Van Oijen, M.; **Cameron, D.R.**; Butterbach-Bahl, K.; Farahbakhshazad, N.; Jansson, P.-E.; Kiese, R.; Rahn, K.-H.; Werner, C.; Yeluripati, J.B.. 2011 A Bayesian framework for model calibration, comparison and analysis: application to four models for the biogeochemistry of a Norway spruce forest. *Agricultural and Forest Meteorology*, 151 (12). 1609-1621. 10.1016/j.agrformet.2011.06.017

A Bayesian framework for model calibration, comparison and analysis: application to four models for the biogeochemistry of a Norway spruce forest

M. van Oijen[1,*], D.R. Cameron[1], K. Butterbach-Bahl[2], N. Farahbakhshazad[3,4], P.-E. Jansson[3], R. Kiese[2], K.-H. Rahn[2], C. Werner[2,5], J.B. Yeluripati[6]

[1] Centre for Ecology and Hydrology, CEH-Edinburgh, Bush Estate, Penicuik EH26 0QB, United Kingdom

[2] Karlsruhe Institute of Technology, Institute of Meteorology and Climate Research, Atmospheric Environmental research (IMK-IFU), Kreuzeckbahnstr. 19, 82467 Garmisch-Partenkirchen, Germany

[3] Department of Land and Water Resources Engineering, Royal Institute of Technology, 100 44 Stockholm, Sweden

[4] Swedish Secretariat for Environmental Earth Systems Sciences (SSEESS), The Royal Swedish Academy of Sciences, Stockholm, Sweden

[5] LOEWE Biodiversity and Climate Research Centre (BiK-F), Frankfurt, Germany

[6] School of Biological Sciences, University of Aberdeen, Cruickshank Building, St Machar Drive, Aberdeen AB24 3UU, United Kingdom

* Corresponding author. Tel. 44-131-4458567, FAX: 44-131-4453943, Email: mvano@ceh.ac.uk

Abstract

Four different parameter-rich process-based models of forest biogeochemistry were analysed in a Bayesian framework consisting of three operations: (1) Model calibration, (2) Model

comparison, (3) Analysis of model-data mismatch.

Data were available for four output variables common to the models: soil water content and emissions of N_2O, NO and CO_2. All datasets consisted of time series of daily measurements. Monthly averages and quantiles of the annual frequency distributions of daily emission rates were calculated for comparison with equivalent model outputs. This use of the data at model-appropriate temporal scale, together with the choice of heavy-tailed likelihood functions that accounted for data uncertainty through random and systematic errors, helped prevent asymptotic collapse of the parameter distributions in the calibration.

Model behaviour and how it was affected by calibration was analysed by quantifying the normalised RMSE and r^2 for the different output variables, and by decomposition of the MSE into contributions from bias, phase shift and variance error. The simplest model, BASFOR, seemed to underestimate the temporal variance of nitrogenous emissions even after calibration. The model of intermediate complexity, DAYCENT, simulated the time series well but with large phase shift. COUP and MoBiLE-DNDC were able to remove most bias through calibration.

The Bayesian framework was shown to be effective in improving the parameterisation of the models, quantifying the uncertainties in parameters and outputs, and evaluating the different models. The analysis showed that there remain patterns in the data - in particular infrequent events of very high nitrogenous emission rate – that are unexplained by any of the selected forest models and that this is unlikely to be due to incorrect model parameterisation.

Keywords: Bayesian calibration; carbon cycle; likelihood of data; nitrogen cycle; NO; N_2O; prior and posterior probability distributions for parameters; uncertainty analysis; water cycle

52

53

54 # 1. Introduction

55

56 ## 1.1 *Rationale*

57 Various recent reviews have assessed the evidence for impacts of environmental change on
58 European forests (Hyvönen et al., 2007; Kahle et al., 2008; Luyssaert et al., 2010). Most
59 studies have focused on changes in growth and carbon balance, but the importance of the
60 interaction with the nitrogen cycle is increasingly recognised (de Vries et al., 2009; Sutton et
61 al., 2008; Van Oijen et al., 2008a; Van Oijen et al., 2004). Research programmes to measure
62 and model emissions of nitrogenous greenhouse gases from European forests and other
63 ecosystems have been set up (Sutton et al., 2007).

64 The measurement of nitrous oxide (N_2O) and nitric oxide (NO) emissions from forest
65 soils is hampered by the large spatial and temporal heterogeneity in the fluxes, and modelling
66 these processes is still limited by availability of data (Kesik et al., 2005). Moreover, the
67 relevant underlying mechanisms have not yet been clarified fully, and large uncertainties are
68 present in both data and models. Available data sets not only suffer from random
69 measurement error, but also from systematic errors associated with the positioning of
70 measurement chambers in the field and their functioning (Butterbach-Bahl et al., 2002; Kroon
71 et al., 2010). When modelling the systems, there is uncertainty about how to represent
72 processes, i.e. model structural uncertainty (de Bruijn et al., 2009). Furthermore, there is
73 uncertainty about environmental drivers and parameter values.

74 To improve the applicability of models to the analysis of the greenhouse gas balance
75 of forests, these uncertainties need to be quantified and reduced. Probabilistic methods of
76 model-data fusion or data-assimilation have come to the fore in recent years, and offer the

77 prospect of improved data use and uncertainty quantification (Fox et al., 2009; Wang et al.,
78 2009). Because these methods are applications of probability theory, they require all
79 uncertainties – in data, model inputs and model structure - to be expressed in the form of
80 probability distributions. Bayes‟ Theorem can then be employed to update the distributions
81 when new information becomes available.

82 In biogeochemical modelling, most Bayesian applications have focused on
83 parameterisation of individual models, with little attention for systematic errors in data and
84 model structure. Wang et al. (2009) thus concluded, in a recent review on model-data fusion
85 studies for terrestrial ecosystems, that there is a need for "developing an integrated Bayesian
86 framework to study both model and measurement errors systematically". The work presented
87 here is intended to contribute to that goal.

88

89 1.2 *Towards a Bayesian framework for dynamic modelling in forest biogeochemistry*

90 We propose a framework which requires that multiple models are used in any given study,
91 and which consists of three operations: (1) Bayesian calibration, (2) Bayesian model
92 comparison, (3) Analysis of model-data mismatch.

93 The overarching objective of this paper is to demonstrate that this three-stage
94 framework is an effective tool for the analysis of models in forest biogeochemistry. For that
95 purpose, we used four different published models and one rich data set from the Norway
96 spruce forest in Höglwald, Germany (Kreutzer et al., 2009). Most of the data were on the
97 nitrogen cycle, with long time series of measurements of emissions of N_2O and NO, but we
98 also used time series of the carbon and water cycles in the form of soil respiration and soil
99 water content.

100 Bayesian calibration, i.e. the first operation in the framework, consists of defining a
101 prior probability distribution for a model‟s parameters and updating that distribution using the

data. The method has not often been applied to parameter-rich nonlinear process-based ecosystem models (Luo et al., 2009). One reason is the high computational demand associated with the technique, which is exacerbated by the long running time of the models. A second issue is the difficulty of quantifying uncertainties about random and systematic measurement errors. We show in this paper how both types of error can be accommodated in a Markov Chain Monte Carlo algorithm for Bayesian calibration.

Bayesian model comparison, the second operation used in the framework, aims to determine the extent to which the data support the different models. This is done by providing a probability distribution over models rather than parameter values. The attempt in this paper to assess whether Bayesian model comparison as a method can be useful for model selection purposes is, as far as we are aware, new for parameter-rich process-based ecosystem models.

Detailed analysis of model-data mismatch, the third operation in our framework, is not a common step in Bayesian model studies, which tend to focus on the probabilistic aspects of model behaviour rather than the internal structure of the models (Gelman and Shalizi, 2010). Bayesian calibration and model comparison effectively treat models as black boxes that convert parameter values into outputs, so this further analysis is needed to facilitate model improvement.

In summary, this paper aims to show the strengths and weaknesses of this three-operation Bayesian framework using a case-study with four models simulating the biogeochemistry of a Central European spruce forest.

2. Materials and Methods

2.1 *Data*

All data were taken from the Norway spruce (*Picea abies* L.) site at Höglwald, Germany,

latitude 48°30'N, longitude 11°10'E, altitude 540 m (Papen and Butterbach-Bahl, 1999). Trees were planted in 1907. Soil C and N were around 90,000 and 5000 kg ha^{-1} (Kreutzer et al., 2009; Rothe, 1997). For the years 1985-1995, mean annual temperature was 7.9 °C, precipitation 888 mm, and atmospheric N-deposition as measured in the throughfall 39.4 kg N ha^{-1} (Rothe, 1997). For 1975-1990, average global radiation was 11.3 MJ m^{-2} d^{-1} and wind speed 2.8 m s^{-1} (data from JRC-Ispra, as cited by (Van Oijen et al., 2008b)).

The primary data from the site were in the form of time series of daily measurements of soil water content and soil emissions of N_2O, NO and carbon dioxide (CO_2) (see e.g.Wu et al., 2010). Measurements at the Höglwald Forest are continuous throughout the year with fluxes being available in sub-daily resolution. However, daily mean values were used here for various years between 1994 and 2003 (1994-1996 and 2001-2003). For use in the calibration, the data were aggregated to monthly averages (Fig. 1). For N_2O and NO, we also calculated intra-annual quantiles of the frequency distribution of daily emission magnitudes (10, 50 and 90%). Monthly averages and annual statistics were only calculated for months and calendar years with more than 75% coverage, no gap-filling being applied. The data transformations led to ten different time series of data being available for use in the Bayesian analysis, four with monthly averages and six with annual quantiles, and with a combined number of data points of n = 225 (Table 1).

[Fig. 1 HERE]

2.2 *Models*

Four different deterministic process-based models of forest biogeochemistry were used in this study: BASFOR, COUP, DAYCENT and MoBiLE-DNDC (Table 2).

BASFOR is the simplest model in the group. It was designed to simulate the

interactive effects of changes in N-deposition, atmospheric [CO_2] and climate on the carbon balance of forests (Van Oijen et al., 2010a; Van Oijen and Thomson, 2010). The model has been subjected to Bayesian calibration before, using data from the United Kingdom (Van Oijen et al., 2005).

COUP and MoBiLE-DNDC are the two most complex models in the group. Both models were originally designed with special focus on soil processes, but recent versions of the models simulate the whole ecosystem. MoBiLE-DNDC calculates soil microclimate and hydrology, plant growth and plant-soil interactions, biogeochemical processes of the C and N cycle in soils, microbial growth and subsequent trace gas emissions. The core functionality follows the concepts developed in the DNDC suite of models (Li et al., 2000; Li et al., 1992; Werner et al., 2007). COUP was subjected to uncertainty quantification by Klemedtsson et al. (2008) and Svensson et al. (2008). The version of the model used in this paper, referred to as CoupModel, includes an N-flux submodel taken from the PnET-N-DNDC model (Norman et al., 2008). A preliminary uncertainty assessment was also carried out for MoBiLE-DNDC (de Bruijn et al., 2009).

DAYCENT, a model of intermediate complexity, traces its origins to the grassland soil model CENTURY (Parton et al., 1993), but like the previous two models it has developed into a full model for various ecosystems (Del Grosso et al., 2001), of which a small part was subjected to Bayesian calibration before (Yeluripati et al., 2009). The model version used in this paper, referred to as DailyDAYCENT, uses a daily time step for all processes.

2.3 *Parameter screening*

In the case of the simplest model, BASFOR, no parameter screening was applied, so all its 48 model parameters and initial constants for state variables were included in the Bayesian

calibration. Initial constants were not included in the calibrations of the other three models.

The COUP model was subjected to informal screening, based on previous work involving the same model and experimental site (Norman et al., 2008). The fraction of COUP"s output variability that was caused by uncertainty regarding the selected parameters was not quantified.

Calibration parameters for DAYCENT were selected using Morris screening (Morris, 1991) on average model outputs for soil water content and emissions of N_2O, NO and CO_2. DAYCENT has over 300 parameters but only 214 were subjected to Morris screening because the majority of the about hundred parameters of the soil water dynamics module were known to contribute about equally to the overall uncertainty, precluding identification of a subset of essential parameters. Morris screening is a global parameter sensitivity analysis, i.e. it explores combinations of parameter values across parameter space rather than just in the neighbourhood of a default parameterisation. Compared to other global sensitivity analyses, the method requires relatively few model runs (proportional to the number of parameters) which permits its use even for models with long runtimes. A subset of 17 DAYCENT parameters was selected in this way. The r^2 of the relations between model output variability (from runs where all parameters were varied) and the selected parameters was 0.88-0.98 for the three emission rates and 0.20 for water content, the latter value reflecting the difficulty in selecting key parameters for soil water dynamics in this model.

In the case of MoBiLE-DNDC, uncertainty in the 67 parameters of the soil chemistry submodel was considered. These were mainly parameters used to adapt or scale physical or chemical processes observed by lab studies to real world conditions, semi-empirical ratios, and k-values (decay rates for the various litter and microbial pools of varying decomposability). A substantial number of those parameters also describe Michaelis-Menten kinetics for the microbial turnover processes. The Morris screening method (Morris, 1991)

was applied to MoBiLE-DNDC using the same selection criteria as for DAYCENT, and 26 parameters were selected which together accounted for more than 60% (NO-emission) and >90% (N_2O- and CO_2-emission, soil water content) of model behaviour.

2.4 *Bayesian calibration*

Bayesian calibration is the application of probability theory to parameter estimation (Jaynes, 2003; Sivia, 2006). The method finds increasing use in ecological modelling (Ogle, 2009; Ogle and Barber, 2008; Van Oijen et al., 2005). Uncertainty about parameters is represented as a joint probability distribution for the possible parameter values. Bayes‟ Theorem is used to determine how this distribution changes in the light of new data:

$$P(\theta|D) \propto P(\theta)\, P(D|\theta) \quad (1)$$

Where $P(\theta)$ and $P(\theta|D)$ are the prior and posterior distributions for the parameters θ, i.e. before and after conditioning on the data. The factor that modifies the prior, $P(D|\theta)$, is the likelihood function, which is the probability of the data for a given θ. A formal likelihood function, integrating to unity in data space, needs to be used to be consistent with the probability calculus, allowing Bayes‟ Theorem to be applied (Rougier, (in press)).

The prior probability distribution for the parameters of a model, $P(\theta)$, reflects a modeller‟s uncertainty about parameter values before using the data. This uncertainty is subjective, and there was no effort in this study to impose any harmonisation on the priors for the four different models (except for data-scaling factors, discussed later). All modellers assigned prior distributions that could be written as the product of independent marginal distributions for individual parameters, but different types of marginal distribution were used: beta for BASFOR and uniform for the other three models.

The likelihood function P(D|θ) is the probability of the data D (the 225 data points) given model output generated by parameter vector θ. It accounts for possible measurement error. The same likelihood function was used for all models to allow formal Bayesian model comparison. As described above, the calibration data were in the form of time series of ten different variables, six for annual quantiles and four for monthly averages of daily measurements (Table 2). Estimates for the uncertainty of these variables, both random and systematic, were elicited from the data-providers: co-authors Butterbach-Bahl, Kiese and Werner. Uncertainty about random measurement error is often represented by the use of independent Gaussian distributions for the data points. However, the squared exponential in the Gaussian tends to cause asymptotic collapse of the parameter distribution even with moderate amounts of data (Clark, 2005), and may represent an overestimate of their information content. We therefore used the more heavy-tailed function proposed by Sivia (2006):

$$P(D|\theta) = \prod_{i=1}^{225} \frac{1}{\sigma_i \sqrt{2\pi}} \frac{1 - Exp(-R_i^2/2)}{R_i^2} \quad (2)$$

Where σ_i is a measure of the uncertainty about random error of the i-th data point, and R_i is the difference between model output and i-th data point, divided by σ_i. The values of the σ_i were considered to be specific to the type and magnitude of the data points, with relative values of 0.2 for the medians (Q50), 0.3 for the tail-quantiles (Q10 and Q90), and 1.0 for the monthly averages. Besides random measurement error, the data-points were considered to be subject to possible systematic error, which could result from unrepresentative positioning of the soil measurement chambers or errors in instrument calibration. This was implemented by means of four multiplicative data-scaling factors γ_j, one each for N_2O, NO, CO_2 and water.

As in other recent studies (Raupach et al., 2005), we treated the γ_j as additional parameters to be calibrated. We considered that errors larger than a factor 2 would be very unlikely, but otherwise no assumptions about the systematic data-scaling factors were made. We therefore used an uninformative Jeffrey‟s prior (Jaynes, 2003; Sivia, 2006) for the marginal prior distribution for each of the four factors:

$$P(\gamma_j) \propto 1 / \gamma_j \quad (3)$$

The Jeffrey‟s priors for the uncertainty about multiplicative error thus are log-uniform distributions on the interval [½, 2]. Keats et al. (2009) also used multiplicative errors, for air pollution data in Bayesian calibration of an atmospheric transport model, but used log-normal distributions instead. As Keats et al. (2009) argued, multiplicative error is often the natural choice for measurements of non-negative quantities.

For each of the four models, the Bayesian calibration was carried out by means of Markov Chain Monte Carlo sampling (MCMC), using the Metropolis algorithm (Metropolis et al., 1953; Van Oijen et al., 2005), but model-specific choices were made of proposal distribution and method of testing chain convergence.

Before and after each model‟s calibration, a preliminary parameter sensitivity analysis was carried out by calculating the partial correlation coefficients (PCC) for the relationships between individual parameters and the average simulated values of N_2O, NO, CO_2 and water. In contrast to the ordinary correlation coefficient (r), the PCC calculates the association between parameter and output after correcting for the linear effects of the other parameters.

2.5 *Bayesian model comparison*

The formal Bayesian model comparison consisted of quantifying the relative probabilities of

correctness of the four models, under the assumption that at least one of them was a correct model for the data. The comparison was, like the calibration described above, based on application of Bayes‟ Theorem (Kass and Raftery, 1995):

$$P(M|D) \propto P(M)\, P(D|M) = P(M) \int P(D|M,\theta)\, P(\theta)\, d\theta \quad (4)$$

Where $P(M)$ and $P(M|D)$ are the prior and posterior distributions for the models M. In contrast to the parameter distributions, these are discrete distributions, over the four models in our comparison. As shown in the right-hand term, the factor that modifies the prior, $P(D|M)$, is the integral of the likelihood function over the space spanned by the prior parameter distribution of each model. We refer to this integral as the model‟s „integrated likelihood‟. Another common name for this quantity is „marginal likelihood‟ which expresses the fact that it is found by marginalising out the parameters θ.

We *a priori* assigned equal probabilities to the different models of being correct, so $P(M)$ is uniform and the integrated likelihoods represented the relative probability for each model of being correct given the information in the data (Kass and Raftery, 1995). We quantified the integrated likelihoods as follows. For each model, 1000 parameter vectors were drawn from its prior parameter distribution. Comparison of the model outputs for these parameter vectors with the data yielded a sample of 1000 values of the likelihood, and the sample mean was taken as the estimate for the integrated likelihood of the model. Because any sampling-based method is subject to sampling error (McCulloch and Rossi, 1992), we additionally calculated the integrated likelihoods using the method suggested by Kass and Raftery (1995), as the harmonic mean of the sample generated by the MCMC.

2.6 *Analysis of model-data mismatch*

Besides the calculation of the likelihood function, the mismatch between model outputs and measurements was also quantified using more classical means. This was done separately for each of the ten output variables (Table 2) by calculating the Normalised Root Mean Square Error (NRMSE = square root of the average squared difference between model output and data, divided by the average of the data) and the squared correlation coefficient (r^2). NRMSE and r^2 are distributed quantities, because they depend on the parameterisation, so they were calculated across the range of prior and posterior parameter distributions.

Additional analysis was carried out for just the modes of the prior and posterior parameter distributions. We ran each model with the two modal parameter vectors and calculated the Mean Squared Error (MSE) for each of the four time series of monthly averages. The MSE-values were then decomposed as suggested by Kobayashi and Salam (2000):

$$+ (\sigma_M-\sigma_D)^2 + 2(\sigma_M \sigma_D) (1-r) \quad (5)$$

Where M is a simulated time series consisting of monthly averages of N_2O, NO, CO_2 or water content, D is the matching data, σ_M and σ_D are their standard deviations, and r is the correlation between the two. The decomposition consists of three terms, which can be interpreted as measures for model-data mismatch due to bias, variance error and phase shift (Kobayashi and Salam, 2000).

3. Results

3.1 *Bayesian calibration*

All four models were calibrated using the same MCMC-algorithm, i.e. Metropolis sampling. Burn-in and convergence were determined visually, by each modelling group separately, but an additional analysis of the Markov chains was carried out to confirm that parameter distributions had properly stabilised. The analysis was based on the fact that, after a chain reaches convergence, subsequent distinct and sufficiently long sub-chains should have similar sample means and variances. We compared the first and second halves of the chains after burn-in. The results showed that convergence was adequate for BASFOR and DAYCENT, with all parameters having similar means and variances in the two halves. However, for COUP and MoBiLe-DNDC, some parameters had not stabilised to the same extent, so the posterior parameter distributions for these two models were likely less accurate.

The calibration modified the means and reduced the variances of most marginal distributions. The average variance reduction for process-parameters was small in BASFOR and DAYCENT (3%, 4%), but larger in COUP and MoBiLE (26%, 29%). The data-scaling factors γ_j showed greater variance reductions except for soil water content (Fig. 2).

[Fig. 2 HERE]

Parameter uncertainty induced output uncertainty. The degree of prior output uncertainty was assessed by determining the quantiles of the output distributions (Table 3). For all models except DAYCENT, prior Q95 was one or two orders of magnitude larger than Q5 for the eight nitrogenous emission variables (Table 3). DAYCENT was already strongly constrained by its prior parameter distribution. For all models, soil respiration was *a priori* slightly more constrained than the N-emissions, whereas soil water content, which was the only state variable in the set of ten output variables, was most narrowly delimited. Overall, prior ranges were widest for BASFOR. The calibration had only little impact on the output

distributions for soil water content and almost no effect on the soil respiration distributions (Table 3). However, the posterior outputs for the nitrogenous emission variables were much more narrowly constrained than the prior distributions, with posterior Q95/Q5 ratio‟s ranging from 2-5 (BASFOR, see also Fig. 1), 2-3 (COUP) and ~1.5 (MoBiLE-DNDC) (Table 3). DAYCENT was the exception with posterior ratio‟s that were similar to the prior.

Using the samples from both the prior and posterior parameter distributions, we calculated the partial correlations between individual parameters and outputs. Posterior PCC-values tended to be higher than prior values for BASFOR and COUP, whereas the calibration decreased PCC-values for DAYCENT and MoBiLE-DNDC. However, for all models and output variables, the PCC-based ranking of the parameters changed little, so we restrict ourselves to reporting on the posterior values.

In the case of BASFOR, only one parameter was strongly correlated ($|PCC| > 0.5$) with N_2O- and NO-emission: the soil water content at which both emissions are equal. Soil respiration was strongly correlated with the parameters that govern decomposition rate of organic matter, and also with the light-use efficiency. Soil water content was mainly correlated with specific leaf area and leaf longevity, both of which affect the active surface area for transpiration. The results for COUP were similar. N-emissions were also mainly correlated with the N_2O-NO balance, soil respiration with decomposition rates and light-use efficiency, and water content with specific leaf area. In the case of DAYCENT, no individual parameters were *a posteriori* strongly correlated with model outputs, although there had been some strong correlations with the prior parameter distribution (i.e. the NO_3-N_2O conversion efficiency for N-emissions and the leaf area ratio for soil water content). In the case of MoBiLE-DNDC, no parameters were strongly correlated with N_2O-emission, but the K_m-value for NO_2 did have a high PCC with NO-emission. Soil respiration and water content were both mainly correlated with the parameter that scales decomposition of active organic

substance as a function of soil porosity.

3.2 *Bayesian model comparison*

The log-transformed integrated likelihood values, calculated from samples from the prior, were as follows (between brackets the values from the alternative calculation using the harmonic mean of the sample from the posterior): BASFOR: -661.7 (-654.7), COUP: -663.5 (-651.2), DAYCENT: -738.5 (-761.2), MoBiLE-DNDC: -657.0 (-758.9). For comparison: a parameter vector whose model outputs would have gone exactly through the 225 data points, would have had a log-likelihood of -581.2. Both methods of calculating the integrated likelihoods showed that the data provided greater support for BASFOR and COUP than for DAYCENT. The two estimates of the integrated likelihood for MoBiLE-DNDC differed strongly, so it is less clear how plausible this model is.

3.3 *Analysis of model-data mismatch*

First, the data were compared with the ranges of model output uncertainty induced by the parameters. All time series averages of measurements were in the central intervals of the prior output distributions, between the 5% and 95% quantiles, except for soil respiration as predicted by BASFOR and MoBiLE-DNDC, and the lower quantiles of daily N_2O emission rates as predicted by COUP (Table 3).

Although the distributions of simulated time series averages were found to cover the data fairly well, inspection of the time series themselves revealed considerable differences between models and measurements (Fig. 3). For example, none of the models was able to reproduce the large peak in N_2O-emissions in early 1996 after a strong freeze-thaw event (Papen and Butterbach-Bahl, 1999), neither with the mode of the prior, not with the posterior mode (Fig. 3). This is likely to be a consequence of incomplete process representation in the

models, although a more recent version of MoBiLE-DNDC showed a possible way to account for the freeze-thaw effect (de Bruijn et al., 2009). Despite these remaining differences between model outputs and data, overall the calibration was able to remove much of the mismatch for the N-emissions and, to lesser extent, for CO_2-emission (compare the lower „posterior" panels of Fig. 3 to the upper „prior" ones, and see also the increased likelihoods depicted in Fig. 4). Note that some of the reduced bias in the posterior results was due to the impact of calibration on data-scaling factors (Fig. 2) rather than on model parameters. There was no apparent improvement in the simulation of soil water content, which reflected the fact that for this variable the least amount of information was available (Table 2). In one specific case, simulation of soil water by model DAYCENT using the mode of its posterior distribution, bias was increased relative to the prior mode (Fig. 3), but this result was not representative of the full posterior distribution for the water data-scaling factor (γ_{WATER}) of this model (Fig. 2). However, it does suggest that a useful – but for this model computationally demanding - additional step in the procedure would have been to determine the mode of the posterior distribution by targeted optimisation, rather than relying on the parameter sample generated by the MCMC.

[Fig. 3 HERE]

Whereas for the prior output distributions of the models most time-series averages were in the central Q5-Q95, the same did not apply to the posterior distributions. Most time-series averages were to be found in the upper tails (>Q95) of the posterior output distributions (Table 3, and compare also the examples for BASFOR in Fig. 1). There were differences, however, in how the likelihood distributions responded to the calibration, as can be seen from the posterior distributions of likelihoods (Fig. 4).

[Fig. 4 HERE]

For each model and each parameter vector sampled from the prior and posterior distributions, we compared the simulated time series of the ten output variables with the corresponding data, by calculating the correlation coefficient (r) and the normalised root mean square error (NRMSE). Each parameter distribution thus induced ten different distributions of r and NRMSE, of which we show the 5, 50 and 95% quantiles (Table 4). The posterior values of the quantiles for r are often not improvements over the prior, except for Q5. So calibration tended to remove only the parameter vectors with the poorest output-data correlation. In contrast, NRMSE was improved for almost every quantile of every variable in each model (Table 4).

The MSE-decompositions for time series with monthly averages, both for the prior and posterior parameter modes, are shown in Fig. 5. Phase shift, variance error and bias were reduced to different extent for the different models.

[Fig. 5 HERE]

4. Discussion

4.1 *Bayesian calibration: methodological issues*

Bayesian calibration uses data to update the joint probability distribution for a model‟s parameters. The Bayesian approach allows for non-Gaussian distributions for both parameter uncertainty and measurement error. Our calibration was therefore based on sampling by means of MCMC rather than on matrix inversion methods. This in turn allowed us to include

451 systematic data error in the calibration, rather than having to estimate error terms in a first
452 separate step, as was done for example by Michalak et al. (2005), using maximum-likelihood
453 estimation.

454 Although the theory is straightforward, it is easy to overestimate the information
455 content of any dataset, and this may lead to unsupported changes in the parameter
456 distributions. For example, when the common assumption is made that each new data point
457 adds independent new information to the calibration, the parameter distributions will
458 asymptotically collapse with sample size (Clark, 2005). Modellers thus need to elicit realistic
459 assessments of measurement uncertainty from the data-providers (Moala and O'Hagan,
460 2010). This issue was important in the case-study presented here because the dataset was
461 fairly large (n=225), covering times series of four variables. We applied four techniques to
462 ensure a realistic, albeit subjective assessment of the information content of the data: (1)
463 using the monthly temporal scale as the one at which the models were supposed to be
464 applicable, together with the frequency distribution of daily emission events, (2) allowing for
465 random errors in the data, (3) allowing for systematic errors in the data by the use of the four
466 scaling factors, (4) using a heavy-tailed likelihood function (Sivia, 2006). The adjustment of
467 temporal scale is a common technique in atmospheric physics, applied whenever models
468 produce more smooth results than measurements and thereby induce apparent correlations
469 between measurement errors (Prinn, 2000). We considered the implementation of these four
470 techniques to be partly the responsibility of the data-providers (in the case of random and
471 systematic errors) and partly that of the modellers and data-providers together (temporal scale
472 and likelihood function). Using this approach, parameter uncertainties were reduced
473 markedly but the distributions did not collapse. The techniques applied are generic and may
474 be widely applicable to calibration of complex dynamic models using long time series.
475

476 4.2 *Bayesian calibration: impact on parameter uncertainty of the forest models*

477 Parameter uncertainties were reduced strongly compared to the prior, and the likelihood
478 distributions were shifted toward higher values for all four models (Fig. 4). The degree of
479 uncertainty reduction varied between the models, as did the balance between changing data-
480 scaling parameters and process parameters. Although our dataset included measurements of
481 the three major biogeochemical cycles (nitrogen, carbon and water), there was still some lack
482 of balance because about 75% of the data points were for emissions of N_2O and NO (Table
483 1). Therefore, most of the improvement of model behaviour (reduction of likelihood and
484 NRMSE, increase in *r*) was for these variables. For all models, the parameters that were
485 changed the most were related to the soil nitrogen dynamics.

486 The results of our preliminary parameter sensitivity analysis, consisting of calculating
487 partial correlations with the different outputs, need to be interpreted with care. A high value
488 of the PCC for a specific parameter-output combination suggests that the parameter – within
489 its range of uncertainty - strongly affects the output. Therefore knowledge about the process
490 governed by that parameter is key to understanding variability in the output. The opposite
491 may not be true: a strong but non-linear effect may yield a low PCC. Also, the importance of
492 a parameter is not an intrinsic property: it depends on the distribution of that parameter.
493 Whenever Bayesian calibration reduces the variance of a parameter, the contribution of that
494 parameter to output variability is expected to decrease. However, PCC-analysis is not a
495 variance-decomposition method (Saltelli et al., 2000), so it may not be able to show that
496 effect, and indeed, for two out of the four models posterior values of PCC were generally
497 larger than the prior values.

498 With these caveats, the results of the PCC-analysis did reveal commonalities between
499 the models. Across the models, N-emissions were mainly correlated with N_2O-NO
500 partitioning, soil respiration mainly with decomposition but also tree productivity, and soil

water content with leaf area dynamics. These agreements between the models suggest that some of the differences between complex process-based models may not overly affect their behaviour. The models compared here all represent, albeit in very different ways, the linkages between the ecosystem C-, N- and H_2O-cycles. Therefore there are inevitable similarities in the overall feedback structure of the models, imposed by constraints of stoichiometry and mass-balance of the three biogeochemical cycles. These similarities may outweigh details of process representation and parameterisation (Van Oijen et al., 2004; Van Oijen et al., 2010b).

The prior PCC-values were, as expected, indicative of which parameters were most informed by the data in the subsequent calibration. For all four models, the relative decrease of marginal variance tended to be greatest for those parameters that had a strong *a priori* correlation with N_2O-emission, the output variable for which the largest number of data points were available. The parameter variance reduction accounted for by the prior PCC-values ranged from 25% (COUP, MoBiLE) to 29% (BASFOR) and even 35% (DAYCENT). PCC-analysis might thus play a useful role in parameter screening before calibration.

4.3 *Bayesian model comparison*

Comprehensive model comparison requires taking into account parameter uncertainty. A complex model might, in principle, be able to predict complex biogeochemical time series more closely than a simple model. But if it is unclear what the parameterisation of the complex model for good prediction should be, then its predictive capacity is reduced. A simple model whose parameters are well-known might then perform better. Regarding model complexity, there is a trade-off between the need to represent the intricacies of the real world and the need to minimise parameter uncertainty. We thus need a method for comparing the behaviour of the models not just at the modes of the parameter distributions or at the maximum likelihood estimates, but across their whole parameter distributions. Bayesian

model comparison is such a method. This method has been used before in environmental modelling, in the elegant study by Tuomi et al. (2008) who compared different functions describing the impact of temperature on soil respiration, but to our knowledge this is the first application to complex process-based ecosystem models. Formally, the relative magnitudes of the integrated likelihoods equate to relative probabilities for the individual models of being correct, conditional on a correct model being present in the comparison. However, in environmental modelling all models are incorrect in some way, so we prefer to use the integrated likelihoods as a guide towards plausible model structures rather than as probabilities of correctness (Gelman and Shalizi, 2010; Kass and Raftery, 1995).

Bayesian model comparison requires that a common likelihood function is used with all models – because the criterion for comparison is the integrated likelihood of the models, where the integration is over the prior parameter distribution. Therefore only those data can appear in the likelihood function, which are part of the output set of each model. For example, in the current study, we could not include vertical profiles of soil temperature in the likelihood function because the simplest model BASFOR has only one soil compartment. Three of the models had remarkably similar values of the integrated likelihood, the exception being the low value for DAYCENT, which was thus identified as the least plausible model. There was only a slight preference for MoBiLE-DNDC. We analysed these prior integrated likelihood values further by considering the underlying distributions of likelihoods associated with the four different categories of output variable (N_2O, NO, respiration, water) (Fig. 4). The lower value of the integrated likelihood for DAYCENT can be seen to be mainly due to poorer performance for N_2O. The similarity between the integrated likelihoods for the other models was seen to extend to the underlying distributions of category-wise likelihoods (Fig. 4).

4.4 *Analysis of model-data mismatch*

The NRMSE and *r* statistics provided useful additional information beyond the formal Bayesian model comparison. When going from prior to posterior, *r* did not improve (apart from Q5) in any of the models, but NRMSE improved throughout (Table 4). The calibration thus was more successful in reducing the average magnitude of the differences between simulations and measurements, than in aligning the distribution of the outputs over time. There clearly remain difficulties for all models in simulating the large interannual variation in nitrogenous emission characteristics. The models were also similar in that the posterior NRMSE was lowest for soil water content and highest for the monthly values of N_2O-emission (Table 4). It was easier to simulate water than nitrogen dynamics.

The decomposition of the MSE for the modes of the prior and posterior parameter distributions gave further information. The MSE-decomposition is only possible for long time series, i.e. the monthly data (Table 1). There were not enough annual quantiles available to allow the reliable estimation of the variance and phase shift terms. The calibration reduced the MSE for the parameter modes of all four models, and all four categories of output variables (Fig. 5), confirming the effectiveness of the calibration. However, the analysis revealed large differences between the models. The simplest model, BASFOR, had the highest variance error for N_2O and NO. This suggests that a simple model may not be able to respond quickly enough to changes in the environment that affect nitrogenous emissions. The low integrated likelihood of DAYCENT has already been attributed to poor simulation of N_2O-emission (Fig. 4), and the MSE-decomposition showed that this was mainly due to a large phase shift for N_2O emission (Fig. 5). In fact, DAYCENT had very low bias and variance error for N_2O, so it was able to capture general characteristics (mean, „peakiness“) of the time series of emission very well, but not the timing of emission events.

Most data were in the central Q5-Q95 ranges of the prior output distributions of the

models (Table 3). However, perhaps surprisingly, most data were to be found in the upper tails (>Q95) of the posterior output distributions (Table 3). This is because of a trade-off between the different variables in the calibration. Such trade-off is inevitable with models that are imperfect and cannot capture the whole range of behaviour of all variables simultaneously. Moreover, there is also the distinct possibility of systematic error in the measurement of the different categories of output variables. Where the models were unable to reconcile all the data, the calibration tended to modify the settings of the scale parameters (Fig. 2). It is important to establish whether the likely underlying cause of the revision of the scale factors was indeed systematic measurement error or model structural error. One way of attempting this is to consider the differences between the models in their posterior estimates for the scaling factors. *A posteriori,* all models suggested that the measurements for CO_2-emissions were unrealistically high (Fig. 2: γ_{CO2} mostly <1), but only the BASFOR model suggested the same for the N-emissions (γ_{N2O} and γ_{NO} <<1). There thus is some doubt about the CO_2-measurements, and likewise about the capacity of BASFOR to simulate N-dynamics.

Note that, in our approach, the data-scaling factors are intended to represent the idiosyncrasies of specific datasets, so we cannot expect the calibrated scaling values to apply elsewhere. However, if calibration of a model using data from multiple sites were shown to consistently lead to scaling factors different from unity, we would expect the error to be predominantly the model‟s rather than that of the data. Here we only had data from one site available, so no strong conclusions can be drawn.

4.5 *The Bayesian framework*

The three-operation Bayesian framework proposed here - calibration, comparison, analysis of model-data mismatch – was shown to work well in this study. Most of the techniques employed in each of these operations are novel in their application to complex dynamic

models of forest biogeochemistry, in particular in combination with the use of data from processes that vary strongly over time (Luo et al., 2009). The application of the framework showed that model parameter uncertainty could be reduced in all models, irrespective of their level of complexity. However, it also showed that the models still suffer from structural deficiencies, even if we allow for the possibility of errors in the data, and that the deficiencies are stronger for the nitrogen cycle than for the carbon and water cycles.

There are limitations associated with this study and a caveat has to be made regarding the use of the results. We only used data from one site, Höglwald, and the general applicability of the models to European forests needs to be tested with data from other sites. When data from these new sites become available, the posterior distributions found here will become prior distributions in further calibration. Another limitation of the study was the use of parameter screening, which was considered necessary for the three models with the highest computational demand, but there may be environmental conditions under which simulated forest biogeochemistry is sensitive to the excluded parameters. Also, uncertainty concerning environmental drivers such as weather conditions was ignored as it was expected to be small compared to the structural and parameter uncertainties. This assumption needs to be verified.

Wang et al. (2009) called for the development of an integrated Bayesian framework that can account for the different sources and types of error arising in environmental modelling. The Bayesian framework proposed here is an integrated one in the sense that its three operations were linked methodologically and in that its three operations provide complementary information. Methodologically, the Bayesian calibration made use of the same likelihood function as the Bayesian model comparison. Calibration was also linked to comparison in the use of MCMC to explore parameter space, with the thus generated sample being used in the model comparison for estimating the integrated likelihood. Because this estimate, based on the harmonic mean of the sampled likelihoods, can be unstable (Chib and

626 Jeliazkov, 2001), we also used the method of directly sampling from the prior.
627 Methodological links further existed between the calibration and the analysis of model-data
628 mismatch, in that the sample generated by the calibration was used to calculate the NRMSE
629 across the posterior parameter distribution, rather than just for one parameter vector.

630 More importantly than these methodological links, the three operations in the
631 framework also complemented each other in how they help improve the modelling.
632 Calibration reduced parameter uncertainty, model comparison reduced uncertainty about the
633 relative plausibility of the different models, and the analysis of model-data mismatch showed
634 which parts of the models needed most improvement, which in this case was the nitrogen
635 dynamics.

636

637

638 ## 5. Conclusions

639 - Bayesian calibration can be used to reduce parametric uncertainty of complex
640 dynamic models for forest biogeochemistry.
641 - Bayesian calibration allows for the use of datasets that contain long time series of gas
642 emissions with high intra- and interannual variability, and with both random and
643 systematic error.
644 - Data need to be compared with models at the appropriate temporal scale. This may
645 involve, as shown here, monthly averaging and the calculation of annual frequency
646 distributions. These transformations, and the use of heavy-tailed likelihood functions
647 that account for uncertainty about random and systematic measurement errors, can
648 help prevent collapse of the parameter distributions in the calibration.
649 - Bayesian model comparison can be used to calculate the relative conditional
650 probabilities of models being correct, irrespective of the type and complexity of the

considered models.

- Bayesian model comparison treats models as black boxes, so it can only identify which models are implausible, but it cannot identify any specific model deficiencies.
- Analysis of model-data mismatch can help identify model weaknesses by decomposition of the MSE and by showing how the NRMSE and the correlation coefficient *r* vary for the different processes simulated by the models.
- Together, the three operations of Bayesian calibration, Bayesian model comparison, and analysis of model-data mismatch, constitute a promising framework for uncertainty reduction and improvement of complex dynamic models in forest biogeochemistry.
- This was confirmed by the case-study analysed here, in which four different parameter-rich process-based models of forest biogeochemistry were confronted with long time-series of biogeochemical data. Parameter uncertainties were reduced in all models and the relative model plausibilities were quantified, with MoBiLE-DNDC having a slight preference over the other models. The simplest model, BASFOR, was shown to underestimate variance of nitrogenous emissions even after calibration. The model of intermediate complexity, DAYCENT, simulated the time series well but with large phase shift. COUP and MoBiLE-DNDC were able to remove most bias through calibration.
- The calibration not only reduced parameter and model uncertainty, but also identified possible systematic error in the measurement of soil respiration, to which all models assigned data-scaling factors less than unity with high posterior probability.
- There remain patterns in the data - in particular infrequent events of very high nitrogenous emission rate - that are unexplained by any of the models, even after calibration. Given the intensive exploration of parameter space in the calibration, this

is unlikely to be due to incorrect model parameterisation.

- The analysis showed that the models still suffer from structural deficiencies, even if we allow for the possibility of errors in the data. The deficiencies are stronger for the nitrogen cycle than for the carbon and water cycles.

Acknowledgements

We thank the European Union for financial support to carry out this work in projects NitroEurope and Carbo-Extreme, and we are grateful to our colleagues in these projects for discussion. J.B.Y. wishes to acknowledge the help of M. Richards in programming the routines for the Bayesian analysis of DAYCENT. We express our thanks to two anonymous reviewers for their constructive comments on the original manuscript.

References

Butterbach-Bahl, K., Rothe, A. and Papen, H., 2002. Effect of tree distance on N_2O and CH_4-fluxes from soils in temperate forest ecosystems. Plant and Soil, 240(1): 91-103.

Chib, S. and Jeliazkov, I., 2001. Marginal likelihood from the Metropolis-Hastings output. Journal of the American Statistical Association, 96(453): 270-281.

Clark, J.S., 2005. Why environmental scientists are becoming Bayesians. Ecology Letters, 8(1): 2-14.

de Bruijn, A.M.G., Butterbach-Bahl, K., Blagodatsky, S. and Grote, R., 2009. Model evaluation of different mechanisms driving freeze-thaw N_2O emissions. Agriculture, Ecosystems & Environment, 133(3-4): 196-207.

de Vries, W. et al., 2009. The impact of nitrogen deposition on carbon sequestration by European forests and heathlands. Forest Ecology and Management, 258(8): 1814-1823.

Del Grosso, S.J. et al., 2001. Simulated interaction of carbon dynamics and nitrogen trace gas fluxes using the

703 DAYCENT model. In: M. Schaffer, et al. (Editor), Modeling Carbon and Nitrogen Dynamics for Soil
704 Management. CRC Press, BOca Raton, pp. 303-332.
705 Fox, A. et al., 2009. The REFLEX project: Comparing different algorithms and implementations for the
706 inversion of a terrestrial ecosystem model against eddy covariance data. Agricultural and Forest
707 Meteorology, 149(10): 1597-1615.
708 Gelman, A. and Shalizi, C.R., 2010. Philosophy and the practice of Bayesian statistics. Working paper,
709 http://arxiv.org/abs/1006.3868.
710 Hyvönen, R. et al., 2007. The likely impact of elevated [CO_2], nitrogen deposition, increased temperature, and
711 management on carbon sequestration in temperate and boreal forest ecosystems. A literature review.
712 New Phytologist, 173: 463-480.
713 Jaynes, E.T., 2003. Probability Theory: The Logic of Science. Cambridge University Press, Cambridge, 758 pp.
714 Kahle, H.P. et al. (Editors), 2008. Causes and Consequences of Forest Growth Trends in Europe. Brill, Leiden,
715 xiv + 261 pp.
716 Kass, R.E. and Raftery, A.E., 1995. Bayes Factors. Journal of the American Statistical Association, 90(430):
717 773-795.
718 Keats, A., Cheng, M.T., Yee, E. and Lien, F.S., 2009. Bayesian treatment of a chemical mass balance receptor
719 model with multiplicative error structure. Atmospheric Environment, 43(3): 510-519.
720 Kesik, M. et al., 2005. Inventories of N_2O and NO emissions from European forest soils. Biogeosciences, 2(4):
721 353-375.
722 Klemedtsson, L. et al., 2008. Bayesian calibration method used to elucidate carbon turnover in forest on drained
723 organic soil. Biogeochemistry, 89(1): 61-79.
724 Kobayashi, K. and Salam, M.U., 2000. Comparing simulated and measured values using mean squared
725 deviation and its components. Agronomy Journal, 92(2): 345-352.
726 Kreutzer, K., Butterbach-Bahl, K., Rennenberg, H. and Papen, H., 2009. The complete nitrogen cycle of an N-
727 saturated spruce forest ecosystem. Plant Biology, 11(5): 643-649.
728 Kroon, P.S. et al., 2010. Uncertainties in eddy covariance flux measurements assessed from CH_4 and N_2O
729 observations. Agricultural and Forest Meteorology, 150(6): 806-816.
730 Li, C.S., Aber, J., Stange, F., Butterbach-Bahl, K. and Papen, H., 2000. A process-oriented model of N_2O and
731 NO emissions from forest soils: 1. Model development. Journal of Geophysical Research-
732 Atmospheres, 105(D4): 4369-4384.

733 Li, C.S., Frolking, S. and Frolking, T.A., 1992. A model of nitrous-oxide evolution from soil driven by rainfall
734 events.1. Model structure and sensitivity. Journal of Geophysical Research-Atmospheres, 97(D9):
735 9759-9776.
736 Luo, Y. et al., 2009. Parameter identifiability, constraint, and equifinality in data assimilation with ecosystem
737 models. Ecological Applications, 19(3): 571-574.
738 Luyssaert, S. et al., 2010. The European carbon balance. Part 3: forests. Global Change Biology, 16(5): 1429-
739 1450.
740 McCulloch, R.E. and Rossi, P.E., 1992. Bayes factors for nonlinear hypotheses and likelihood distributions.
741 Biometrika, 79(4): 663-676.
742 Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E., 1953. Equation of state
743 calculations by fast computing machines. Journal of Chemical Physics, 21: 1087-1092.
744 Michalak, A.M. et al., 2005. Maximum likelihood estimation of covariance parameters for Bayesian
745 atmospheric trace gas surface flux inversions. Journal of Geophysical Research-Atmospheres,
746 110(D24): 16.
747 Moala, F.A. and O'Hagan, A., 2010. Elicitation of multivariate prior distributions: A nonparametric Bayesian
748 approach. Journal of Statistical Planning and Inference, 140(7): 1635-1655.
749 Morris, M.D., 1991. Factorial sampling plans for preliminary computational experiments. Technometrics, 33(2):
750 161-174.
751 Norman, J. et al., 2008. Simulation of NO and N_2O emissions from a spruce forest during a freeze/thaw event
752 using an N-flux submodel from the PnET-N-DNDC model integrated to CoupModel. Ecological
753 Modelling, 216(1): 18-30.
754 Ogle, K., 2009. Hierarchical Bayesian statistics: merging experimental and modeling approaches in ecology.
755 Ecological Applications, 19(3): 577-581.
756 Ogle, K. and Barber, J.J., 2008. Bayesian data-model integration in plant physiological and ecosystem ecology.
757 Progress in Botany, 69: 281-311.
758 Papen, H. and Butterbach-Bahl, K., 1999. A 3-year continuous record of nitrogen trace gas fluxes from
759 untreated and limed soil of a N-saturated spruce and beech forest ecosystem in Germany. 1. N_2O
760 emissions. Journal of Geophysical Research-Atmospheres, 104(D15): 18487-18503.
761 Parton, W.J. et al., 1993. Observations and modeling of biomass and soil organic-matter dynamics for the
762 grassland biome worldwide. Global Biogeochemical Cycles, 7(4): 785-809.

763 Prinn, R.G., 2000. Measurement equation for trace chemicals in fluids and solution of its inverse. In: P.
764 Kasibhatla et al. (Editors), Inverse Methods in Global Biogeochemical Cycles. Geophysical Mongraph.
765 American Geophysical Union, Washington DC, pp. 3-18.
766 Raupach, M.R. et al., 2005. Model-data synthesis in terrestrial carbon observation: methods, data requirements
767 and data uncertainty specifications. Global Change Biology, 11(3): 378-397.
768 Rothe, A., 1997. Einfluss des Baumartenanteils auf Durchwurzelung, Wasserhaushalt, Stoffhaushalt und
769 Zuwachsleistung eines Fichten-Buchen-Mischbestandes am Standort Höglwald. Forstliche
770 Forschungsberichte: 163 pp.
771 Rougier, J.C., (in press). Formal Bayes methods for model calibration with uncertainty. In: K. Beven and J. Hall
772 (Editors), Applied Uncertainty Analysis for Flood Risk Management. Imperial College Press / World
773 Scientific, London. Draft version available at http://www.maths.bris.ac.uk/~mazjcr/FRMbox2-4.pdf.
774 Saltelli, A., Chan, K. and Scott (Eds.), E.M. (Editors), 2000. Sensitivity Analysis. Wiley, Chichester, xv + 475.
775 pp.
776 Sivia, D.S., 2006. Data Analysis: A Bayesian Tutorial. Second edition. Oxford University Press, Oxford, 260
777 pp.
778 Sutton, M.A. et al., 2007. Challenges in quantifying biosphere-atmosphere exchange of nitrogen species.
779 Environmental Pollution, 150: 125-139.
780 Sutton, M.A. et al., 2008. Uncertainties in the relationship between atmospheric nitrogen deposition and forest
781 carbon sequestration. Global Change Biology, 14(9): 2057-2063.
782 Svensson, M. et al., 2008. Bayesian calibration of a model describing carbon, water and heat fluxes for a
783 Swedish boreal forest stand. Ecological Modelling, 213(3-4): 331-344.
784 Tuomi, M., Vanhala, P., Karhu, K., Fritze, H. and Liski, J., 2008. Heterotrophic soil respiration--Comparison of
785 different models describing its temperature dependence. Ecological Modelling, 211(1-2): 182-190.
786 Van Oijen, M. et al., 2008a. Evaluation of past and future changes in European forest growth by means of four
787 process-based models. In: H.P. Kahle et al. (Editors), Causes and Consequences of Forest Growth
788 Trends in Europe. Brill, Leiden, pp. 183-199.
789 Van Oijen, M. et al., 2008b. Methodology for the application of process-based models to analyse changes in
790 European forest growth. In: H.P. Kahle et al. (Editors), Causes and Consequences of Forest Growth
791 Trends in Europe. Brill, Leiden, pp. 67-80.
792 Van Oijen, M., Cannell, M.G.R. and Levy, P.E., 2004. Modelling biogeochemical cycles in forests: state of the

art and perspectives. In: F. Andersson, Y. Birot and R. Päivinen (Editors), Towards the Sustainable Use of Europe's Forests - Forest Ecosystem and Landscape Research: Scientific Challenges and Opportunities. EFI Proceedings. European Forest Institute, pp. 157-169.

Van Oijen, M., Dauzat, J., Harmand, J.-M., Lawson, G. and Vaast, P., 2010a. Coffee agroforestry systems in Central America: II. Development of a simple process-based model and preliminary results. Agroforestry Systems, 80(3): 361-378.

Van Oijen, M., Rougier, J. and Smith, R., 2005. Bayesian calibration of process-based forest models: bridging the gap between models and data. Tree Physiology, 25(7): 915-927.

Van Oijen, M., Schapendonk, A. and Höglind, M., 2010b. On the relative magnitudes of photosynthesis, respiration, growth and carbon storage in vegetation. Annals of Botany, 105(5): 793-797.

Van Oijen, M. and Thomson, A., 2010. Toward Bayesian uncertainty quantification for forestry models used in the United Kingdom Greenhouse Gas Inventory for land use, land use change, and forestry. Climatic Change, 103(1): 55-67.

Wang, Y.-P., Trudinger, C.M. and Enting, I.G., 2009. A review of applications of model-data fusion to studies of terrestrial carbon fluxes at different scales. Agricultural and Forest Meteorology, 149(11): 1829-1842.

Werner, C., Butterbach-Bahl, K., Haas, E., Hickler, T. and Kiese, R., 2007. A global inventory of N_2O emissions from tropical rainforest soils using a detailed biogeochemical model. Global Biogeochem. Cycles, 21(3): GB3010.

Wu, X. et al., 2010. Environmental controls over soil-atmosphere exchange of N_2O, NO, and CO_2 in a temperate Norway spruce forest. Global Biogeochem. Cycles, 24(2): GB2012.

Yeluripati, J.B. et al., 2009. Bayesian calibration as a tool for initialising the carbon pools of dynamic soil models. Soil Biology & Biochemistry, 41(12): 2579-2583.

Table 1. Overview of data used for calibration. Variables are annual quantiles of daily emission rates where indicated, and monthly averages otherwise. The period of measurement indicates the years of first and last measurement, and *n* is the total number of data points over that period. The columns marked *min*, *mean* and *max* show the extreme values and the mean of the *n* values.

Variable	Unit	Period of measurement	*n*	*min*	*mean*	*max*
N_2O (Q10)	kg N ha^{-1} y^{-1}	1994-2003	6	0.08	0.21	0.31
N_2O (Q50)	kg N ha^{-1} y^{-1}	1994-2003	6	0.27	0.52	0.85
N_2O (Q90)	kg N ha^{-1} y^{-1}	1994-2003	6	0.55	2.32	9.14
NO (Q10)	kg N ha^{-1} y^{-1}	1994-2002	5	1.65	2.33	2.79
NO (Q50)	kg N ha^{-1} y^{-1}	1994-2002	5	4.81	6.64	8.30
NO (Q90)	kg N ha^{-1} y^{-1}	1994-2002	5	10.65	13.65	18.52
N_2O	kg N ha^{-1} y^{-1}	1994-2003	70	0.03	0.99	16.55
NO	kg N ha^{-1} y^{-1}	1994-2003	61	1.74	7.72	24.04
Rsoil	mg C m^{-2} h^{-1}	1995-1997	36	23.9	113.3	255.2
Water	% (vol)	1994-1996	25	27.1	33.6	37.0

830

Table 2. Overview of the models.

Property	BASFOR	COUP	DAYCENT	MoBiLE-DNDC
# State variables (trees, soil)	14 (6,8)	56 (8,48)	20 (10,10)	38 (10,28)
# Parameters (in calibration)	48 (48)	>300 (23)	>300 (17)	67* (26)
Time step	Daily	Hourly	Daily	Daily (but less than 1 minute for diffusion processes)
Inputs: Environmental time series	Radiation, temperature, precipitation, humidity, wind speed, N-deposition, [CO_2]	Radiation, temperature, precipitation, humidity, wind speed, N-deposition	Radiation, temperature, precipitation, humidity, wind speed, N-deposition, [CO_2]	Radiation, temperature, precipitation, N-deposition, [CO_2]
Inputs: environmental constants	Soil water retention curve, rooting depth	Soil water retention curve, rooting depth	Soil water retention curve, rooting depth	Layer-specific texture, bulk density, field capacity, pH

831 ***** For the soil chemistry module only.

832

833

Table 3. Summary of model output, with prior and posterior uncertainties. All values shown refer to time series averages, to be compared with the last-but-one column of Table 1. For each model, the table entries are three quantiles (5, 50 and 95%) of the output distributions generated by the prior and posterior parameter distributions. In **bold**: posterior distributions for which the posterior width (Q5-Q95) is at least an order of magnitude less than for the prior.

Var.	Unit	Dist.	BASFOR			COUP			DAYCENT			MoBiLE-DNDC		
			Q5	Q50	Q95	Q5	Q50	Q95	Q5	Q50	Q95	Q5	Q50	Q95
N_2O (Q10)	kg N ha^{-1} y^{-1}	Prior	0.01	0.05	1.78	0.01	0.03	0.13	0.01	0.02	0.02	0.09	0.68	2.82
		Post.	**0.02**	**0.03**	**0.04**	0.03	0.05	0.07	0.01	0.01	0.02	**0.13**	**0.16**	**0.21**
N_2O (Q50)	kg N ha^{-1} y^{-1}	Prior	0.03	0.18	3.83	0.02	0.10	0.49	0.34	0.41	0.50	0.19	1.64	6.82
		Post.	**0.07**	**0.12**	**0.17**	0.06	0.10	0.16	0.29	0.33	0.39	**0.25**	**0.30**	**0.34**
N_2O (Q90)	kg N ha^{-1} y^{-1}	Prior	0.12	1.13	8.08	0.16	0.69	4.49	2.00	2.13	2.28	0.40	3.21	12.85
		Post.	**0.29**	**0.46**	**0.65**	**0.11**	**0.18**	**0.33**	1.90	1.99	2.08	**0.41**	**0.51**	**0.62**
NO (Q10)	kg N ha^{-1} y^{-1}	Prior	0.03	0.10	8.99	3.42	7.24	14.26	0.72	0.75	0.78	0.27	2.88	12.22
		Post.	**0.05**	**0.11**	**0.18**	1.60	2.10	2.77	0.71	0.72	0.75	**1.84**	**2.28**	**2.72**
NO (Q50)	kg N ha^{-1} y^{-1}	Prior	0.09	0.45	12.21	5.95	13.02	28.90	1.16	1.26	1.37	0.92	5.83	17.91
		Post.	**0.25**	**0.79**	**1.43**	**2.41**	**3.20**	**4.64**	1.08	1.15	1.23	**3.52**	**4.19**	**4.86**
NO (Q90)	kg N ha^{-1} y^{-1}	Prior	0.31	2.38	19.99	10.56	22.88	48.30	0.51	0.57	0.63	1.89	9.39	23.70
		Post.	0.99	2.69	4.57	4.41	6.22	8.70	0.48	0.51	0.55	5.61	6.65	8.02
N_2O	kg N ha^{-1} y^{-1}	Prior	0.05	0.39	4.42	0.07	0.27	1.44	1.55	1.82	2.11	0.24	1.81	7.42
		Post.	**0.11**	**0.18**	**0.24**	**0.08**	**0.12**	**0.20**	1.40	1.55	1.74	**0.27**	**0.32**	**0.37**
NO	kg N ha^{-1} y^{-1}	Prior	0.15	0.92	13.25	6.05	14.78	32.82	4.04	4.97	6.00	1.21	6.00	17.58
		Post.	0.43	1.16	1.99	2.96	4.11	5.68	3.54	4.05	4.74	**3.68**	**4.32**	**5.01**
CO_2	mg C m^{-2} h^{-1}	Prior	44.2	69.4	106.8	76.5	97.0	127.0	25.3	48.2	73.0	46.4	66.3	96.5
		Post.	56.2	76.5	99.5	67.6	84.4	101.2	26.9	49.6	76.2	49.1	55.4	64.2
Water	% (vol)	Prior	28.3	31.6	33.7	33.2	34.5	36.1	32.3	32.3	32.3	34.5	34.5	34.5
		Post.	28.7	31.0	33.0	33.3	34.6	35.9	32.3	32.3	32.3	34.5	34.5	34.5

836

Table 4. Comparison of data with model outputs: correlation coefficient (*r*) and normalised root mean square error (NRMSE). The table shows quantiles (Q5, Q50, Q95) of the distributions of *r* and NRMSE induced by prior and posterior parameter distributions. In **bold**: posterior values that are improvements over the prior (r increased, NRMSE reduced).

Var.	Dist.	Statistic	BASFOR			COUP			DAYCENT			MoBiLE-DNDC		
			Q5	Q50	Q95	Q5	Q50	Q95	Q5	Q50	Q95	Q5	Q50	Q95
N_2O (Q10)	Prior	*r*	-0.37	0.42	0.72	-0.83	-0.09	0.47	-0.36	-0.27	-0.19	-0.27	0.54	0.76
		NRMSE	0.44	1.07	11.95	0.41	0.93	1.91	0.52	1.01	1.95	0.42	2.47	12.95
	Post.	*r*	-0.75	-0.19	0.49	-0.60	-0.22	0.20	-0.39	-0.29	-0.21	**0.21**	**0.70**	**0.87**
		NRMSE	**0.41**	**0.47**	**0.53**	**0.31**	**0.38**	**0.48**	0.54	**0.77**	**1.05**	**0.20**	**0.39**	**1.32**
N_2O (Q50)	Prior	*r*	-0.46	0.65	0.88	-0.62	0.28	0.77	0.45	0.60	0.70	-0.67	-0.14	0.33
		NRMSE	0.40	1.02	7.64	0.35	0.87	1.86	0.23	0.40	1.25	0.48	2.46	13.32
	Post.	*r*	-0.53	-0.04	0.78	-0.73	-0.37	0.38	0.45	**0.62**	**0.71**	**-0.06**	**0.15**	**0.50**
		NRMSE	**0.31**	**0.38**	**0.46**	**0.32**	**0.40**	**0.48**	**0.19**	**0.28**	**0.48**	**0.22**	**0.47**	**1.51**
N_2O (Q90)	Prior	*r*	-0.51	0.32	0.80	-0.57	0.22	0.96	0.60	0.60	0.61	-0.58	-0.41	-0.08
		NRMSE	0.83	1.64	3.13	0.64	1.50	3.01	0.72	1.06	2.34	0.94	2.04	5.43
	Post.	*r*	**-0.35**	**0.55**	**0.86**	-0.68	-0.33	0.70	0.60	0.60	0.61	**-0.38**	**-0.25**	-0.13
		NRMSE	**0.67**	**0.75**	**0.86**	0.77	**0.82**	**0.91**	**0.71**	**0.82**	**1.13**	**0.78**	**1.34**	**3.05**
NO (Q10)	Prior	*r*	0.20	0.52	0.78	-0.11	0.20	0.49	-0.44	-0.44	-0.43	-0.73	0.37	0.55
		NRMSE	0.53	1.13	4.54	0.72	2.26	5.37	0.37	0.75	1.62	0.30	0.97	4.37
	Post.	*r*	**0.39**	**0.60**	**0.84**	**0.05**	**0.41**	**0.67**	-0.44	-0.44	-0.44	**-0.62**	-0.01	**0.66**
		NRMSE	**0.45**	**0.48**	**0.54**	**0.18**	**0.35**	**0.55**	**0.35**	**0.40**	**0.51**	**0.18**	**0.37**	**1.00**
NO (Q50)	Prior	*r*	-0.60	-0.32	0.37	-0.90	-0.68	-0.36	0.14	0.18	0.22	-0.92	-0.21	0.39
		NRMSE	0.50	1.04	1.96	0.53	1.26	3.58	0.38	0.82	1.71	0.24	0.67	1.84

	Post.	*r*	**-0.50**	-0.38	-0.25	**-0.78**	**-0.51**	**-0.20**	0.13	0.16	0.20	**-0.80**	-0.36	0.03
		NRMSE	**0.38**	**0.44**	**0.51**	**0.19**	**0.27**	**0.40**	**0.36**	**0.42**	**0.56**	**0.16**	**0.39**	**1.31**
NO (Q90)	Prior	*r*	-0.59	-0.34	0.11	-0.93	-0.70	-0.36	-0.59	-0.53	-0.43	-0.89	-0.30	0.27
		NRMSE	0.41	0.90	1.78	0.42	1.00	2.79	0.51	0.96	1.86	0.20	0.56	1.44
	Post.	*r*	**-0.51**	-0.35	-0.09	**-0.90**	**-0.53**	**-0.09**	-0.61	-0.58	-0.53	**-0.86**	-0.52	0.01
		NRMSE	**0.28**	**0.37**	**0.48**	**0.14**	**0.25**	**0.42**	**0.48**	**0.54**	**0.69**	**0.15**	**0.49**	1.44
N_2O	Prior	*r*	-0.15	0.07	0.41	-0.14	0.25	0.66	-0.08	-0.08	-0.07	-0.24	-0.19	-0.06
		NRMSE	1.34	2.64	5.00	1.16	2.32	4.46	3.28	4.10	5.29	1.56	3.40	8.58
	Post.	*r*	**-0.13**	-0.04	0.21	-0.26	-0.05	0.36	-0.08	**-0.07**	-0.07	**-0.19**	**-0.12**	**-0.02**
		NRMSE	**1.14**	**1.20**	**1.35**	**1.14**	**1.20**	**1.33**	**2.79**	**3.18**	**3.74**	**1.19**	**2.00**	**4.47**
NO	Prior	*r*	-0.59	-0.32	0.26	0.42	0.55	0.65	0.69	0.69	0.70	0.09	0.52	0.67
		NRMSE	0.62	1.24	2.21	0.50	1.23	3.74	0.39	0.58	1.48	0.37	0.86	1.83
	Post.	*r*	**-0.43**	**-0.20**	0.05	0.42	0.50	0.58	0.69	0.69	0.70	**0.52**	**0.67**	**0.74**
		NRMSE	**0.50**	**0.54**	**0.60**	**0.28**	**0.35**	**0.51**	**0.31**	**0.36**	**0.42**	**0.26**	**0.61**	**1.66**
CO_2	Prior	*r*	0.76	0.80	0.82	0.85	0.87	0.89	0.87	0.89	0.90	0.85	0.89	0.91
		NRMSE	0.27	0.64	1.49	0.25	0.44	1.25	0.24	0.63	1.67	0.17	0.55	1.53
	Post.	*r*	0.76	0.79	0.81	**0.86**	**0.88**	0.89	0.87	0.89	0.90	**0.86**	0.88	0.89
		NRMSE	0.28	**0.39**	**0.64**	**0.20**	**0.28**	**0.53**	**0.22**	**0.34**	**0.76**	0.17	**0.42**	**1.36**
Water	Prior	*r*	-0.16	0.03	0.55	0.36	0.38	0.39	0.53	0.53	0.53	0.55	0.56	0.56
		NRMSE	0.04	0.10	0.26	0.08	0.13	0.26	0.11	0.15	0.28	0.02	0.10	0.25
	Post.	*r*	-0.17	-0.03	0.27	0.36	0.38	0.39	0.53	0.53	0.53	0.55	0.55	0.56
		NRMSE	0.04	**0.07**	**0.13**	0.08	**0.10**	**0.17**	0.11	**0.14**	**0.21**	0.02	**0.09**	**0.24**

837

838

839

840

841

842 FIGURES

843

844 Fig. 1. Dots: Monthly averages of measured emissions of N_2O, NO, CO_2 and of soil water
845 content. The lines represent output of model BASFOR. Dashed red line: output for the
846 mode of the prior parameter distribution. Thick black line: output for the mode of the
847 posterior. Thin black lines: 5% and 95% quantiles of the posterior output distribution.

848

849 Fig. 2. Posterior marginal distributions for the four data-scaling factors that represent
850 systematic multiplicative error in the data according to the different models.

851

852 Fig. 3. Differences between measurements and simulations (positive values indicating
853 underestimates by the models) for time series with monthly averages. Top 4 panels:
854 simulations using the mode of the prior parameter distribution. Bottom 4 panels:
855 simulations with the posterior mode.

856

857 Fig. 4. Distributions of log-likelihoods for each of the four models, for the four categories of
858 output variables. Grey: prior, black: posterior.

859

860 Fig. 5. Decomposition of the Mean Squared Error (MSE) associated with the modes of the
861 prior and posterior parameter distributions, for the time series with monthly data. The
862 MSE-values for the N_2O and NO are in the same units (kg N ha^{-1} y^{-1} squared), the

MSE-values for soil respiration and soil water content are in squared mg C m^{-2} h^{-1} and squared %, respectively.

Figure 1

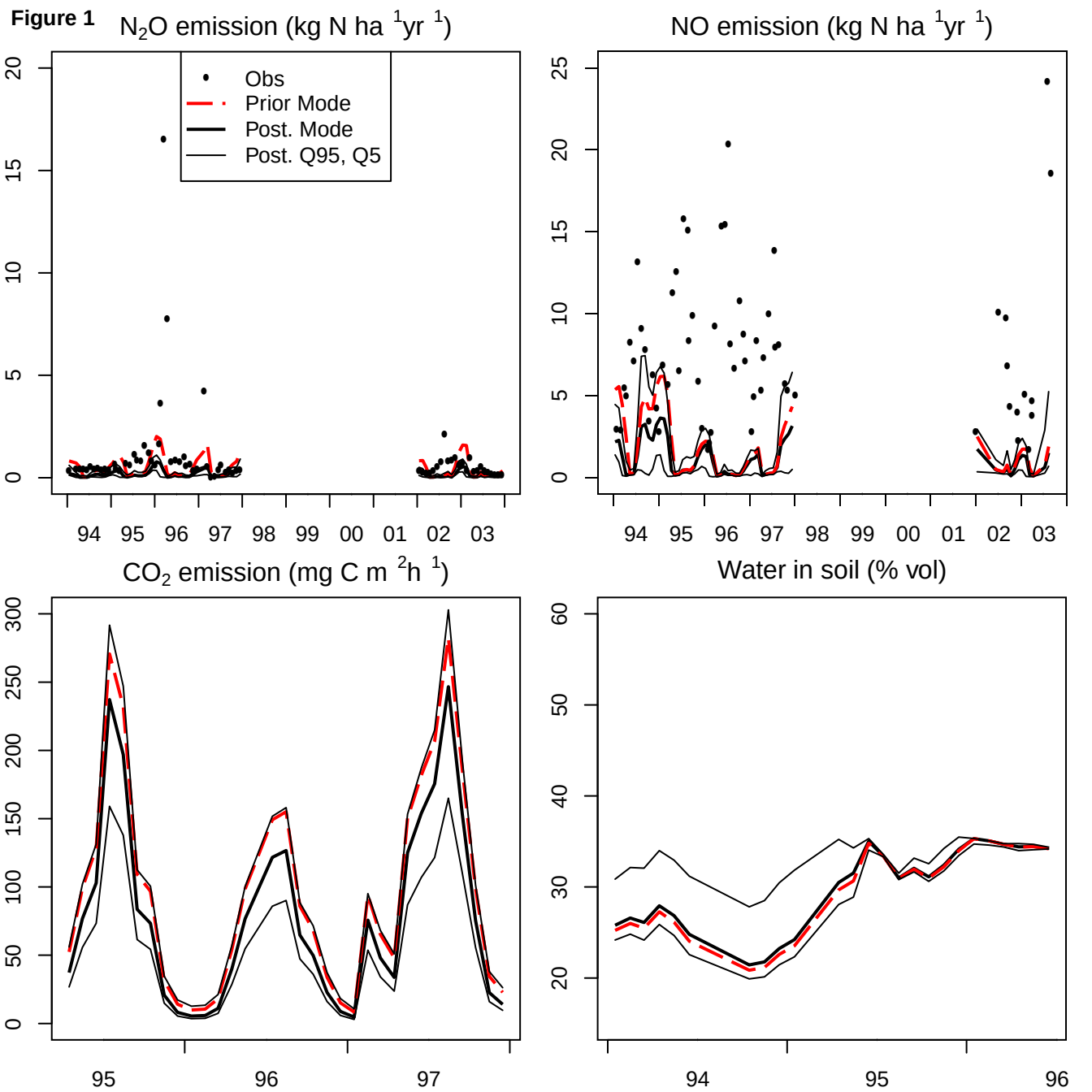

N2O emission (kg N ha 1yr 1)
Obs
Prior Mode
Post. Mode
Post. Q95, Q5
0
5
10
15
20
94
95
96
97
98
99
00
01
02
03
NO emission (kg N ha 1yr 1)
0
5
10
15
20
25
94
95
96
97
98
99
00
01
02
03
CO2 emission (mg C m 2h 1)
0
50
100
150
200
250
300
95
96
97
Water in soil (% vol)
20
30
40
50
60
94
95
96

Figure 2

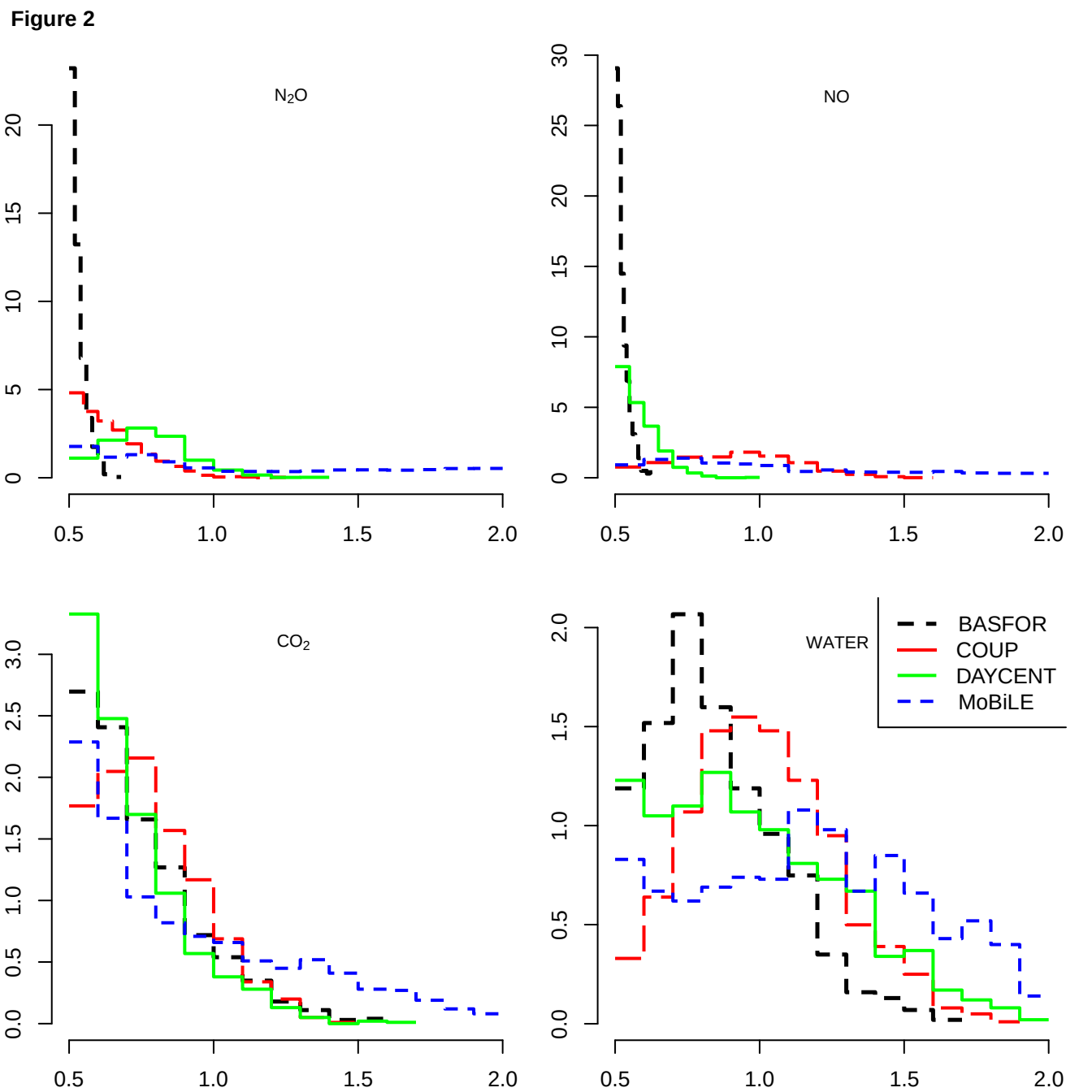
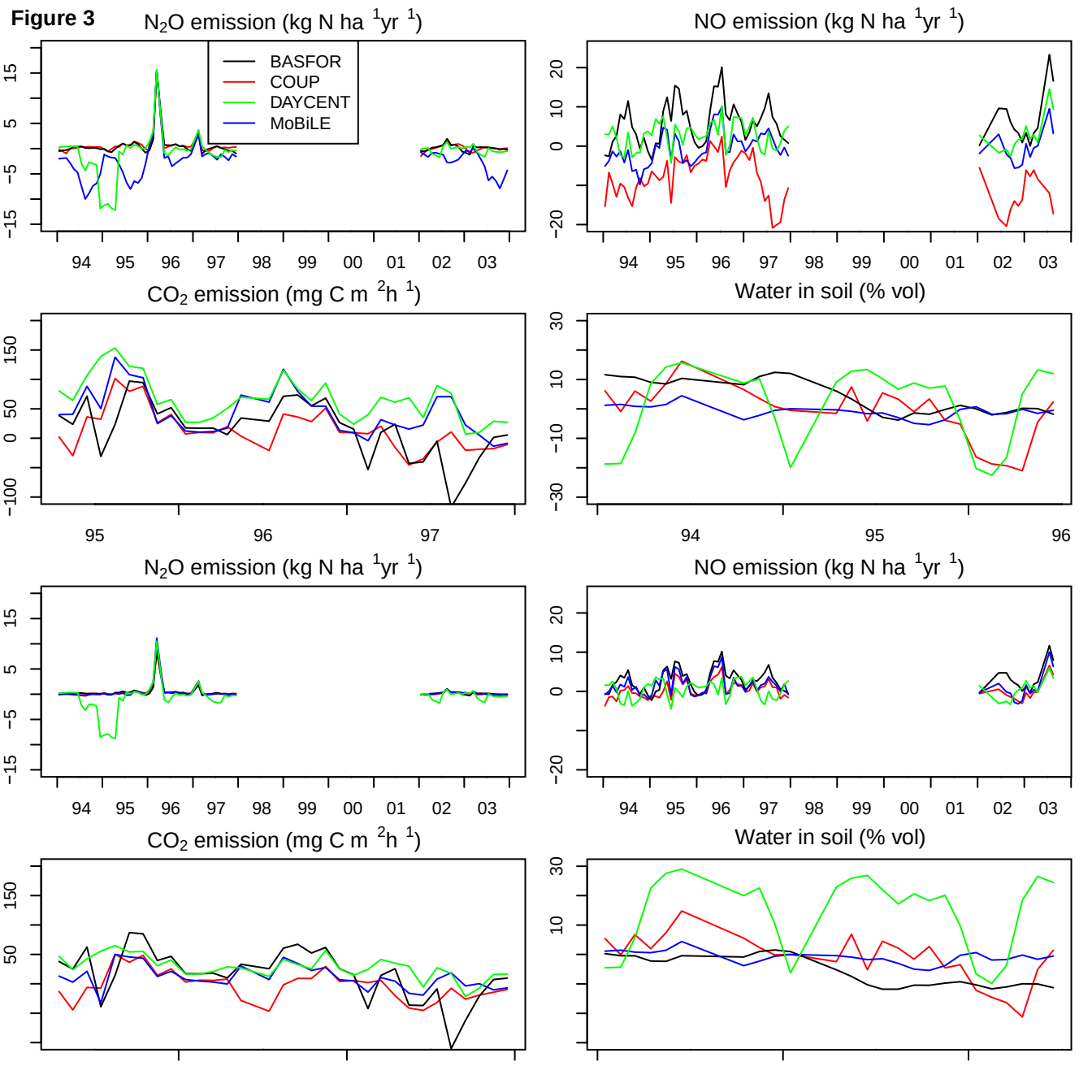


Figure 3


N2O emission (kg N ha 1yr 1)
NO emission (kg N ha 1yr 1)
CO2 emission (mg C m 2h 1)
Water in soil (% vol)
N2O emission (kg N ha 1yr 1)
NO emission (kg N ha 1yr 1)
CO2 emission (mg C m 2h 1)
Water in soil (% vol)
BASFOR
COUP
DAYCENT
MoBiLE
94 95 96 97 98 99 00 01 02 03
95 96 97
94 95 96

0
−100
−10
−30
95
96
97
94
95
96

Figure 4

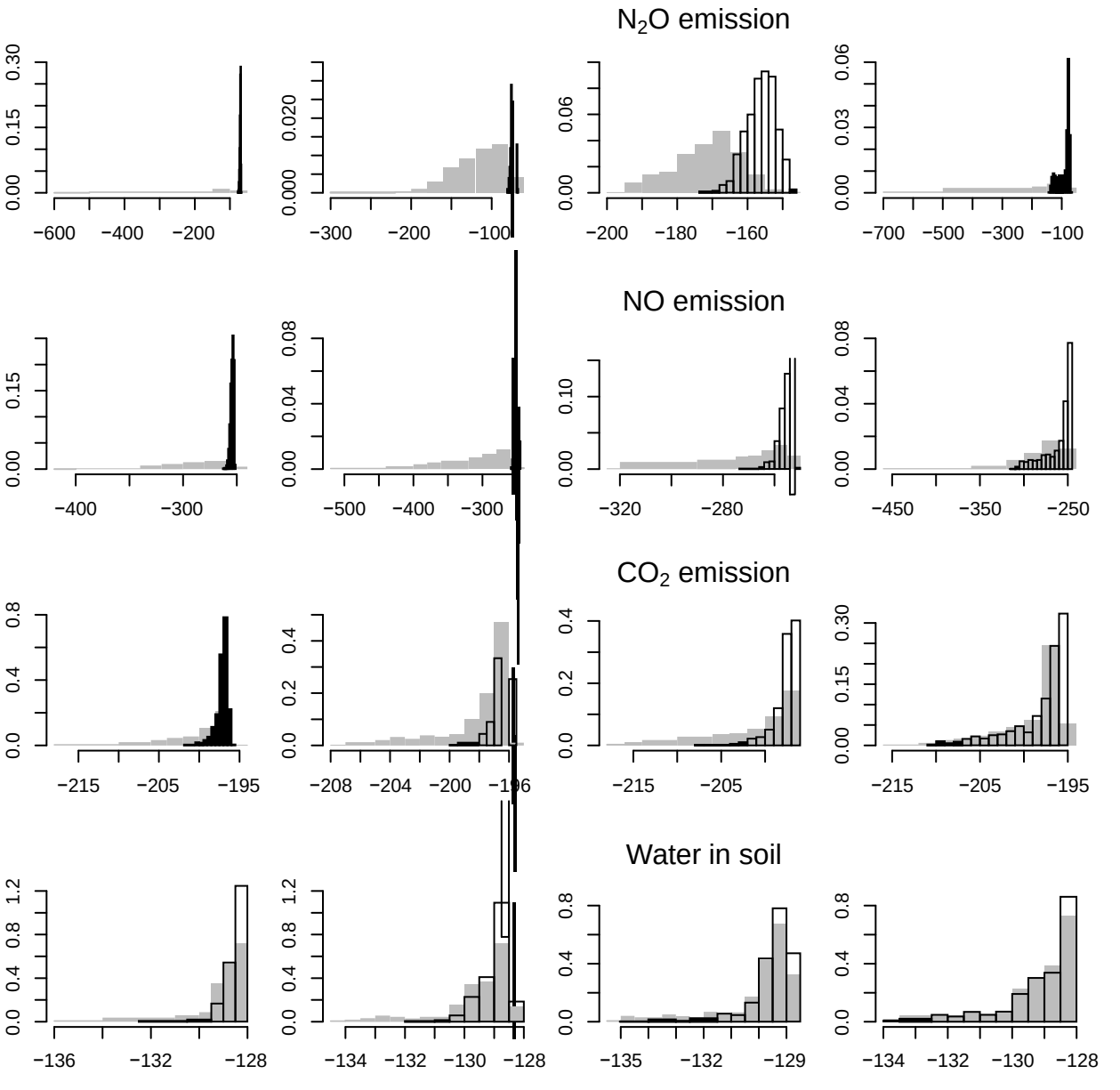
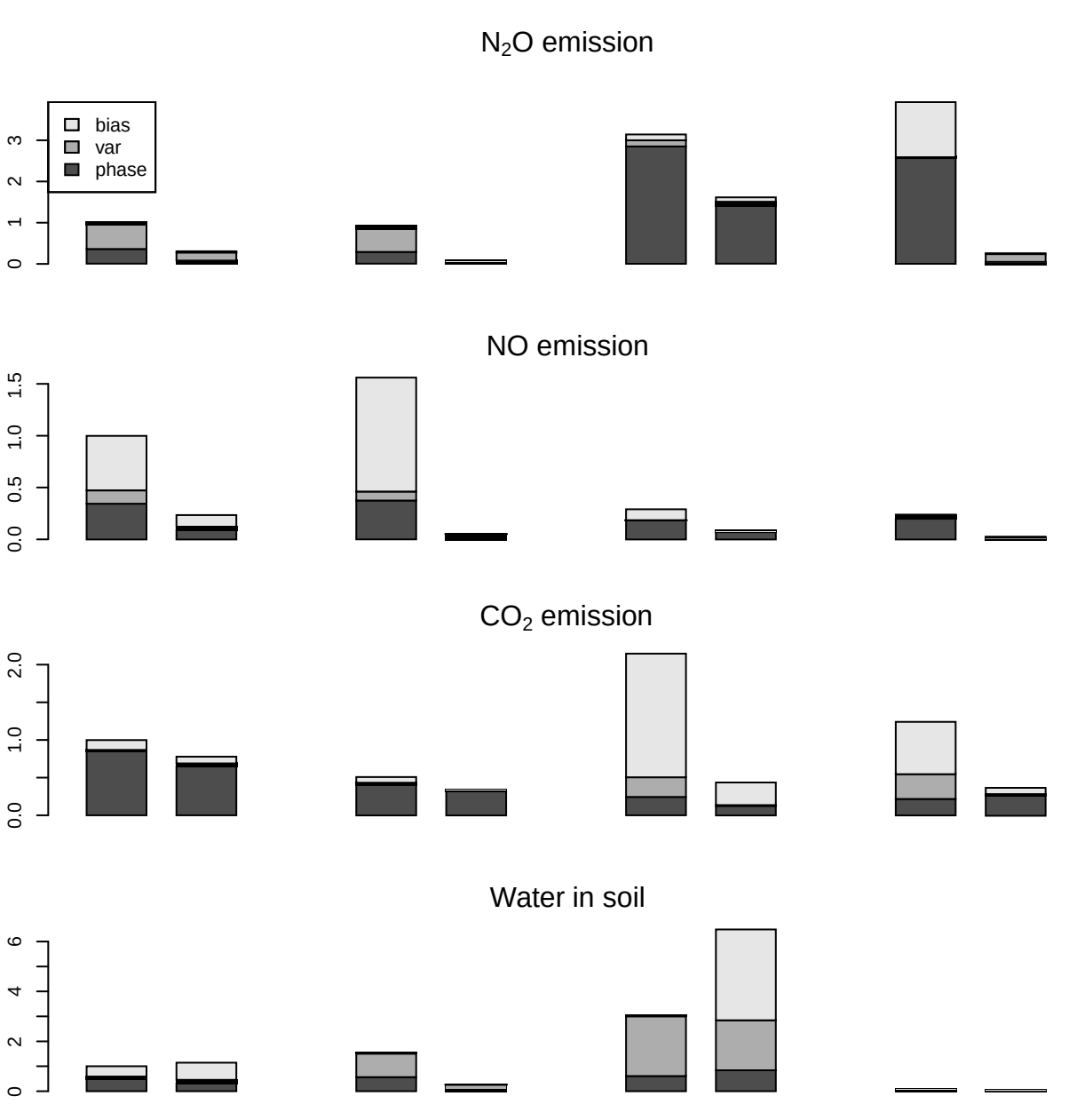

BASFOR
COUP
DAYCENT
MoBiLE
N2O emission
NO emission
CO2 emission
Water in soil

Figure 5


BASFOR
COUP
DAYCENT
MoBiLE
N2O emission
bias
var
phase
0
1
2
3
NO emission
0.0
0.5
1.0
1.5
CO2 emission
0.0
1.0
2.0
Water in soil
0
2
4
6