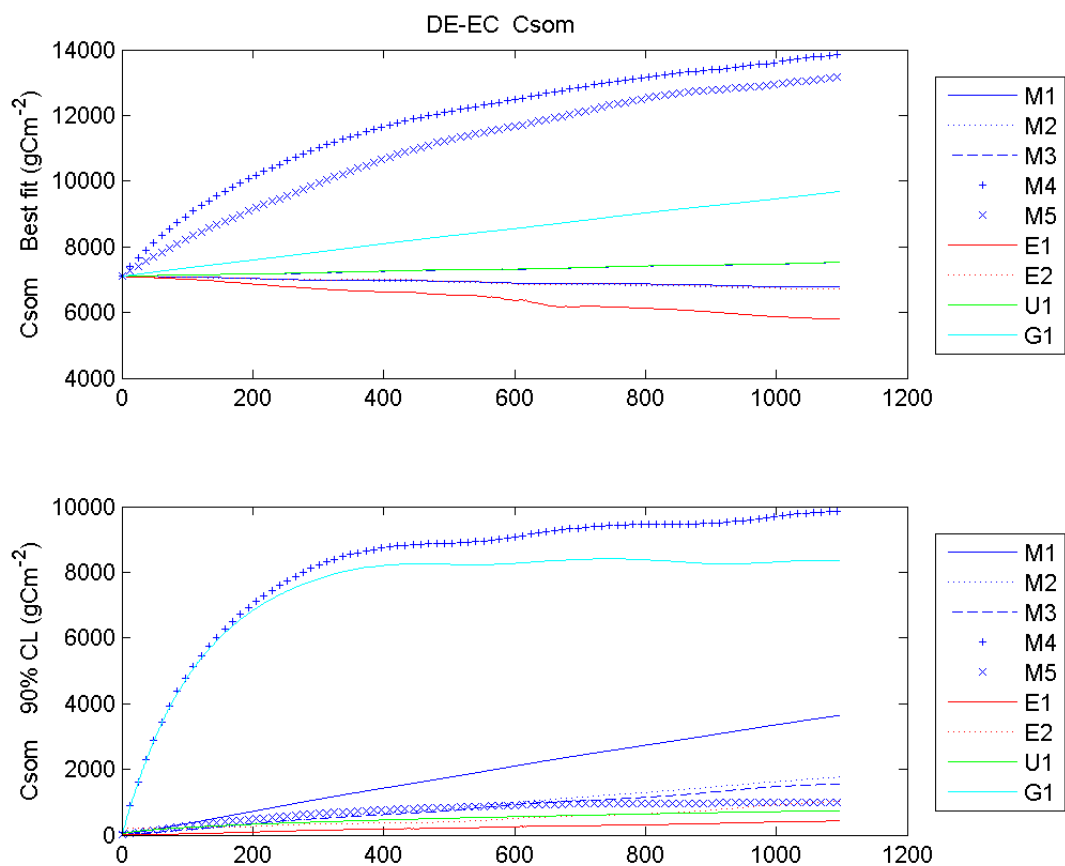
**Figure 6. Annual analyses of NEE, GPP and  $R_e$  for 2000, 2001 and prognoses for 2002 generated with the DE-EC dataset from Hesse, France. Results are shown for each algorithm for NEE and for 8 algorithms for GPP and  $R_e$ , with 90% confidence intervals indicated. The dashed lines show the best 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).**



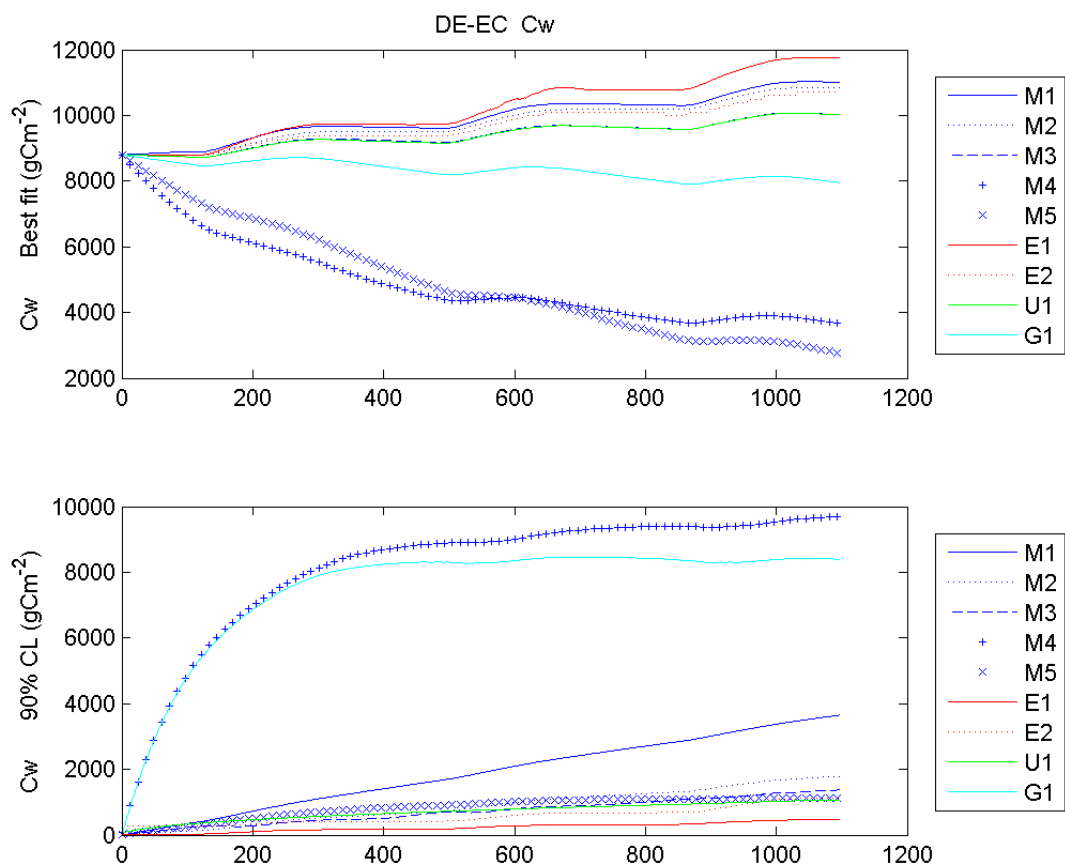
**Figure 7 Annual analyses of NEE, GPP and  $R_e$  for 2000, 2001 and prognoses for 2002 generated with the EV-EC dataset from Loobos, Netherlands. Results are shown for each algorithm for NEE and for 8 algorithms for GPP and  $R_e$ , with 90% confidence intervals indicated. The dashed lines show the 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  $C_{\text{som}}$ , 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.**