Tree Physiology 25, 839–857
© 2005 Heron Publishing—Victoria, Canada

On the validation of models of forest CO2 exchange using eddy covariance data: some perils and pitfalls

BELINDA E. MEDLYN,[1,2] ANDREW P. ROBINSON,[3] ROBERT CLEMENT[4] and ROSS E. MCMURTRIE[1]

[1] *School of Biological, Earth and Environmental Sciences, University of NSW, UNSW Sydney 2052, Australia*

[2] *Corresponding author (b.medlyn@unsw.edu.au)*

[3] *Department of Forest Resources, University of Idaho, Moscow, ID 83843, USA*

[4] *School of GeoSciences, University of Edinburgh, King's Buildings, Mayfield Road, Edinburgh, EH9 3JG, U.K.*

Received August 6, 2004; accepted November 5, 2004; published online May 2, 2005

Summary With the widespread application of eddy covariance technology, long-term records of hourly ecosystem mass and energy exchange are becoming available for forests around the world. These data sets hold great promise for testing and validation of models of forest function. However, model validation is not a straightforward task. The goals of this paper were to: (1) review some of the problems inherent in model validation; and (2) survey the tools available to modelers to improve validation procedures, with particular reference to eddy covariance data. A simple set of models applied to a data set of ecosystem CO_2 exchange is used to illustrate our points.

The major problems discussed are equifinality, insensitivity and uncertainty. Equifinality is the problem that different models, or different parameterizations of the same model, may yield similar results, making it difficult to distinguish which is correct. Insensitivity arises because the major sources of variation in eddy covariance data are the annual and diurnal cycles, which are represented by even the most basic models, and the size of the response to these cycles can mask effects of other driving variables. Uncertainty arises from three main sources: parameters, model structure and data, each of which is discussed in turn. Uncertainty is a particular issue with eddy covariance data because of the lack of replicated measurements and the potential for unquantified systematic errors such as flux loss due to advection.

We surveyed several tools that improve model validation, including sensitivity analysis, uncertainty analysis, residual analysis and model comparison. Illustrative examples are used to demonstrate the use of each tool. We show that simplistic comparisons of model outputs with eddy covariance data are problematic, but use of these tools can greatly improve our confidence in model predictions.

Keywords: eddy flux, identifiability, model comparison, model validation, residual analysis, sensitivity, uncertainty.

Introduction

Models of forest functioning have traditionally been tested against temporally sparse data, such as annual values of volume increment (e.g., McMurtrie and Landsberg 1992). Recently, however, micrometeorological techniques have developed to the point where hourly estimates of ecosystem carbon and water exchange can be calculated, and an increasing number of ecosystems are being measured in this fashion (Baldocchi et al. 2001). These data hold great promise for forest modelers because they allow models to be tested on a fine time scale. They also allow representations of exchange processes, such as gross primary productivity (GPP) and net ecosystem exchange of CO_2 (NEE) to be tested independently of those of less well understood processes such as carbon allocation.

There has been much discussion and disagreement among modelers as to what constitutes validation (Rykiel 1996). Statistical theory provides tools for validation and testing of, for example, regression models that are fitted to data, but offers little help for the validation of mechanistic models that aim to predict system behavior from underlying principles (Berk et al. 2002). As a consequence, many modelers use somewhat ad hoc approaches to validation that are open to question (Mitchell 1997). In the first part of this paper, we review some of the difficulties inherent in model validation, with particular reference to eddy covariance data. We illustrate some of the flaws in common validation procedures, using a set of simple models of ecosystem carbon exchange applied to an eddy covariance data set.

It is important to consider what is meant by "model validation." Many authors argue that the term "validation" is inappropriate, because it implies that the model is being shown to be correct, whereas models (as with hypotheses) can only be falsified, not proved (Oreskes et al. 1994, Vanclay and Skovsgaard 1997). In fact, it can be difficult to falsify models as well, as many comparisons of models show (Amthor et al. 2001, Kramer et al. 2002).

Mayer and Butler (1993) defined validation as "comparison of the model's predictions with the real world to determine whether the model is suitable for its intended purpose." This definition highlights the need to specify the purpose of the model before validation is carried out, a point stressed by many authors (Rykiel 1996, Berk et al. 2002). Some potential purposes for models validated against eddy covariance data include: identifying the driving variables of a particular ecosystem; extrapolating from the measured system to estimate regional fluxes; or modeling responses to changes in climate (van Wijk et al. 2002). Models that are judged suitable for one purpose may be unsuitable for another. Mayer and Butler's (1993) definition also highlights the need for criteria against which to judge the model's performance. How do we judge whether a model is "suitable" or not? In some cases, the criteria will be quantitative. If, for example, the aim is extrapolation for regional estimation, the criteria may be based on the uncertainty or overall error of model predictions as compared to data. In other cases, the criteria may be more qualitative, such as whether the model is able to reproduce responses to a given driving variable.

Robinson and Ek (2000) argue that validation is the responsibility of the model user, as the model developer will not necessarily have the appropriate data or insight into the intended applications. Equally, however, model users may not have sufficient insight into the working of the model to validate it adequately. We lean towards the viewpoint of Vanclay and Skovsgaard (1997), who suggest that validation for a single purpose is too limited in any case, and that modelers should carry out a full evaluation of their model, providing as much information as possible about the model's behavior and predictive ability. This approach would then enable users to decide how suitable the model is for their own purpose.

The full evaluation advocated by Vanclay and Skovsgaard (1997) includes attempting to establish the following points about the model. First, is its structure adequate, i.e., does it represent the processes involved and capture responses to driving variables? Does it provide realistic predictions for its most likely applications? Second, are the model parameters estimated correctly for the evaluation data set, and how can they best be estimated for another data set? Last what is the likely error of a model application? In the second part of this paper, we survey some of the main techniques available to modelers to carry out such an evaluation of their models. The techniques are illustrated on our set of example models, and we discuss how each technique can be used to answer these questions about model performance.

Illustrative example

In this section, we carry out a standard validation test of two illustrative models against the data, and then give a critique of the results.

We modeled NEE, which is given by the difference between ecosystem respiration (RE) and gross primary productivity (GPP). Gross primary productivity was modeled using a simple sun–shade model (Medlyn et al. 2000, 2003). Ecosystem respiration was modeled using two alternative models. The first is a Q_{10} relationship with soil temperature, which encapsulates the hypothesis that respiration is driven largely by temperature. The second respiration model is a substrate-recycling model (Dewar et al. 1999, Cannell and Thornley 2000). This model is based on the hypothesis that respiration is limited by substrate availability in the longer term. It represents two carbon pools, a "sucrose" pool and a "protein" pool. Photosynthate is assumed to enter the sucrose pool; a fraction of the sucrose pool becomes protein, which is then recycled into sucrose. The respiration rate depends on the recycling rate, which is a function of both temperature and carbon availability. The equations of the models used are given in Appendix A.

In what follows, Model 1 denotes the combination of the GPP and Q_{10} submodels, and Model 2 denotes the combination of the GPP and substrate-recycling submodels. The models were run on a half-hourly time step. Model outputs were compared to 1998 CO_2 flux data from the Griffin forest, a 17-year-old Sitka spruce (*P. sitchensis* (Bong.) Carr.) plantation in Scotland (Bernhofer et al. 2003, Clement 2004). The aims of the exercise were to: (1) validate the sun–shade GPP model; and (2) decide which is the better model of ecosystem respiration.

The models were parameterized as far as possible with values from the literature, preferring values for Sitka spruce or those from the Griffin site itself where available. A list of parameter values is given in Table 1. Sources for the parameter values are described in the section on Uncertainty in Parameterization. Parameters for the respiration submodels could not be estimated directly and, hence, were fitted to nighttime CO_2 flux data from the Griffin eddy covariance data set. Only "good" data were used, that is, data judged reliable by the system operator, and obtained under turbulent conditions. Nonlinear least squares was used for the fitting, using the package PEST (http://www.sspa.com/pest/).

The model output was then compared with the CO_2 flux data. Note that the data set used was not gap-filled, which is the practice of replacing missing or uncertain values using a fitted empirical model (Falge et al. 2001). As noted by van Wijk and Bouten (2002), the use of gap-filled data to test models would very likely increase the correspondence between data and model, since the gap-filling procedure is based on models that are structurally similar to those being tested.

It is common practice to aggregate the data in some way when making such comparisons. Aggregation reduces the data set to a manageable size and smooths out much of the noise inherent in eddy covariance measurements (Moncrieff et al. 1996, Baldocchi and Wilson 2001). A common means of aggregation is to calculate average values for each half-hour period over a fortnight or a month to give ensembles or bin-averaged data (Baldocchi and Wilson 2001). In Figure 1A, outputs from Model 1 are compared with the monthly ensembles of half-hourly CO_2 flux data from Griffin. Such comparisons are often judged by the slope and intercept of the regression, and the r^2, shown in Figure 1. In this case, the comparison appears excellent, with a high r^2, a slope close to 1 and

Table 1. Initial parameter set. For the gross primary productivity (GPP) model, parameters were obtained from the literature (see text for sources). For the respiration (RE) models, parameters were fitted using nonlinear least squares to nighttime CO_2 flux data obtained under turbulent conditions.

Parameter	Definition	Value
GPP Model		
$J_{\max 25}$	Mean maximum electron transport rate at 25 °C	79.5 μmol m^{-2} s^{-1}
E_{aJ}	Activation energy of $J_{\max}$	50.03 kJ mol^{-1}
E_{dJ}	De-activation energy of $J_{\max}$	201 kJ mol^{-1}
ΔS_J	Entropy factor for $J_{\max}$	660.13 J mol^{-1} K^{-1}
$V_{\mathrm{cmax}25}$	Mean maximum Rubisco activity at 25 °C	43 μmol m^{-2} s^{-1}
E_{aV}	Activation energy of V_{cmax}	59.45 kJ mol^{-1}
α_J	Quantum yield of electron transport	0.385 mol mol^{-1}
g_1	Ball-Berry stomatal conductance parameter	8.2
a	Leaf absorptance	89.5%
k	Canopy light extinction coefficient	0.5 m^2 m^{-2}
LAI	Leaf area index	6.5 m^2 m^{-2}
RE Model 1		
R_0	Ecosystem respiration rate at 0 °C	1.033 μmol m^{-2} s^{-1}
Q_{10}	Q_{10} of ecosystem respiration	6.44
RE Model 2		
k_c	Turnover rate of “sucrose” pool	3.97
k_p	Turnover rate of “protein” pool	0.884
a_p	Allocation to “protein” pool	0.984
Y_p	Fraction of “sucrose” turnover not lost to respiration	0.955

an intercept close to zero. Such a result is often taken to imply that the model performs well and is likely to give good predictions of CO_2 fluxes. However, this procedure can be criticized on several grounds.

From a statistical viewpoint, the procedure is flawed in many ways. First, as many authors have pointed out, the use of the r^2 statistic to judge model performance is incorrect, because it fails to account for model bias (Mayer and Butler 1993, Mitchell 1997). Several alternatives to this statistic are available, including the normalized mean average error (NMAE) and root mean squared error (RMSE), both of which measure the mean deviation of the model predictions from data, and the model efficiency (ME), given by:

$$\mathrm{ME} = 1 - \frac{\sum (y_{\mathrm{i}} - \hat{y}_i)^2}{\sum (y_{\mathrm{i}} - \bar{y})^2} \tag{1}$$

which estimates the proportion of variance of the data explained by the 1:1 line (Mayer and Butler 1993). Another approach is to test for model bias, by testing whether the regression line is significantly different from the 1:1 line. Such a test assumes, however, that data are independent and normally distributed, and neither of these assumptions holds for eddy covariance data. It can also be shown that it is always possible to obtain a significant difference between data and model as long as sufficient data are available (Robinson et al. 2005). Finally, several alternative statistical tests can be used to test for bias, and these may not be consistent in outcome (Yang et al. 2004).

In addition to these statistical issues, there are several important practical problems with this approach to validation. Equifinality (Franks et al. 1997) occurs when different parameterizations or formulations of a model give similar results. Thus, obtaining a good fit between model and data does not imply that the model is “correct,” because other parameterizations or formulations may lead to an equally good fit. Part of this problem is the identifiability of parameters, which is analogous to collinearity in regression. If a model is over-parameterized, some parameters may compensate for each other, or model output may be insensitive to them. A comparison with data does not allow these parameters to be estimated (in the case of fitted models) or tested (in the case of literature-based parameters) accurately. One criterion for a good regression model is that the parameters be identifiable (Reichert and Omlin 1997). However, for ecological models, the need to represent mechanisms in the model means that over-parameterization and poor identifiability are common. There is an analogous problem with model structure. Models may be said to represent hypotheses about system functioning. Good agreement between a model and data does not mean that we can accept the hypothesis embodied in the model, because alternative hypotheses and model structures may agree with data just as well. To test a hypothesis, it is necessary to formulate different models that do and do not include that hypothesis, and observe whether the data can be used to discriminate between the different formulations.

Insensitivity of the validation test is another problem, as for example in Figure 1A. The data used for comparison are monthly ensembles of half-hourly data over the course of

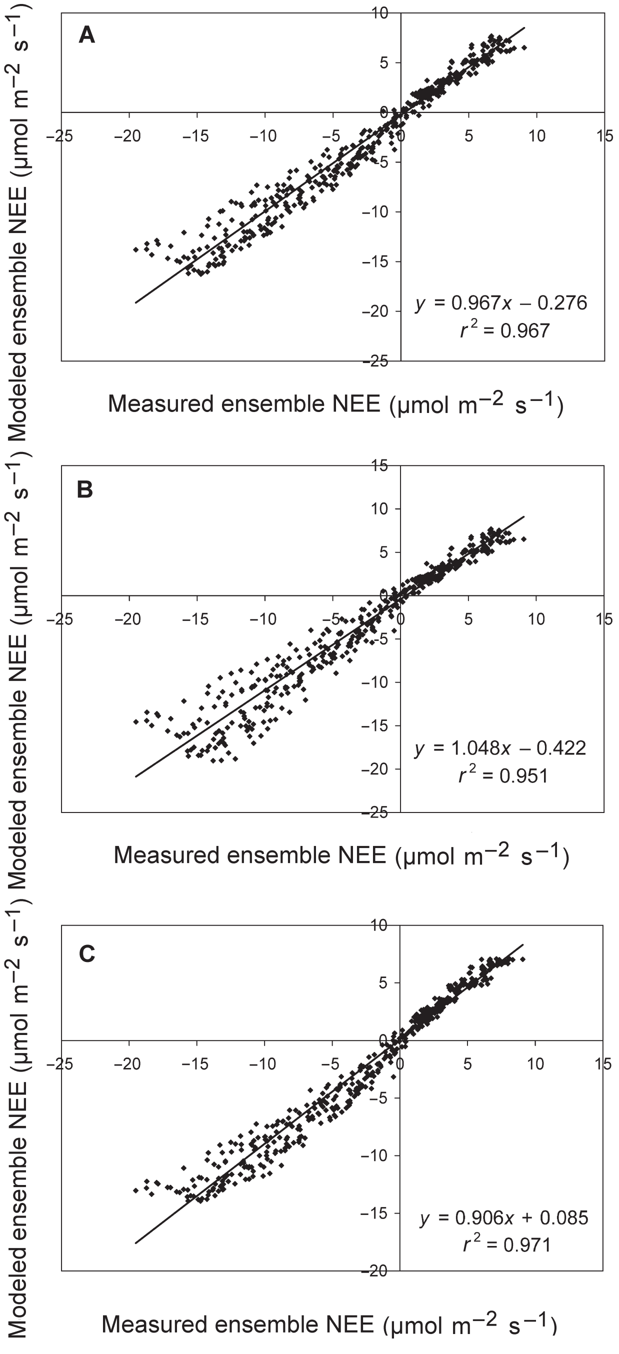

Figure 1. Comparison of modeled with measured monthly ensembles of half-hourly net ecosystem exchange of CO_2 (NEE) for the Griffin forest in 1998. (A) Model 1 (sun–shade gross primary productivity (GPP) model with Q_{10} respiration model) with parameters given in Table 1. (B) Model 1 with incorrect temperature response parameters (see text). (C) Model 2 (sun–shade GPP model with substrate-recycling respiration model) with parameters given in Table 1. Solid lines in each figure indicate best-fit line.

1 year. The main causes of variability in such data are the seasonal and diurnal variability in fluxes, and any remotely realistic model will capture this variability (Loehle 1997, Mitchell 1997). Thus, a reasonably good regression result may be expected of any model and may be misleading. Important divergences between model and data may be overlooked simply because the model is predicting CO_2 efflux at nighttime and uptake during the day, which may be enough to give a statistically significant regression with high r^2.

Uncertainty is a third important practical problem. The validation test in Figure 1A does not take uncertainty into account, and hence gives little idea of the magnitude of the error associated with model predictions. There are three main sources of uncertainty in modeling exercises (Chatfield 1995): the parameter values; the model structure; and the data used to build or validate the model. In a model application to eddy covariance data, all three of these are likely to be important. Uncertainty in data is a particular issue, since estimation of errors in the data is difficult owing to the lack of spatially replicated measurements, and because the methodology used to derive flux data from measurements is still developing, especially for non-ideal sites (Finnigan et al. 2003).

Some of these issues are illustrated in Figures 1B and 1C. In Figure 1B, an error was made in entering parameters. The model parameters E_{aJ}, E_{dJ} and E_{aV} were accidentally entered in kJ mol^{-1} rather than J mol^{-1} as required by the model, and thus were 1,000 times too small. At the same time, values of J_{max25} and V_{cmax25} were slightly lower, having not been corrected to 25 °C from the measurement temperature of 22 °C, and α_J was set to 0.3 rather than 0.385, a value used by A. Ibrom et al., Risø National Laboratory, Denmark (unpublished results) for the Griffin site. Despite the gross error in the temperature parameters, the comparison of the model output with data still looks extremely good. This example illustrates the problem of compensating errors among parameter values, and also shows the insensitivity to gross errors in parameterization of regression of modeled against predicted values. This error was made by the senior author of this paper (Medlyn 2004) and was only found through residual analysis as described below.

In Figure 1C, we show the results of the comparison of Model 2 with the CO_2 flux data. Again, the comparison appears excellent. The slope of the regression between model and data is lower than that for Model 1, but the r^2 is higher. The RMSE is marginally higher for Model 1 (1.196 cf. 1.189) and the model efficiency is the same for both models (0.966). It is not immediately obvious which of Models 1 and 2 gives the better fit to the data. Either model could be accepted as "valid." However, while the two models appear to give similar results for current environmental conditions at Griffin, the model predictions diverge greatly if temperatures (T) are assumed to rise by 2 °C, as shown in Figure 2. Under Model 1, respiration rates increase considerably with increasing T. The annual predicted NEE of the forest is reduced from 660 g C m^{-2}, under ambient conditions, to 97 g C m^{-2}. On the other hand, under Model 2, respiration rates are determined by photosynthetic uptake, rather than temperature, and do not change greatly. The annual predicted NEE stays essentially the same, changing from 513 to 503 g C m^{-2}. Thus, the use of different models, each "validated" by data, leads to dramatically different conclusions about the effect of increasing temperature on the carbon stor-

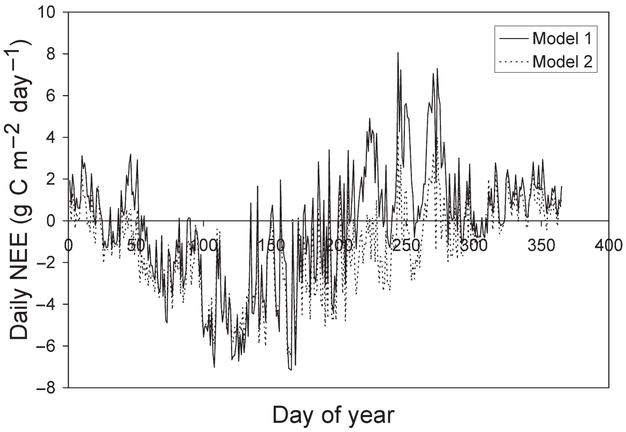

Figure 2. Modeled daily net ecosystem exchange (NEE) for the Griffin site with air and soil temperatures increased by 2 °C. Solid line = Model 1 (sun–shade gross primary productivity (GPP) model with Q_{10} respiration model); and dashed line = Model 2 (sun–shade GPP model with substrate-recycling respiration model).

age by the forest. It is important to distinguish between these models if possible. This example also illustrates why a model which performs well under current conditions may be unreliable when used to predict responses to future conditions, and highlights the need to specify the purpose of the model when carrying out the validation.

We have shown that a simple comparison between model output and flux data as a test of model validity is open to criticism. In the following sections, we describe some of the tools that can be used by modelers to evaluate model performance more critically. We begin by considering the parameterization of the model. We first estimate the error due to uncertainty in parameterization. We then evaluate the sensitivity of the model to individual parameters, and use identifiability analysis to work out which parameters are really tested by a comparison with data. Uncertainty in the data is then discussed and estimated. We fit parameters to the data, taking uncertainty into account, and show how the fitted parameters can be used to evaluate the values obtained from the literature. Finally, we consider two methods used to test model structure: residual analysis and comparison of different model structures.

Uncertainty in parameterization

We begin with a set of tools that evaluate the adequacy of parameter values. Our first concern is to evaluate the uncertainty introduced into model predictions due to uncertainty in the parameter values. A standard approach used to quantify the effects of parameter uncertainty is Monte Carlo simulation, whereby probability density functions (pdfs) are specified for each parameter value and the model is run many times over, sampling the parameters from these pdfs, to generate a pdf for the model outputs. The idea behind this approach is fairly straightforward; the key problems in implementation are: (1) the computing time required to run many simulations (although evidently this is becoming less onerous); and (2) specifying the parameter probability density functions.

The first step needed to specify the pdfs is to define the domain of reference. Here, we shall concentrate on determining the uncertainty of model predictions for Sitka spruce forests in Scotland. An alternative domain of reference might be coniferous forests in northern Europe, for example; in that case, we would need to include a much wider range of studies when determining the uncertainty in parameter values.

We used several methods described by Radtke et al. (2001) to generate parameter pdfs. One standard method, however, was ruled out by the narrowness of our domain of reference. This method involves collating parameter values from many different studies and fitting pdfs to the distribution of these parameters. In this case, there have been insufficient studies on Scottish Sitka spruce to use this approach. Instead, where possible, the uncertainty is based on the variability observed within individual experiments. The pdfs obtained are given in Table 2.

For example, to estimate mean canopy maximum electron transport rate (J_{max}), we took the mean of six values of J_{max} measured at different heights (i) in a Sitka spruce canopy by Meir et al. (2002). Each value $J_{max,i}$ was derived from an $A–C_i$ (photosynthesis–leaf intercellular CO_2 concentration) curve and, hence, had a standard error (SE_i) associated with it. The variance of each value was calculated as $n_i SE_i^2$ where n_i was the number of points in the $A–C_i$ curve. The variance of the mean of the six values was then calculated as:

$$\mathrm{Var}(J_{max}) = \frac{1}{6^2}\sum_{i=1}^{6}\mathrm{Var}(J_{max,i}) \qquad (2)$$

A similar procedure was followed for V_{cmax}.

The Ball-Berry parameter, g_1, was estimated from a regression of stomatal conductance against the term, ARH/C_a, where RH is relative humidity and C_a is ambient CO_2 concentration. The data used in the regression were obtained from branch

Table 2. Probability density distributions for model parameter values used in uncertainty assessment. See Table 1 for definitions of the parameters.

Parameter	Distribution type	Distribution parameters
J_{max25}	Normal	μ = 79.5, σ = 3.7
E_{aJ}	Uniform	min = 50, max = 70
E_{dJ}	Constant	200
ΔS_J	Uniform	min = 630, max = 670
V_{cmax25}	Normal	μ = 43, σ = 5.7
E_{aV}	Uniform	min = 55, max = 65
α_J	Uniform	min = 0.3, max = 0.4
g_1	Normal	μ = 8.2, σ = 0.25
a	Uniform	min = 0.88, max = 0.91
k	Uniform	min = 0.4, max = 0.6
LAI	Normal	μ = 5.9, σ = 0.212
R_0	Normal	μ = 1.0334, σ = 0.023
Q_{10}	Normal	μ = 0.1863, σ = 0.0027

bags established at the Griffin site (Wingate 2003). The standard error of the parameter g_1 was used as an estimate of its standard deviation.

For leaf area index (LAI), estimates based on basal area were made for forty-six 10 × 10-mm plots across the fetch of the tower (Wingate 2003). The parameter required by the model is mean LAI. We assumed that the variance of the mean of these data was the variance of the data divided by the sample size.

For the respiration parameters, which are fitted to the night-time CO_2 flux data, we assumed that the variance was given by the square of the standard error of the parameters.

For the other model parameters, little information was directly available on variance. In such cases, it is common practice to assume a uniform distribution, with bounds based on expert opinion. That approach was used here for the parameters E_{aJ}, E_{dJ}, ΔS_{J}, E_{aV}, k, a and α_{J}.

The temperature response parameters were estimated from the temperature dependence of light-saturated photosynthesis of Sitka spruce given by Neilson et al. (1972). No uncertainty values were available for these data. Instead, bounds for the uniform distribution were obtained from a review of these parameters for conifers (Medlyn et al. 2002). Uncertainty bounds were not specified for the parameter E_{dJ} because it was held constant in that review.

The light extinction coefficient, k, was assumed, in the absence of any data, to take the typical value of 0.5. The foliage is neither erectophile nor planophile, so this coefficient was assigned conservative bounds of 0.4 and 0.6. Leaf absorptance was roughly estimated by Norman and Jarvis (1974), and does not vary greatly, so a small range of values was assumed.

The uncertainty in the quantum yield parameter α_{J} was the most difficult to specify. The value of 0.385 given in Table 1 is the theoretical value for the quantum yield of electron transport given by Farquhar and Wong (1984). However, some much lower values have been obtained empirically from light-response curves. For Sitka spruce, a value of 0.3 was estimated from photosynthetic light-response curves by Ibrom et al. (unpublished results); a value of 0.22 was obtained for deciduous trees by Harley and Baldocchi (1995); and Porté and Loustau (1998) give 0.2 for maritime pine. Comparison of these values is complicated by the fact that the value obtained from a light-response curve depends on the form of the curve fitted to the data (Friend 1998), so that the parameter is not, in fact, model invariant. Here we specified bounds on the parameter as (0.3, 0.4), the lower bound based on the value from Ibrom et al. (unpublished results) and the upper bound just above the theoretical maximum.

Using the pdfs specified in Table 2, 1000 parameter samples were generated by the Latin hypercube sampling strategy, and Model 1 was run with these samples. A histogram of total annual NEE is shown in Figure 3. The minimum and maximum values of annual NEE were –713 g C m^{-2} and +132 g C m^{-2}, respectively. The 95% confidence interval was (–583, +23) g C m^{-2}, and the mean across all the simulations was –280 g C m^{-2}. These figures compare with a best estimate from the data of –590 g C m^{-2} (Clement 2004) and a best estimate from the model of –664 g C m^{-2}. Note that the best estimate from the model is not the same as the mean over the Monte Carlo simulations because some of the parameter pdfs are not centered on the best parameter estimates (α_{J} and ΔS_{J}; compare Tables 1 and 3). This analysis illustrates the range in possible model outputs given plausible parameter values, and gives some idea of the potential error in model prediction arising from uncertainty in parameter values.

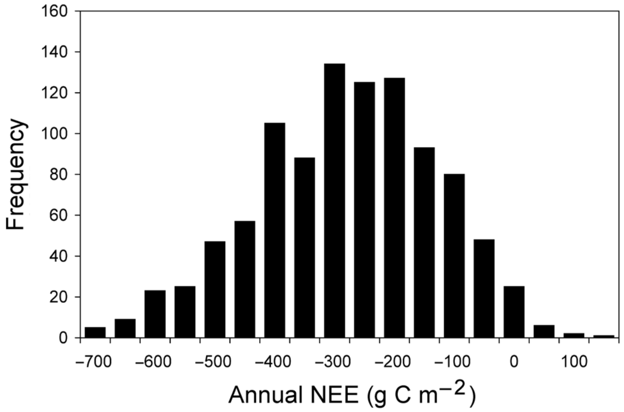

Figure 3. Uncertainty analysis of Model 1. Parameter values were sampled with the probability density functions given in Table 3 and used in 1000 simulation runs of the model. Figure shows a histogram of the frequencies of totals of annual net ecosystem exchange (NEE).

It is useful to investigate which parameters contribute most to the uncertainty, because the accuracy of model predictions will be most improved by reducing the uncertainty in those parameters. The partial correlation coefficients between the parameter values and simulated results give an approximate idea of the parameters contributing the most to uncertainty. These

Table 3. Partial correlation coefficients for uncertainty analysis. Each value is the Pearson correlation coefficient (r) for the outputs of 1000 model simulation runs against the corresponding parameter values. Parameters are presented in descending order of the absolute value of the correlation with annual NEE. See Table 1 for definitions of the parameters. Abbreviations: GPP = gross primary productivity; RE = ecosystem respiration; and NEE = net ecosystem exchange.

Parameter	Annual GPP	Annual RE	Annual NEE
α_{J}	0.528	–0.010	–0.515
ΔS_{J}	0.496	0.008	–0.478
E_{aJ}	–0.508	–0.085	0.469
LAI	0.254	–0.064	–0.263
Q_{10}	0.002	0.663	0.174
R_0	0.025	0.724	0.169
J_{max25}	0.179	0.023	–0.168
V_{cmax25}	0.135	–0.039	–0.141
k	0.104	–0.052	–0.115
a	0.074	–0.013	–0.076
E_{aV}	–0.028	0.049	0.040
g_1	0.013	0.002	–0.013

coefficients were calculated from our simulation runs and are presented in Table 3. This analysis suggests that the largest sources of uncertainty are the parameters of the temperature response of $J_{\max}$ and the quantum yield parameter.

Sensitivity and identifiability analysis

Sensitivity analysis is a basic tool used by modelers to evaluate the response of model output to changes in parameter values. In this section, we compare three approaches to sensitivity analysis. Each is applied to Model 1 and the results are compared in Table 4.

The most common sensitivity analysis is the constant fraction analysis, which involves quantifying the change in model output following a small change in the parameter values from an initial parameter set. Here, the sensitivities were calculated by increasing each parameter by 10%, re-running the model, and calculating the difference in modeled annual NEE. The model is considerably more sensitive to some parameters, such as R_0, Q_{10} and α_J, than others, such as $J_{\max 25}$ and $V_{\mathrm{cmax}25}$. This type of sensitivity analysis can be extended by evaluating the model outputs at a wider range of parameter values, rather than one small increment (e.g., Williams et al. 1998), an approach that is particularly useful for highlighting nonlinearities in the model response to different parameters.

Constant fraction analysis is useful because it highlights the parameters that have most influence on model outputs. From the point of view of making good predictions, it is useful to know the most sensitive parameters, as more effort may be devoted to determining them accurately. From the point of view of model validation, the values of the most sensitive parameters may be most stringently tested, whereas the values of insensitive parameters may not be tested at all.

There are several limitations to this analysis, however. First the perturbation of each parameter by a constant fraction is misleading, because some parameters are more uncertain than others. The model may be highly sensitive to a particular parameter, but if that parameter has a narrow range, it will have little influence on the model outcome. Second, the sensitivity analysis tests responses to parameters individually and does not indicate collinearity, or compensating effects, between parameters. Both limitations can be overcome by identifiability analysis, which can rank parameters by their influence on the model and identify sets of compensating parameters. This analysis was developed for the situation in which parameters are fitted to data using nonlinear regression and is intended to identify the subsets of parameters that can be fitted with any confidence to a given data set, but is also useful in the situation where we have taken parameters from the literature and wish to know how well-bounded these parameters are by comparison with data.

Here, we apply the formal identifiability analysis proposed by Brun et al. (2001, 2002), using the associated software package UNCSIM (available at http://www.uncsim.eawag.ch). The mathematical details of the analysis are presented briefly in Appendix B. We note that this analysis is essentially a formalization of the method used by Wang et al. (2001) to estimate parameters of a surface exchange model from eddy covariance data obtained above a wheat canopy.

The identifiability analysis involves the calculation of three different indices. Sample values of these indices for subsets of parameters from our illustrative model are given in Tables 4 and 5. The first index is a sensitivity measure, δ_j^{msqr}. This measure differs from the constant fraction analysis in Table 4 because it is weighted by the prior uncertainty range of the parameter, Δp_j, here obtained from the uncertainty analysis described above. The sensitivity of the model to highly uncertain parameters is increased. The two sensitivity measures are

Table 4. Comparison of three approaches to sensitivity analysis. In each analysis, parameters are ranked in order of importance. For the constant fraction sensitivity analysis, parameters were increased by 10% and the total change in net ecosystem exchange (NEE) over the year is given. The identifiability sensitivity measure is calculated according to Equation B2, from the uncertainty ranges given in the column Δp_j. Details of the Hornberger-Spear analysis are given in the text. Column p gives the probability that the parameter distribution for good predictive simulations is the same as that for poor predictive simulations.

Constant fraction sensitivity analysis		Identifiability sensitivity measure			Hornberger-Spear sensitivity analysis	
Parameter	ΔNEE	Parameter	Δp_j	δ_j^{msqr}	Parameter	p
Q_{10}	126.47	α_J	0.1	0.416	α_J	< 2E–16
α_J	–100.77	LAI	2	0.396	ΔS_J	< 2E–16
a	–100.77	E_{aJ}	10	0.239	E_{aJ}	3E–11
E_{dJ}	90.47	E_{aV}	10	0.216	$V_{\mathrm{cmax}25}$	1.8E–06
R_0	79.73	Q_{10}	0.01863	0.205	LAI	2E–05
LAI	–69.58	E_{dJ}	20	0.173	$J_{\max 25}$	0.04
E_{aV}	67.19	$V_{\mathrm{cmax}25}$	15	0.164	R_0	0.07
E_{aJ}	59.01	g_1	4	0.110	k	0.11
ΔS_J	–15.03	R_0	0.10334	0.109	a	0.13
$J_{\max 25}$	–15.03	$J_{\max 25}$	20	0.096	Q_{10}	0.17
$V_{\mathrm{cmax}25}$	–12.18	a	0.02	0.036	E_{aV}	0.20
g_1	–10.42	k	0.1	0.032	g_1	0.75
k	–10.13	ΔS_J	20	0.012	–	–

compared in Table 4, and we see that, for example, the parameter α_J is identified as sensitive by both approaches, whereas the parameter Q_{10} has a large sensitivity but a low uncertainty, and so has a low δ_j^{msqr}. The comparison of model output with data provides a much better bound for the higher-ranked parameters than for the lower-ranked ones.

The second index, γ, indicates the degree of collinearity of a subset of parameters. Brun et al. (2001) suggest that subsets with a collinearity index above a threshold of 10 have a high degree of collinearity. Table 5 shows that, for Model 1, the collinearity indices are extremely high for the two pairs of parameters $J_{\max}$ and ΔS, and α_J and a, indicating that the structure of the model is such that these pairs of parameters compensate strongly for each other. Indeed, inspection of Equation (A1) shows that α_J and a are multiplicative; commonly they are combined into one parameter known as the effective quantum yield. There is also a fairly high degree of collinearity among some subsets of three parameters.

The third index, ρ, indicates the identifiability of a subset of parameters and combines the sensitivity and collinearity measures. If this index is high (where "high" is a relative term in this case (Brun et al. 2001)), the subset is said to be highly identifiable because the model is sensitive to these parameters and there is a low degree of collinearity among them. According to the analysis in Table 5, the maximum number of parameters that can be identified for the sun–shade model from the Griffin data set is four. There are some subsets of four parameters that have a high ρ and yet a γ less than 10, but subsets of five parameters all have a high degree of collinearity. The subset of parameters likely to be best identified by the data is the subset ($J_{\max25}$, V_{cmax25}, α_J, LAI).

Table 5. Selection of results from identifiability analysis. The sensitivity measure was calculated for all parameters. Collinearity and identifiability indices were calculated for subsets of parameters of the GPP model only, because parameters of the respiration model have already been fitted to nighttime data. See text for further details. Definitions of the parameter are shown in Table 1.

Collinearity and identifiability	γ	ρ
Subsets of two parameters		
$J_{\max25}$, ΔS_J	945008	0.004
α_J, a	496789	0.020
E_{dJ}, E_{aV}	8.489	7.715
E_{aV}, LAI	8.091	11.951
E_{aJ}, V_{cmax25}	1.071	19.331
$J_{\max25}$, V_{cmax25}	1	12.316
Subsets of three parameters		
Most identifiable (highest ρ)		
V_{cmax25}, α_J, LAI	3.812	19.485
E_{aJ}, α_J, LAI	4.494	19.103
E_{aJ}, α_J, V_{cmax25}	2.972	19.005
Most collinear (highest γ)		
V_{cmax25}, E_{dJ}, LAI	17.741	9.089
V_{cmax25}, E_{dJ}, E_{aV}	17.067	7.577
E_{dJ}, E_{aV}, g_1	14.569	6.230
Subsets of four parameters		
Most identifiable (highest ρ)		
$J_{\max25}$, V_{cmax25}, α_J, LAI	4.20911	14.6689
E_{aJ}, V_{cmax25}, α_J, LAI	8.68471	14.0044
Subsets of five parameters		
Most identifiable (highest ρ)		
$J_{\max25}$, V_{cmax25}, α_J, LAI, E_{aJ}	17.333	9.952

In summary, the identifiability analysis highlights how sensitive the model is to each parameter, and indicates the extent to which parameters compensate for each other within the model. Only four parameters can be identified by the data, indicating that a test against data gives us a good bound on those parameters, but little further information about the other parameters.

The identifiability analysis can be carried out on any model irrespective of the system to which it is to be applied. An alternative type of sensitivity analysis, which takes into account the predictive ability of the model, was proposed by Hornberger and Spear (1981). This approach makes use of the Monte Carlo uncertainty analysis described above. The Monte Carlo simulations are divided into those that do or do not adequately predict system behavior, and the distributions of the parameter values in each category are compared. This approach highlights the parameters whose values are important in modeling a particular system behavior. As with constant fraction analysis, this approach does not take into account compensatory effects between parameters; an algorithm for dealing with these effects is given by Hornberger and Cosby (1985), but is not implemented here.

For this example, we defined a simulation with adequate predictive ability as one where the model efficiency was greater than 0.85 when simulations were compared with individual data points. Of our 1000 Monte Carlo runs, 142 fell into this category. The distributions of the parameter values of these 142 simulations were compared with the distributions for the other simulations, and a Kolmogorov-Smirnov statistic used to test whether the distributions were the same. The probability levels of these tests are given in Table 4, where the results of this sensitivity analysis are compared with the other sensitivity analyses described above. The importance ranking of parameters in this sensitivity analysis is somewhat different from that in previous analyses. The quantum yield α_J is highlighted as important by all three analyses, but the temperature parameter ΔS_J is found to be far more important in this approach. We can conclude that, although the model is not highly sensitive to ΔS_J, the value of this parameter is important when attempting to simulate this particular data set.

Error and uncertainty in data

Eddy covariance measurements are subject to both uncertainty, due to random natural fluctuations in the system under study, and error, which can arise through instrumentation or through assumptions inherent in the eddy covariance technique. Ideally, both should be taken into account in model–data comparisons, but quantification presents a number of problems.

The uncertainty due to random natural fluctuations can be estimated from short-term measurements. For example, Law et al. (1999) estimated the standard deviation of a half-hour measurement from the standard deviation of the six 5-min measurements during that half-hour. This calculation is straightforward, but unfortunately is rarely published, so the magnitude of this uncertainty is generally poorly known.

The sources of error in eddy covariance measurements are described in detail by Moncrieff et al. (1996) and Aubinet et al. (2000). The major instruments required for eddy covariance measurements are the sonic anemometer, gas analyzer and software, each having associated errors. For example, both the sonic anemometer and gas analyzer must have sufficiently high frequency responses to capture all turbulent fluctuations. Drift in the calibration of the gas analyzer and inaccuracy of the sonic anemometer will contribute errors. In general, the magnitude of the errors associated with instrumentation can be easily estimated from a working knowledge of instrument performance (Moncrieff et al. 1996). For example, Anthoni et al. (1999) estimated the overall uncertainty of daytime CO_2 flux due to instrumentation of their system as ± 12% of the flux.

The errors associated with the assumptions of the eddy covariance technique are considerably more difficult to estimate. These errors include departures from the assumptions of steady-state conditions, homogenous terrain and no advection (Moncrieff et al. 1996). Failure to account for low-frequency turbulence is also a potential source of error (Finnigan et al. 2003). For latent heat and sensible heat fluxes, the magnitude of error can be estimated by calculating the energy balance of the system, but there is no equivalent constraint that can be used to test the CO_2 flux (Moncrieff et al. 1996).

For modeling purposes, we can estimate the uncertainty in the CO_2 flux from replicates of measurements. This uncertainty will include random natural fluctuations and some error. Errors with the eddy covariance technique can be classified as either random or systematic (Moncrieff et al. 1996). Random errors will be included, but not systematic errors, which are extremely difficult to quantify.

Replicate eddy covariance measurements are rarely made, but one exception is at the Howland forest (Hollinger et al. 2004) where two towers were established 800 m apart. The towers were in similar forest areas but did not have overlapping fetches. From these measurements, Hollinger et al. (2004) calculated the standard deviation of CO_2 fluxes and showed that it varied with the magnitude of the flux. From the graphs presented therein, the standard deviation was approximately $0.9 + 0.3F_c$ µmol m^{-2} s^{-1} for positive CO_2 fluxes (F_c), and $1.5 - 0.1F_c$ µmol m^{-2} s^{-1} for negative fluxes. Standard deviation also decreased with increasing wind speed, suggesting that system error decreases with increasing turbulence.

At other sites, where only one tower has been established, "replicate" measurements can be obtained by binning measurements made under similar conditions. For example, Berbigier et al. (2001) fitted a relationship between daytime CO_2 uptake and incident photosynthetically active radiation (PAR) and calculated the error in CO_2 uptake as a function of the residuals of this relationship. Van Wijk and Bouten (2002) binned CO_2 fluxes under similar conditions and calculated the variance of each bin. The latter authors obtained an estimate of daily variance of CO_2 flux of 0.4 F_c + 20 g CO_2 m^{-2} day^{-1} (M. van Wijk, Wageningen University, The Netherlands, personal communication).

Here, we followed the approach suggested by van Wijk and Bouten (2002). Carbon dioxide measurements were divided into daytime and nighttime measurements. Measurements can be binned in several ways: we chose to bin daytime measurements by incident PAR and nighttime measurements by soil temperature, as fluxes showed strong relationships with both these variables. The mean and standard deviation of the fluxes associated with each bin were calculated and are shown in Figure 4. Daytime standard deviation was approximately 2.7 – 0.16 F_c µmol m^{-2} s^{-1}, whereas nighttime standard deviation was approximately 0.36 + 0.27 F_c µmol m^{-2} s^{-1}. These error estimates are somewhat higher than those found by van Wijk and Bouten (2002), but are comparable to those of Hollinger et al. (2004).

We also derived the following relationships between the standard deviation of the daytime flux, $F_{c,d}$, and incident PAR and Q, and between the standard deviation of nighttime flux, $F_{c,n}$, and friction velocity, U^*:

$$\mathrm{SD}(F_{c,d}) = \begin{cases} -2.5\times10^{-6}Q^2 + 5.71\times10^{-3}Q + 2.4 & Q < 1200\ \mu\mathrm{mol\ m^{-2}\ s^{-1}} \\ 5.7 & Q \geq 1200\ \mu\mathrm{mol\ m^{-2}\ s^{-1}} \end{cases} \quad (3)$$

$$\mathrm{SD}(F_{c,n}) = 2.61 - 0.678U^* \quad (4)$$

These relationships were used to estimate a standard deviation associated with each flux measurement. The standard deviations can then be used as an estimate of measurement error,

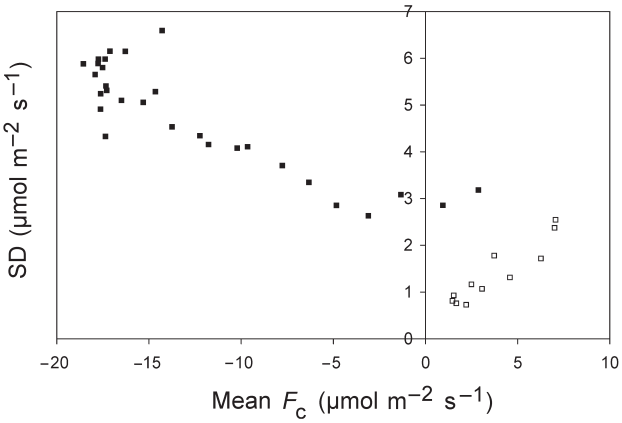

Figure 4. Error estimates for CO_2 flux data. Data were separated into daytime (■) and nighttime (□) measurements. Daytime measurements were binned by incident photosynthetically active radiation, whereas nighttime measurements were binned by friction velocity. The figure shows bin standard deviation (SD) plotted against bin mean flux.

although it should be borne in mind that the data may be subject to further errors not captured by this analysis. This estimate of measurement error can be used by modelers in several ways, including confidence testing of the model fit to data, as suggested by Loehle (1997), and the maximum likelihood methods discussed below.

Comparing the literature and best-fit parameters

Parameter values can either be obtained from the literature, as in Table 1, or fitted to data. In general, only one of these approaches is used, as they are seen as serving different purposes. If the aim is to link the model output to underlying mechanistic processes, or to use the model to predict outcomes in new situations, literature-based parameters are seen as most appropriate, and modelers are often at pains to stress that their model has not been "calibrated" to the data (e.g., Ibrom et al., unpublished results). If, on the other hand, the aim is to identify patterns in a given data set, or to predict outcomes for similar situations, then parameter values are commonly obtained by fitting (Lloyd et al. 1995, Arneth et al. 1998).

It can be useful to combine these approaches, and compare the literature-based parameter values with those obtained by fitting. It is possible, for example, that the value used in the model does not correspond exactly to the value derived from measurements, particularly where the functions involved are nonlinear. In our illustrative example, LAI varies considerably over the fetch of the measurements. We use mean LAI as a parameter to the model, and thereby introduce bias to the model because the response of GPP to LAI is nonlinear (Duursma and Robinson 2003). It can be argued that we should use an "effective" LAI rather than a mean LAI (Medlyn et al. 2003). Fitting the parameters to the data gives an estimate of the effective LAI and hence indicates the error involved in using the mean LAI instead.

In this section, therefore, we consider methods for fitting models to data. The first point to consider is that, in general, not all parameters are identifiable from data. As described above, identifiability analysis can be used to select subsets of parameters that can be fitted successfully to data. For our illustrative model, the identifiability analysis suggested that a maximum of four parameters can be fitted, and the most identifiable subset of four parameters is J_{max25}, V_{cmax25}, α_{J} and LAI.

The second point for consideration is the choice of method for fitting. Typically, nonlinear least-squares minimization is used (Wang et al. 2001). This method minimizes the sum of squares of the residuals:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (5)$$

where y_i is the observation and $\hat{y}_i$ is the model output. An alternative, more general, approach is maximum likelihood. Likelihood theory argues that some of the possible values of model parameters are more likely, or are more strongly suggested by the observed data than others. The goal is to find the values for the parameters that are the most likely given the observations: the maximum likelihood estimates (Pawitan 2001). These two methods, least squares and maximum likelihood, give identical results if the data values are normally distributed and have homogeneous variances, and the model residuals are independent. As noted above, the variances of eddy covariance data are not homogeneous, but tend to increase with the magnitude of the flux. The maximum likelihood approach will thus give different, and arguably better, parameter estimates. This approach involves minimizing the sum of squares of residuals, weighted by an estimate of the variance of the corresponding measurement error (van Wijk and Bouten 2002), i.e. minimizing:

$$\sum_{i=1}^{n}\frac{1}{\sigma_i^2}(y_i - \hat{y}_i)^2 \qquad (6)$$

This is effectively a weighted least squares approach, where the weights decrease with the variance of the flux, thus giving more weight to more accurate measurements when fitting parameters. We compared these two approaches visually by calculating contour plots of the likelihood profiles of pairs of parameters (holding all other parameters fixed). Examples of these plots are shown in Figure 5. The unweighted plots show the solution assuming constant variance, which is the same as the least-squares solution, whereas the weighted plots show the solution with variable variance. These plots show how the shape of the function to be minimized is affected by the introduction of variable weights.

We used PEST, the nonlinear parameter estimation package, to obtain best estimates of the parameters for both approaches. The estimates of standard deviation derived in Equation 7 were used for the maximum likelihood approach. The maximum likelihood parameters can also be calculated using Monte Carlo methods (van Wijk and Bouten 2002, Hollinger et al. 2004). PEST calculates 95% confidence intervals for parameter estimates for nonlinear models by treating the model as though it were locally linear, that is, linear in the region of the best parameter estimates. This requires a Taylor-series approximation of the function, evaluated at the best parameter estimates. The approximation is then treated as a linear model for estimating standard errors and confidence intervals, which require the usual regression assumptions to be true, to be reasonable. Fitted parameters and their confidence intervals are compared with the literature values in Table 6. These confidence intervals are somewhat misleading because they imply that the parameters are uncorrelated, whereas it is clearly shown in Figure 5 that the correlation between parameters affects the shape of the confidence limits. However, it is difficult to visualize four-dimensional confidence limits, so we use the linear model confidence intervals to give a rough idea of the uncertainty of each fitted parameter.

The maximum likelihood parameters are largely similar to the nonlinear least squares parameters, and both are similar to the parameter values obtained from the literature. In particular, literature estimates of J_{max25} and V_{cmax25} are close to those ob-

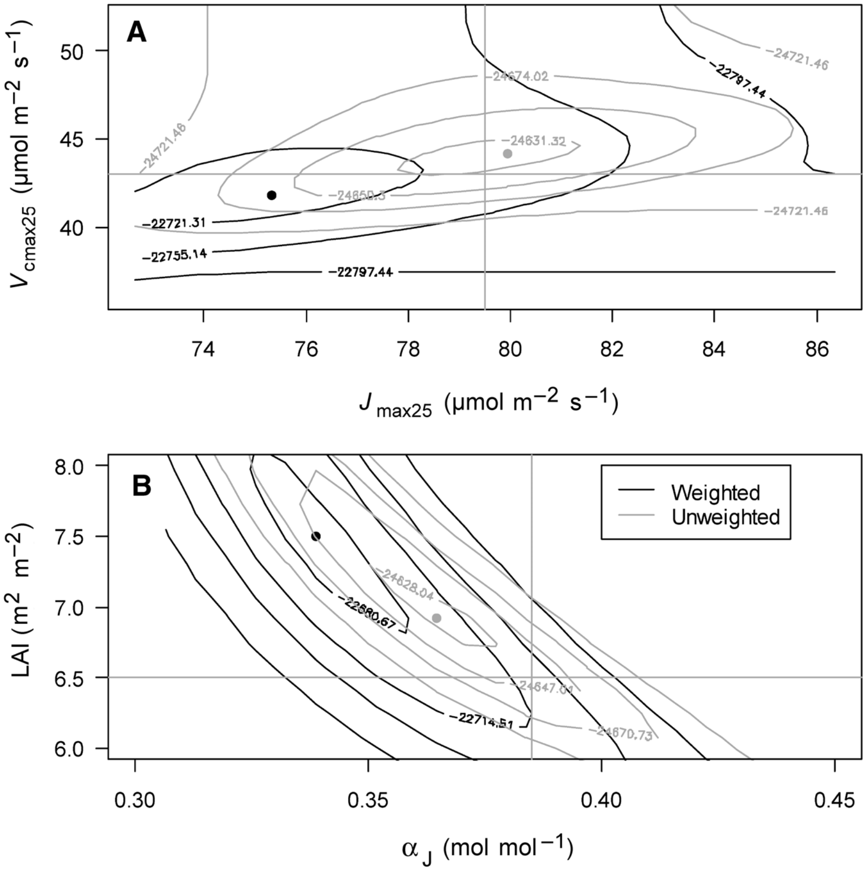

Figure 5. Contour plots for nonlinear least squares (unweighted) and maximum likelihood (weighted) fits of pairs of parameters to the data set. (A) Mean maximum electron transport rate at 25 °C (J_{max25}) and mean maximum Rubisco activity at 25 °C (V_{cmax25}) fitted, while other parameters were held constant; (B) quantum yield of electron transport (α_J) and leaf area index (LAI) fitted. The dot is the highest point of the surface, and the innermost ring is the approximate 95% confidence interval; numbers show the value of the objective function being maximized.

tained by fitting the model, giving confidence in those parameter estimates. The fitted value of α_J is somewhat less than the literature value. As noted above, our confidence in this literature value was poor, and the results of the fitting suggest that a smaller value may be more appropriate. On the other hand, the fitted value of LAI was higher than the literature value, a result that is somewhat surprising. Because the actual LAI varies greatly within the fetch, and the response of GPP to LAI is convex, it was anticipated that the effective (fitted) value of LAI would be lower than the mean value.

We suggested above that maximum likelihood estimates of parameters were better than least-squares estimates because they take into account that the variances of the eddy covariance data are not identical. However, maximum likelihood esti-

Table 6. Parameters derived from the literature compared with parameters fitted to data using nonlinear least squares (NLLS) and maximum likelihood (ML) approaches. See Table 1 for definitions of the parameters.

Parameter	Literature value	NLLS Value and 95% CI	ML Value and 95% CI
GPP Model			
J_{max25} (µmol m^{-2} s^{-1})	79.5	78.8 (43.0, 114.6)	78.6 (–6.7, 163.9)
V_{cmax25} (µmol m^{-2} s^{-1})	43 µmol m^{-2} s^{-1}	43.6 (24.3, 63.0)	44.1 (–2.2, 90.5)
α_J	0.385 mol mol^{-1}	0.363 (0.343, 0.382)	0.334 (0.298, 0.370)
LAI	6.5 m^2 m^{-2}	7.04 (3.77, 10.3)	7.55 (–0.75, 15.8)
RE Model 1			
R_0	–	1.033 (0.988, 1.079)	1.05 (1.006, 1.095)
Q_{10}	–	6.44 (6.12, 6.78)	6.36 (6.04, 6.70)
RE Model 2			
k_c	–	3.97 (–0.8, 8.7)	3.98 (–0.34, 8.31)
k_p	–	0.884 (0.08, 1.69)	0.787 (0.15, 1.42)
α_p	–	0.984 (0.96, 1.01)	0.983 (0.96, 1.00)
Y_p	–	0.955 (0.90, 1.01)	0.949 (0.90, 1.00)

mates also assume that residuals are independent, a second assumption which is not met by eddy covariance data. It can be shown that fits of this model to half-hourly flux data have autocorrelated residuals, with a lag of one half hour. Statistical methods do exist to treat this situation, but are outside the scope of this paper.

Residual analysis

We now turn to methods of testing model structure, as opposed to parameter values. As noted above, testing goodness-of-fit between model output and data can be a fairly insensitive test of model structure, especially where the main source of variability in the data is annual and diurnal variation. Residual analysis is a useful tool for investigating the performance of model structure more closely. Although there are some formal techniques for residual analysis (Cook and Weisberg 1982, Draper and Smith 1998), most of these involve assumptions that are not met by eddy covariance data, such as normality. The techniques most useful for the analysis of fitting of models to eddy covariance data mainly involve informal interpretation of graphs of model residuals. However, use of these techniques is complicated by the large quantities of data involved and correlations between driving variables, so they must be applied with care.

The residuals are the difference between model output and data at each time point. The residuals can first be used to test for model bias by plotting the residuals against the predicted values (Draper and Smith 1998). In an unbiased model, this plot will show no correlation, as in Figure 6A.

Secondly, the residuals can be plotted over time, to attempt to identify the periods when the model performs particularly well or badly. It is common to plot the model output and data together, rather than the residuals. However, as shown by Mayer and Butler (1993), this type of graph can be deceptive, because the eye is drawn to the closest distance between two lines, rather than the perpendicular distance. A plot of residuals is less open to misinterpretation. Plots of residuals over time are most informative when environmental conditions show progressive change. For example, Williams et al. (1998) modeled CO_2 flux in an Amazonian forest over a period of 7 weeks encompassing the transition from dry to wet periods. Incident radiation and temperature did not vary greatly over this period. There was a consistent overestimation in the first part of the period, allowing them to identify soil water availability as the likely cause of the overestimation.

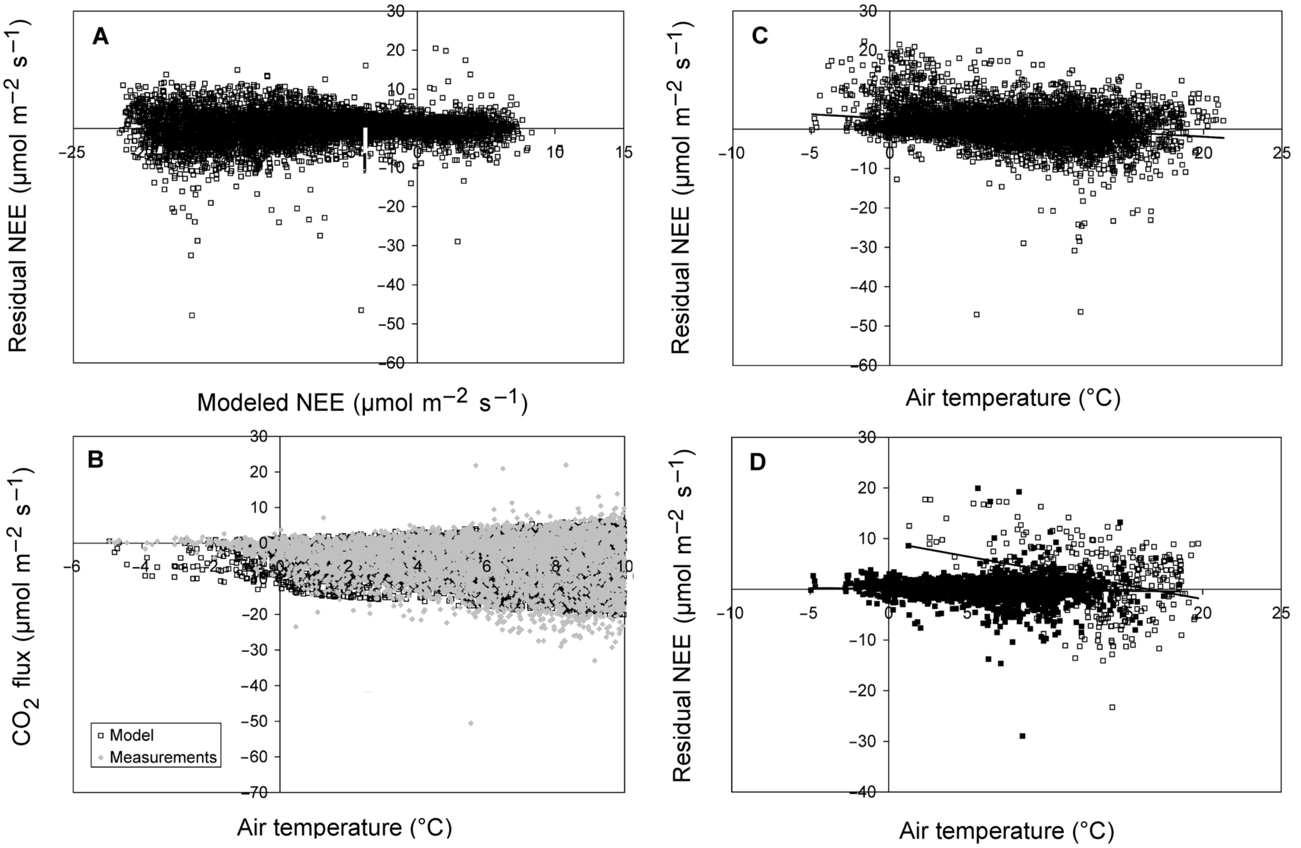

Figure 6. Residual analysis. (A) Checking Model 1 residuals for bias: plot of daytime residuals (measured–modeled flux) against modeled fluxes. (B) Modeled (□) and measured (◆) daytime CO_2 fluxes plotted against air temperature. (C) Residual net ecosystem exchange (NEE) of incorrectly parameterized Model 1 plotted against air temperature. The r^2 of the fitted line is 0.06. (D) As for Figure 5C, but data are shown separately for photosynthetically active radiation (PAR) < 100 µmol m^{-2} s^{-1} (■, r^2 = 0.003) and PAR > 1000 µmol m^{-2} s^{-1} (□, r^2 = 0.103).

In many cases, however, environmental drivers (such as incident PAR, temperature and vapor pressure deficit) co-vary, making it difficult to identify any particular driver as the main cause of large residuals. Plotting residuals, or model output and data, against individual driving variables (as recommended by Draper and Smith (1998)) can be helpful. For example, Figure 6B illustrates output of Model 1 and data plotted against air temperature. It can be seen that the model overestimates the magnitude of the flux below temperatures of about 1 °C. Above that temperature, however, the large number of data points make the graph extremely difficult to interpret. It is possible to fit a mean response of the data to any given driver and observe deviations of the model around this mean response. This approach is useful for highlighting gross discrepancies between models and data (Kramer et al. 2002) but, again, is complicated by correlations between environmental variables.

Alternative approaches that allow for correlation between environmental variables include stratifying the data and using added variable plots. Stratification can be carried out on individual data points or on averaged data. An example is given in Figures 6C and 6D. Here, the model predictions were obtained by dividing the temperature response parameters E_{aJ}, E_{dJ} and E_{aV} by 1000—the error referred to earlier in this paper. The residuals of model predictions against air temperature are shown in Figure 6C and show only a slight bias. In Figure 6D, the residuals are shown for half hours when incident PAR was < 100 μmol m^{-2} s^{-1} and > 1000 μmol m^{-2} s^{-1}. The residuals are unbiased at low incident PAR but strongly biased at high PAR, information that was used to identify the error in the parameters of the light-saturated temperature response. It can also be useful to average stratified data. For example, Ogée et al. (2003) calculate ensembles of data and model predictions for environmental conditions classified according to season, humidity and cloudiness. This approach allowed them to identify conditions under which their model performed best.

Added-variable plots are commonly discussed in texts on residual analysis (Weisberg 1985, Draper and Smith 1998), but have not, as far as we know, been applied to flux models. These plots can be used to test whether an additional variable will increase the explanatory power of the model, taking into account correlations between the additional variable and those variables already included in the model. As an example, we consider soil water content (SWC) as an added variable for Model 1. Soil water content measurements were made, in addition to all other flux measurements, at Griffin in 2001. We ran the model, using the same parameter values, for the meteorological conditions in 2001. To remove any covariance, SWC was regressed against incident PAR, air and soil temperature, and relative humidity, and the residuals of the regression calculated. The residuals of Model 1 were separated into daytime and nighttime and each was plotted against the regression residuals (Figure 7). If the added variable has no explanatory power, we would expect there to be no correlation in this plot. In order to highlight the trend of the residuals, the plot was simplified into a series of means and standard deviations. The daytime model residuals are not correlated with the SWC residuals, suggesting no influence of SWC on GPP during that year, other than that already captured by the covariance of SWC with PAR, air and soil temperature, and relative humidity. However, the nighttime residuals show a correlation, with the model underestimating measurements (residuals > 0) where the SWC residuals are large in magnitude and negative. This result suggests that modeled respiration rates could be improved by including a dependence on soil water content.

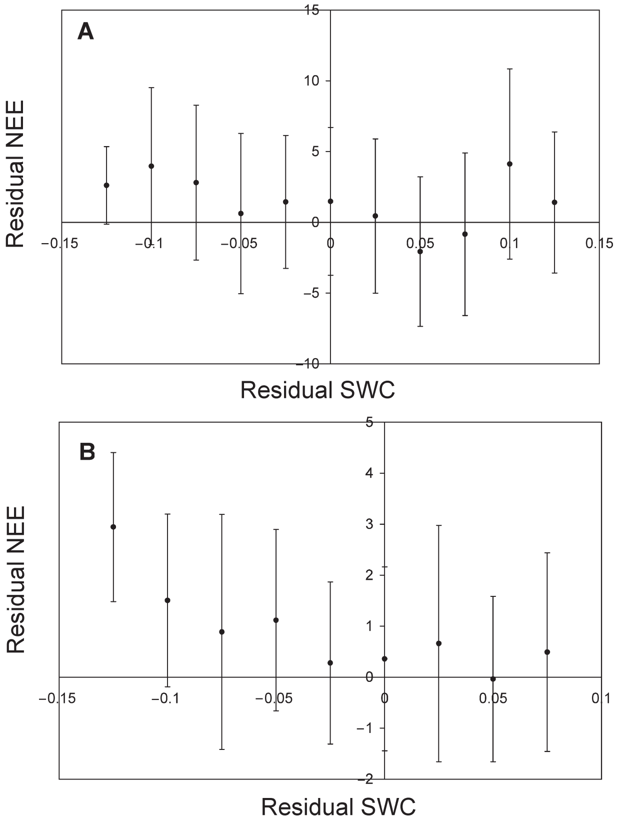

Figure 7. Added-variable plots for (A) daytime and (B) nighttime data in 2001. Residuals of net ecosystem exchange (NEE) from Model 1 are plotted against residuals of soil water content (SWC) obtained after linear regression against photosynthetically active radiation, air temperature, soil temperature and relative humidity. Data are summarized into mean and standard deviations of residual NEE to clarify presentation.

Finally, we note that artificial neural networks have also been used as a means of interpreting the complex patterns of residuals of forest flux models (Dekker et al. 2001).

Uncertainty in model structure

We now turn to model comparison as a means of evaluating model structure. Model structure is a major source of model uncertainty, but this uncertainty is extremely difficult to quantify (Chatfield 1995). A given model may provide a reasonable fit to data, but there may exist many other models that give an equally good, or better, fit. If we are interested only in extrapolating the model to make predictions under similar conditions, model uncertainty may be unimportant. If, however, the aim is to extrapolate to different conditions, or to identify key drivers of system behavior, then model uncertainty can be important. This point is illustrated in Figures 1 and 2, where models assuming different drivers for respiration give similar results under current conditions but dramatically different results when environmental conditions are changed.

Model structure can be tested through residual analysis, as described above, or by specifying alternative model structures and comparing their performance. Residual analysis is useful for finding differences between models and data, that is, for identifying conditions under which the model does not perform well. However, it is difficult to use residual analysis to demonstrate the overall acceptability of a model because such a judgement is subjective, and it does not preclude other alternatives. For example, Grant et al. (2001) set out to test the hypothesis, embodied in their model, that low ecosystem water and CO_2 exchange in boreal coniferous forests are caused by slow nitrogen cycling. They compared model output to eddy covariance data, found a reasonable fit between the two, and therefore concluded that their hypothesis was supported. But such a test is, strictly speaking, inconclusive: they have not confirmed their hypothesis, merely failed to reject it. A better test would be to encode an alternative to their hypothesis and test which model gives a better fit to data. For example, Baldocchi (1997) tested the hypothesis that stomatal conductance declines during a drought by comparing the performance of models that do and do not include an effect of drought on stomatal conductance. The model that includes the effect of drought on stomatal conductance was clearly shown to perform better. We still cannot conclude that this model is "correct"—particularly as Baldocchi (1997) showed there are still discrepancies between its output and the data, but progress has been made, because the alternative model can definitely be rejected.

Having shown the usefulness of comparing alternative models, we now consider how such a comparison can be made, using our alternative Models 1 and 2 as an example. Essentially, we want to know which model provides a better fit to data. There are some inherent difficulties involved in such a comparison. Parameterization is one issue, since we want to be sure that we are comparing model structures, not how well each model is parameterized. Further, the model with more parameters may perform best simply because it has more degrees of freedom. A second major issue is the uncertainty in the data, which, ideally, should be taken into account when estimating the goodness-of-fit of each model.

The simplest means of comparing the models is to compare the goodness-of-fit of each model to data, using a statistic such as root mean square error (RMSE) or model efficiency (ME). These statistics are shown in Table 7 for the 1998 data set, considering all half hours or nighttime half hours only. Model 2 appears to perform marginally better than Model 1 in both cases.

It is often argued that models should be parameterized and tested on different data sets—a procedure known as data splitting or cross validation (Power 1993). The argument is that the model best fitting a given data set will not necessarily give the best predictions for the future behavior of the system. Here, we tested the predictive validity of the two models by applying both to data from a second year, 1999. Statistics of the comparison with these data are given in Table 7 and show that the models perform similarly at nighttime, but Model 1 performs better on the full data set. However, this procedure may not be appropriate here because, in this case, we are most interested in which model structure can best explain the variability in the data sets. We want to test model structure, rather than parameterization, and hence should be using parameter values that have been fitted to the data.

Table 7. Statistics of comparison between Model 1 (M1) and Model 2 (M2). Statistics in the first two columns were obtained by comparing the model to the full data set, whereas those in the last two columns were obtained with the nighttime data only. Abbreviations: n = number of points in each data set; RMSE = root mean squared error; ME = mean error; SSQ_k = residual sum of squares of a model with k parameters; and AIC = Akaike Information coefficient.

Statistic	M1	M2	M1 Night	M2 Night
1998				
n	9598	–	2535	–
RMSE	3.151	3.142	1.232	1.170
ME	0.844	0.845	0.667	0.7
$\ln(SSQ_k) + 2k/n$	11.465	11.460	8.256	8.154
AIC	23082	23007	4671	4629
1999				
n	8877	–	2335	–
RMSE	3.967	4.284	1.484	1.474
ME	0.809	0.778	0.621	0.626

One problem with comparing statistics such as RMSE and ME is that Model 2 has more parameters than Model 1 (four versus two) and may provide a better fit to data because it has more degrees of freedom. There are some indices available for model comparison that take the number of parameters of the model into account. For example, Hilborn and Mangel (1997) give the index $\ln(SSQ_k + 2k/n)$ where SSQ_k is the residual sum of squares of a model with k parameters and n is the number of points. This index was evaluated for both models for the 1998 data set and was marginally better (smaller) for Model 2. The larger number of parameters is essentially irrelevant here because of the large number of data points used for fitting.

A second problem is the failure of any of these statistics to take into account the uncertainty in the data. Our uncertainty estimates can be used in another index, the Akaike Information Coefficient (AIC), which is based on the maximum likelihood calculations (Hilborn and Mangel 1997). The AIC is given by the negative log-likelihood of the model plus twice the number of its parameters. Here, we calculated the negative log-likelihood of each model as:

$$\sum_{i=1}^{n} \frac{1}{2}\ln(2\pi) + \ln(\sigma_i) + \frac{(y_i - \hat{y}_i)^2}{2\sigma_i^2} \qquad (7)$$

where the model outputs were calculated with the maximum likelihood parameters given in Table 6 and the standard deviations, σ_i, were calculated from Equation 7. The values are given in Table 7 and again suggest that Model 2 is a marginally better fit to the data, even when uncertainty estimates are taken into account.

However, the similarity of all the figures given in Table 7 still leaves room for doubt that either model is substantially better than the other, and for this reason, we report on a further test. This test was proposed by Sun (1994) and allows for the possibility that the models are indistinguishable with the data at hand. The test suggests that Model 1 should only be rejected in favor of Model 2 if the following inequality holds:

$$\sqrt{\sum_{i=1}^{n}(\hat{y}_{i,1} - \hat{y}_{i,2})^2} > \sqrt{\sum_{i=1}^{n}(\hat{y}_{i,2} - y_i)^2} + \sqrt{\sum_{i=1}^{n}\sigma_i^2} \qquad (8)$$

where $\hat{y}_{i,1}$ and $\hat{y}_{i,2}$ are the outputs of Model 1 and 2, respectively. The test thus states that Model 1 should only be rejected if the difference between the outputs of the two models is greater than the sum of the residuals of Model 2 and the observation error. For the 1998 nighttime data set, the left hand side of this inequality is 34 and the right hand side is 59 + 109 = 168, and thus Model 1 should not be rejected. This test is therefore inconclusive: essentially, it states that the differences between the models are smaller than the uncertainty in the data, making it impossible to distinguish between them.

All these tests apply to the full year's data. It may be possible to distinguish between models if we can identify periods when one model performs significantly better than the other. Here we have plotted the residuals of Model 2 against Model 1 (Figure 8) in an attempt to identify such periods. The figure demonstrates that the residuals are strongly correlated. The underlying problem is that the major drivers in the alternative models, temperature and GPP, are strongly correlated in this data set. Hence, these data cannot be used to infer that one model is better than another. To distinguish between these models, a different data set with little or no correlation between temperature and GPP would be required. What we can state is that a prediction of CO_2 fluxes under increased temperature for this system is subject to uncertainty due to model structure.

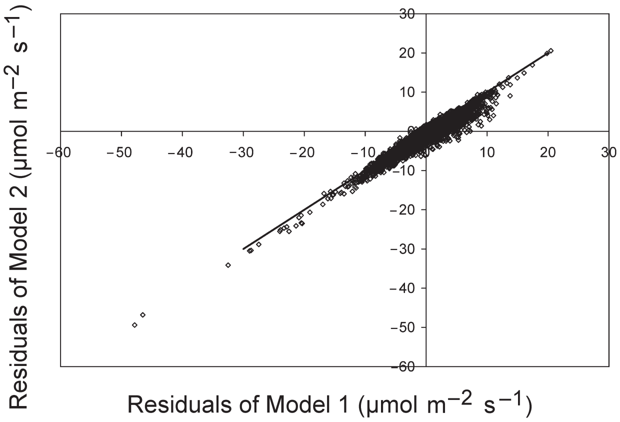

Figure 8. Residuals of Model 2 plotted against residuals of Model 1. The solid line shows the 1:1 line.

Conclusions

We showed, in the first part of the paper, that standard methods of evaluating model performance against eddy covariance data are subject to several problems. Chief among these are equifinality, or the possibility that different models may yield similar results; the dominance of the effects of incident PAR over annual and diurnal cycles; and the uncertainty inherent in parameters, data and model structure. In the second part of the paper, we reviewed a number of methods for improving our evaluation of model performance, and demonstrated them with a set of simple models of ecosystem carbon exchange, which included a model of GPP and two alternative models of ecosystem respiration.

Our initial aims were to validate the GPP model and to decide which of the two respiration models better represented the data. Following the full evaluation of the model, we have considerable confidence in the GPP model. We identified major sources of uncertainty in the model, using sensitivity and uncertainty analysis, and quantified the effect of uncertainty in the parameter values. By comparing the literature-based parameters with fitted parameters, we demonstrated that several key parameters were approximately correct. Residual analysis, including added-variable plots, demonstrated that the major environmental effects on GPP are adequately captured by the model. However, we were unable to distinguish between the two respiration models, despite careful comparison of their performance. The two models give similar outputs for current environmental conditions, but differ greatly under changed environmental conditions. We conclude that our simple models are adequate for modeling CO_2 exchange of Scottish Sitka spruce forests under current conditions, but should not be used to extrapolate into future conditions.

Acknowledgments

BM was supported by an ARC Post-doctoral Fellowship. We thank Annikki Mäkelä and Hubert Hasenauer for the invitation to present this manuscript. We also thank Marcel van Oijen, David Hollinger, Mark van Wijk, Ying-Ping Wang, Peter Levy and David Pepper for helpful discussions.

References

Amthor, J.S., J.M. Chen, J.S. Clein et al. 2001. Boreal forest CO_2 exchange and evapotranspiration predicted by nine ecosystem process models: intermodel comparisons and relationships to field measurements. J. Geophys. Res. Atmos. 106:33,623–33,648.

Anthoni, P.M., B.E. Law and M.H. Unsworth. 1999. Carbon and water vapor exchange of an open-canopied ponderosa pine ecosystem. Agric. For. Meteorol. 95:151–168.

Arneth, A., F.M. Kelliher, T.M. McSeveny and J.N. Byers. 1998. Assessment of annual carbon exchange in a water-stressed *Pinus radiata* plantation: an analysis based in eddy covariance measurements and an integrated biophysical model. Global Change Biol. 5: 531–545.

Aubinet, M., A. Grelle, A. Ibrom et al. 2000. Estimates of the annual net carbon and water exchange of forests: the EUROFLUX methodology. Adv. Ecol. Res. 30:113–175.

Baldocchi, D. 1997. Measuring and modeling carbon dioxide and water vapor exchange over a temperate broad-leaved forest during the 1995 summer drought. Plant Cell Environ. 20:1108–1122.

Baldocchi, D., E. Falge, L.H. Gu et al. 2001. FLUXNET: a new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor and energy flux densities. Bull. Am. Meteorol. Soc. 82:2415–2434.

Baldocchi, D.D. and K.B. Wilson. 2001. Modeling CO_2 and water vapor exchange of a temperate broadleaved forest across hourly to decadal time scales. Ecol. Model. 142:155–184.

Ball, J.T., I.E. Woodrow and J.A. Berry. 1987. A model predicting stomatal conductance and its contribution to the control of photosynthesis under different environmental conditions. *In* Progress in Photosynthesis Research. Ed. J. Biggins. Martinus-Nijhoff Publishers, Dordrecht, The Netherlands, pp 221–224.

Berbigier, P., J.M. Bonnefond and P. Mellmann. 2001. CO_2 and water vapour fluxes for 2 years above Euroflux forest site. Agric. For. Meteorol. 108:183–197.

Berk, R.A., P. Bickel, K. Campbell et al. 2002. Workshop on statistical approaches for the evaluation of complex computer models. Stat. Sci. 17:173–192.

Bernhofer, C., M. Aubinet, R. Clement et al. 2003. Spruce forests (Norway and Sitka spruce, including Douglas-fir): carbon and water fluxes and balances, ecological and ecophysiological determinants. *In* Fluxes of Carbon, Water and Energy of European Forests. Ed. R. Valentini. Springer-Verlag, Berlin, pp 100–123.

Brun, R., M. Kühni, H. Siegrist, W. Gujer and P. Reichert. 2002. Practical identifiability of ASM2d parameters—systematic selection and tuning of parameter subsets. Water Res. 36:4113–4127.

Brun, R., P. Reichert and H.R. Kunsch. 2001. Practical identifiability analysis of large environmental simulation models. Water Res. Res. 37:1015–1030.

Cannell, M.G.R. and J.M.H. Thornley. 2000. Modelling the components of plant respiration: some guiding principles. Ann. Bot. 85: 45–54.

Chatfield, C. 1995. Model uncertainty, data mining and statistical inference. J. Roy. Stat. Soc. Ser. A 158:419–466.

Clement, R. 2004. Mass and energy exchange of a plantation forest in Scotland using micrometeorological methods. Ph.D. Thesis, Univ. Edinburgh, U.K., 597 p.

Cook, R.D. and S. Weisberg. 1982. Residuals and influence in regression. Chapman and Hall, New York, 230 p.

Dekker, S.C., W. Bouten and M.G. Schaap. 2001. Analyzing forest transpiration model errors with artificial neural networks. J. Hydrol. 246:197–208.

Dewar, R.C., B.E. Medlyn and R.E. McMurtrie. 1999. Acclimation of the respiration photosynthesis ratio to temperature: insights from a model. Global Change Biol. 5:615–622.

Draper, N.R. and H. Smith. 1998. Applied regression analysis. John Wiley & Sons, New York, 706 p.

Duursma, R.A. and A.P. Robinson. 2003. Bias in the mean tree model as a consequence of Jensen's inequality. For. Ecol. Manage. 186: 373–380.

Falge, E., D. Baldocchi, R. Olson et al. 2001. Gap filling strategies for defensible annual sums of net ecosystem exchange. Agric. For. Meteorol. 107:43–69.

Farquhar, G.D. and S.C. Wong. 1984. An empirical model of stomatal conductance. Aust. J. Plant Physiol. 11:191–120.

Farquhar, G.D., S. von Caemmerer and J.A. Berry. 1980. A biochemical model of photosynthetic carbon dioxide assimilation in leaves of 3-carbon pathway species. Planta 149:78–90.

Finnigan, J.J., R. Clement, Y. Malhi, R. Leuning and H.A. Cleugh. 2003. A re-evaluation of long-term flux measurement techniques. Part I. Averaging and coordinate rotation. Boundary-Layer Meteorol. 107:1–48.

Franks, S.W., K.J. Beven, P.F. Quinn and I.R. Wright. 1997. On the sensitivity of soil–vegetation–atmosphere transfer (SVAT) schemes: equifinality and the problem of robust calibration. Agric. For. Meteorol. 86:63–75.

Friend, A.D. 1998. Appendix: biochemical modelling of leaf photosynthesis. *In* European Forests and Global Change: The Likely Impacts of Rising CO_2 and Temperature. Ed. P.G. Jarvis. Cambridge University Press, Cambridge, U.K., pp 335–346.

Grant, R.F., P.G. Jarvis, J.M. Massheder, S.E. Hale, J.B. Moncrieff, M. Rayment, S.L. Scott and J.A. Berry. 2001. Controls on carbon and energy exchange by a black spruce–moss ecosystem: testing the mathematical model ecosys with data from the BOREAS experiment. Global Biogeochem. Cycles 15:129–147.

Harley, P.C. and D. Baldocchi. 1995. Scaling carbon dioxide and water vapour exchange from leaf to canopy in a deciduous forest. I. Leaf model parametrization. Plant Cell Environ. 18:1146–1156.

Hilborn, R. and M. Mangel. 1997. The ecological detective: confronting models with data. Princeton University Press, Princeton, NJ, 315 p.

Hollinger, D.Y., J.D. Aber, B. Dail et al. 2004. Spatial and temporal variability in forest–atmosphere CO_2 exchange. Global Change Biol. 10:1689–1706.

Hornberger, G.M. and B.J. Cosby. 1985. Selection of parameter values in environmental models using sparse data: a case study. Appl. Math. Comp. 17:335–355.

Hornberger, G.M. and R.C. Spear. 1981. An approach to the preliminary analysis of environmental systems. J. Environ. Manage. 12: 7–18.

Kramer, K., I. Leinonen, H.H. Bartelink et al. 2002. Evaluation of six process-based forest growth models using eddy-covariance measurements of CO_2 and H_2O fluxes at six forest sites in Europe. Global Change Biol. 8:213–230.

Law, B.E., M.G. Ryan and P.M. Anthoni. 1999. Seasonal and annual respiration of a ponderosa pine ecosystem. Global Change Biol. 5:169–182.

Lloyd, J., J. Grace, A.C. Miranda, P. Meir, S.C. Wong, H.S. Miranda, I.R. Wright, J.H.C. Gash and J. McIntyre. 1995. A simple calibrated model of Amazon rainforest productivity based on leaf biochemical properties. Plant Cell Environ. 18:1129–1145.

Loehle, C. 1997. A hypothesis testing framework for evaluating ecosystem model performance. Ecol. Model. 97:153–165.

Mayer, D.G. and D.G. Butler. 1993. Statistical validation. Ecol. Model. 68:21–32.

McMurtrie, R.E. and J.J. Landsberg. 1992. Using a simulation model to evaluate the effects of water and nutrients on the growth and carbon partitioning of *Pinus radiata*. For. Ecol. Manage. 52:243–260.

Medlyn, B.E. 2004. Models of forest GPP and NPP: issues of parameterisation and validation. *In* International Conference of Modeling Forest Production. Ed. H. Hasenauer. University of Natural Resources and Applied Life Sciences, Vienna, Austria, pp 264–273.

Medlyn, B.E., R.E. McMurtrie, R.C. Dewar and M.P. Jeffreys. 2000. Soil processes dominate the long-term response of forest net primary productivity to increased temperature and atmospheric CO_2 concentration. Can. J. For. Res. 30:873–888.

Medlyn, B.E., E. Dreyer, D. Ellsworth et al. 2002. Temperature response of parameters of a biochemically-based model of photosynthesis. II. A review of experimental data. Plant Cell Environ. 25: 1167–1179.

Medlyn, B.E., D.J. Barrett, J.J. Landsberg, P. Sands and R. Clement. 2003. Conversion of canopy intercepted radiation to photosynthate: modelling approaches at regional scales. Funct. Plant Biol. 30:153–169.

Meir, P., B. Kruijt, M. Broadmeadow, E. Barbosa, O. Kull, F. Carswell, A. Nobre and P.G. Jarvis. 2002. Acclimation of photosynthetic capacity to irradiance in tree canopies in relation to leaf nitrogen concentration and leaf mass per unit area. Plant Cell Environ. 25:343–357.

Mitchell, P.L. 1997. Misuse of regression for empirical validation of models. Agric. Sys. 54:313–326.

Moncrieff, J.B., Y. Malhi and R. Leuning. 1996. The propagation of errors in long-term measurements of land–atmosphere fluxes of carbon and water. Global Change Biol. 2:231–240.

Neilson, R.E., M.M. Ludlow and P.G. Jarvis. 1972. Photosynthesis of Sitka spruce (*Picea sitchensis* (Bong.) Carr.). II. Response to temperature. J. Appl. Ecol. 9:721–745.

Norman, J.M. and P.G. Jarvis. 1974. Photosynthesis in Sitka spruce (*Picea sitchensis* (Bong.) Carr.). III. Measurements of canopy structure and interception of radiation. J. Appl. Ecol. 11:375–398.

Ogée, J., Y. Brunet, D. Loustau, P. Berbigier and S. Delzon. 2003. MuSICA, a CO_2, water and energy multilayer, multileaf pine forest model: evaluation from hourly to yearly time scales and sensitivity analysis. Global Change Biol. 9:697–717.

Oreskes, N., K. Schrader-Frechette and K. Belitz. 1994. Verification, validation and confirmation of numerical models in the earth sciences. Science 263:641–645.

Pawitan, Y. 2001. In all likelihood: statistical modelling and inference using likelihood. Oxford University Press, Oxford, U.K., 528 p.

Porté, A. and D. Loustau. 1998. Variability of the photosynthetic characteristics of mature needles within the crown of a 25-year-old *Pinus pinaster*. Tree Physiol. 18:223–232.

Power, M. 1993. The predictive validation of ecological and environmental models. Ecol. Model. 68:33–50.

Radtke, P.J., T.E. Burk and P.V. Bolstad. 2001. Estimates of the distributions of forest ecosystem model inputs for deciduous forests of eastern North America. Tree Physiol. 21:505–512.

Reichert, P. and M. Omlin. 1997. On the usefulness of over-parameterized ecological models. Ecol. Model. 95:289–299.

Robinson, A.P. and A.R. Ek. 2000. The consequences of hierarchy for modelling in forest ecosystems. Can. J. For. Res. 30:1837–1846.

Robinson, R.P., R.A. Duursma and J.D. Marshall. 2005. A regression-based equivalence test for model validation: shifting the burden of proof. Tree Physiol. 25:903–913.

Rykiel, Jr., E.J. 1996. Testing ecological models: the meaning of validation. Ecol. Model. 90:229–244.

Sun, N.Z. 1994. Inverse problems in groundwater modelling. Kluwer Academic Publishers, Dordrecht, 337 p.

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

van Wijk, M.T., W. Bouten and J.M. Verstraten. 2002. Comparison of different modelling strategies for simulating gas exchange of a Douglas-fir forest. Ecol. Model. 158:63–81.

Vanclay, J.K. and J.P. Skovsgaard. 1997. Evaluating forest growth models. Ecol. Model. 98:1–12.

Wang, Y.-P., R. Leuning, H.A. Cleugh and P.A. Coppin. 2001. Parameter estimation in surface exchange models using nonlinear inversion: how many parameters can we estimate and which measurements are most useful? Global Change Biol. 7:495–510.

Weisberg, S. 1985. Applied linear regression. Wiley, New York, 324 p.

Williams, M., Y. Malhi, A. Nobre, E.B. Rastetter, J. Grace and J.S. Pereira. 1998. Seasonal variation in net carbon exchange and evapotranspiration in a Brazilian rain forest: a modelling analysis. Plant Cell Environ. 21:953–968.

Wingate, L. 2003. The contribution of photosynthesis and respiration to the net ecosystem exchange and ecosystem ^{13}C discrimination of a Sitka spruce plantation. Ph.D. Thesis. Univ. Edinburgh, U.K., 268 p.

Yang, Y., R.A. Monserud and S. Huang. 2004. An evaluation of diagnostic tests and their roles in validating forest biometric models. Can. J. For. Res. 34:619–629.

Appendix A: Model Equations

The model of canopy GPP is a sun–shade type model, taken from Medlyn et al. (2000). Leaf photosynthesis, *P*, is assumed to be a non-rectangular hyperbolic function of incident radiation, *I*:

$$P = \frac{\alpha a I A_{\max}}{\alpha a I + A_{\max}} \tag{A1}$$

where α is the quantum yield of photosynthesis, *a* is the leaf absorptance, and $A_{\max}$ is the maximum leaf photosynthetic rate. The parameters α and $A_{\max}$ are based on the Farquhar et al. (1980) model of leaf photosynthesis as follows:

$$\alpha = \frac{\alpha_J}{4}\frac{C_i - \Gamma^*}{C_i + 2\Gamma^*} \tag{A2}$$

$$A_{\max} = \min\left(\frac{J_{\max}}{4}\frac{C_i - \Gamma^*}{C_i + 2\Gamma^*}, V_{\mathrm{cmax}}\frac{C_i - \Gamma^*}{C_i + K_m}\right) \tag{A3}$$

where α_J is the quantum yield of electron transport, C_i is the intercellular CO_2 concentration, Γ^* is the CO_2 compensation point in the absence of mitochondrial respiration, K_m is the Michaelis-Menten coefficient for carboxylation by Rubisco, $J_{\max}$ is the potential rate of electron transport, and V_{cmax} is the maximum Rubisco activity. The variables Γ^*, K_m, $J_{\max}$ and V_{cmax} have temperature dependences as given by Medlyn et al. (2002) (see their Equations 5, 6, 12, 16 and 17). The intercellular CO_2 concentration is calculated from an adaptation of the Ball-Berry stomatal conductance model (Ball et al. 1987) as follows:

$$C_i = C_a - \frac{C_a - \Gamma^*}{\mathrm{RH}}\frac{1.6}{g_1} \tag{A4}$$

where C_a is the atmospheric CO_2 concentration, RH is the relative humidity, and g_1 is the stomatal conductance parameter.

Under the assumptions that leaves are either sunlit or shaded and that the sunlit leaf fraction, incident diffuse radiation and photosynthetic capacity all decrease exponentially through the canopy with rate constant *k*, Equation (A1) can be integrated over the canopy to give the following expression for canopy GPP:

$$\mathrm{GPP} = \frac{1-\exp(-kL_c)}{k}\left(\frac{bA_{\max 0}^2 + A_{\max 0}d(A_{\max 0} + d)}{(A_{\max 0} + d)^2}\right) + \frac{1}{k}\frac{(bA_{\max 0})^2}{(A_{\max 0} + d)^3}\ln\left(\frac{(A_{\max 0} + d)\exp(-kL_c) + b}{A_{\max 0} + d + b}\right) \tag{A5}$$

where L_c is the total canopy leaf area index, $A_{\max 0}$ is the photosynthetic capacity at the top of the canopy, $b = a\alpha I_b k$ and $d = a\alpha I_d k$, and I_b and I_d are the beam and diffuse fractions of incident radiation, respectively. The photosynthetic capacity at the top of the canopy is given by:

$$A_{\max 0} = \frac{A_{\max} k L_c}{1-\exp(-kL_c)} \tag{A6}$$

where $A_{\max}$ is the mean photosynthetic capacity of the canopy and is given by Equation A3.

The first model of respiration is a simple exponential function of soil temperature, T_s:

$$R = R_0 \exp(k_R T_s) \tag{A7}$$

with parameters R_0 and k_R.

The second model of respiration is shown diagrammatically in Figure A1. The model assumes two pools of substrate, a "sucrose" pool, W_c, and a "protein" pool, W_p. The dynamics of both pools are given by:

$$W_c = W_c + \mathrm{GPP} + k_p f(T) W_p - k_c f(T) W_c \tag{A8}$$

$$W_p = W_p + a_p Y_p k_c f(T) W_c - k_p f(T) W_p \tag{A9}$$

and respiration is given by:

$$R = (1 - Y_p) k_c f(T) W_c \tag{A10}$$

The temperature function $f(T)$ is assumed to be an exponential function of air temperature with Q_{10} fixed at 2.

Net ecosystem exchange is then calculated as:

$$\mathrm{NEE} = \mathrm{RE} - \mathrm{GPP} \tag{A11}$$

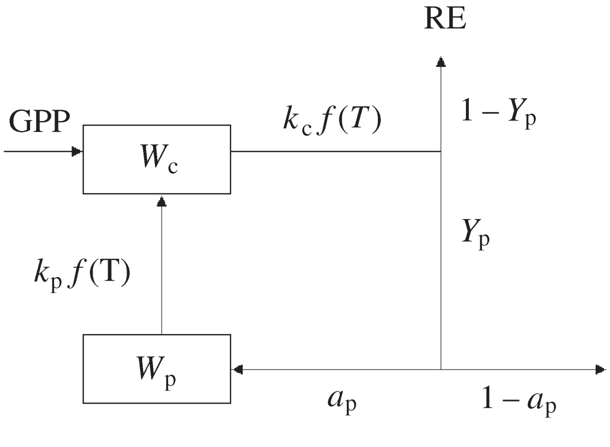

Figure A1. Diagrammatic representation of substrate-recycling model of respiration (Dewar et al. 1999, Cannell and Thornley 2000).

Appendix B: Identifiability Analysis

We give here a brief overview of the mathematical details of the identifiability analysis of Brun et al. (2001, 2002); we refer the reader to those articles for further details.

The identifiability analysis involves the calculation of three different indices. The first index is a sensitivity measure that takes into account the prior uncertainty of each parameter. A sensitivity matrix $\mathbf{S} = \{s_{ij}\}$ is calculated, with:

$$s_{ij} = \frac{\partial \eta_i}{\partial p_j} \frac{\Delta p_j}{w_i} \tag{B1}$$

where $\partial \eta_i / \partial p_j$ is the partial derivative of the ith model output with respect to the jth parameter. The weighting factors, represented by w_i, account for different scales of model outputs and the Δp_j's are prior uncertainties, that is, they give an uncertainty range for each parameter based on expert opinion. Here, the values for w_i were set to one, as all outputs are in the same units; but, for example, if both CO_2 and water vapor fluxes were used, these values might be set to the total standard deviations of each type of measurement. The sensitivity measure is then defined as:

$$\delta_j^{\mathrm{msqr}} = \sqrt{\frac{1}{n} \sum_{i,j} s_{ij}^2} \tag{B2}$$

This measure can be used to rank the sensitivity of the model to individual parameters.

The second index is a collinearity index for a subset of k parameters, defined as:

$$\gamma_k = \frac{1}{\sqrt{\tilde{\lambda}_k}} \tag{B3}$$

where $\tilde{\lambda}_k$ is the smallest eigenvalue of the matrix $\tilde{S}_K^T \tilde{S}_K$. $\tilde{S}_K$ is the $n \times k$ submatrix of the k columns of $\tilde{S}$ corresponding to the parameters in the subset, and $\tilde{S}$ is the matrix $\mathbf{S}$ with each element normalized by $\|s_\mathrm{j}\|$, the norm of the jth column of $\mathbf{S}$. The collinearity index measures the degree of near-linear dependence of the k columns of $\tilde{S}_K$. It is 1 if the columns are orthogonal and becomes infinity if the columns are linearly dependent. Near-linearly dependent columns imply a high degree of collinearity among parameters, and thus the potential for compensatory effects. A threshold of approximately $\gamma_\mathrm{k} = 10$ is suggested by Brun et al. (2001) to identify subsets of parameters with a high degree of collinearity.

Finally, Brun et al. (2001, 2002) define an identifiability index:

$$\rho_k = \left(\prod_{j=1}^{k} \lambda_j \right)^{\frac{1}{2k}} \tag{B4}$$

where λ_j are the eigenvalues of $\tilde{S}_K^T \tilde{S}_K$. This index combines the sensitivity and collinearity measures defined above: a subset of k parameters with high ρ_k is said to be highly identifiable because the model is sensitive to these parameters and there is a low degree of collinearity among them.