Article (refereed) - postprint

Van Oijen, M.; Reyer, C.; Bohn, F.J.; Cameron, D.R.; Deckmyn, G.; Flechsig, M.; Härkönen, S.; Hartig, F.; Huth, A.; Kiviste, A.; Lasch, P.; Mäkelä, A.; Mette, T.; Minunno, F.; Rammer, W. 2013. **Bayesian calibration, comparison and averaging of six forest models, using data from Scots pine stands across Europe.**

Bayesian calibration, comparison and averaging of six forest models, using data from Scots pine stands across Europe

M. van Oijen[1,*], C. Reyer[2], F.J. Bohn[3], D.R. Cameron[1], G. Deckmyn[4], M. Flechsig[2], S. Härkönen[5], F. Hartig[3], A. Huth[3], A. Kiviste[6], P. Lasch[2], A. Mäkelä[7], T. Mette[8], F. Minunno[9], W. Rammer[10]

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

[2] Potsdam Institute for Climate Impact Research, Telegrafenberg, P.O. Box 601203, Potsdam, Germany

[3] UFZ – Helmholtz-Centre for Environmental Research, Department of Ecological Modeling, Permoserstr. 15, 04318 Leipzig, Germany

[4] Plant and Vegetation Ecology, University of Antwerp, Universiteitsplein 1, 2610 Wilrijk/Antwerpen, Belgium

[5] Finnish Forest Research Institute, PL 68, FI-80101 Joensuu, Finland

[6] Institute of Forestry and Rural Engineering, Estonian University of Life Sciences, Kreutzwaldi 5, 51014 Tartu, Estonia

[7] Department of Forest Sciences, P.O. Box 27, FI-00014 University of Helsinki, Finland

[8] Forest Growth and Yield Science, Technical University of Munich, 85354 Freising, Germany

[9] Institute of Agronomy, Forest Research Centre, Tapada da Ajuda, 1349-017 Lisbon, Portugal

[10] Institute of Silviculture, University of Natural Resources and Life Sciences (BOKU) Vienna, Austria

* Corresponding author. Email: mvano@ceh.ac.uk

Abstract

Forest management requires prediction of forest growth, but there is no general agreement about which models best predict growth, how to quantify model parameters, and how to assess the uncertainty of model predictions. In this paper, we show how Bayesian calibration (BC), Bayesian model comparison (BMC) and Bayesian model averaging (BMA) can help address these issues.

We used six models, ranging from simple parameter-sparse models to complex process-based models: 3PG, 4C, ANAFORE, BASFOR, BRIDGING and FORMIND. For each model, the initial degree of uncertainty about parameter values was expressed in a prior probability distribution. Inventory data for Scots pine on tree height and diameter, with estimates of measurement uncertainty, were assembled for twelve sites, from four countries: Austria, Belgium, Estonia and Finland. From each country, we used data from two sites of the National Forest Inventories (NFI), and one Permanent Sample Plot (PSP). The models were calibrated using the NFI-data and tested against the PSP-data. Calibration was done both per country and for all countries simultaneously, thus yielding country-specific and generic parameter distributions. We assessed model performance by sampling from prior and posterior distributions and comparing the growth predictions of these samples to the observations at the PSP's.

We found that BC reduced uncertainties strongly in all but the most complex model. Surprisingly, country-specific BC did not lead to clearly better within-country predictions than generic BC. BMC identified the BRIDGING model, which is of intermediate complexity, as the most plausible model before calibration, with 4C taking its place after calibration. In this BMC, model plausibility was quantified as the relative probability of a model being correct given the information in the PSP-data. We discuss how the method of model initialisation affects model performance. Finally, we show how BMA affords a robust way of predicting forest growth that accounts for both parametric and model structural uncertainty.

Keywords

Dynamic modelling; Forest management models; Growth prediction; National Forest Inventories; Permanent Sample Plots; Uncertainty

1. Introduction

Ecological models are built for a variety of purposes. One general motivation is trying to integrate our understanding of the processes underlying natural phenomena. At a time when the earth system is subject to substantial changes in land use and climate, however, it also becomes of increasing importance to be able to make quantitative predictions, supported by a quantification of uncertainty, about the future of our ecosystems.

Forest ecosystems are a prominent example where quantitative predictions are of particular ecological and economic importance, but for which there is considerable uncertainty because different modelling approaches, models and parameters are available (Mäkelä et al., 2012). We focus here on weather-sensitive dynamic models, which simulate the growth of forest stands over time. Dynamic models that have been considered for forest management range from fairly simple, parameter-sparse empirical models to complex models with many parameters (Fontes et al., 2010). None of these models has found widespread application across Europe, which may be due to problems of parameterisation and a lack of knowledge about the generalisability of the models. Given the increasing availability of forest data from National Forest Inventories (NFI) and Permanent Sample Plots (PSP), and other data sources, however, it can be hoped that limitations of dynamic forest models with respect to data availability can be substantially reduced in the future (Hartig et al., 2012). These data can help in parameterisation and evaluation of the models, if we can find robust ways of comparing models and accounting for measurement and modelling uncertainties. In this paper, we use methods based on probability theory, more specifically Bayesian calibration (BC), Bayesian model comparison (BMC) and Bayesian model averaging (BMA), to address these issues. A strength of these methods is that they can be applied to any type of model. Although we do restrict our focus here to dynamic, weather-sensitive models, we have included models of widely differing structure, complexity and data needs, providing a broad practical test of the methods.

Bayesian methods have been used before to calibrate the parameter distributions of dynamic forest models, starting with the work of Green et al. (1999), but application to parameter-rich process-based models is still rare (Luo et al., 2009). The use of BMC to compare and evaluate dynamic forest models – or any other vegetation models – is a more recent application. Van Oijen et al. (2011) included BMC in their analysis of four models for forest biogeochemistry and Fu et al. (2012) used BMC to identify the most plausible models for predicting tree budburst. Here we present, as far as we know, the first applications of BMC and BMA to dynamic forest growth models that include both parameter-sparse semi-empirical models and complex process-based models with many parameters. Using NFI- and PSP-data on Scots pine (*Pinus sylvestris* L.) from four European countries, we compared the results of calibration and testing of these models using

the combined dataset with the results where the same methods were applied to within-country data only. The purpose of this was to assess whether the models would be most effectively calibrated and applied at smaller or larger spatial scales. Similar comparisons of Bayesian approaches applied locally and generically have been made for a simple soil ionic concentration model by Reinds et al. (2008) and for a model of N_2O-emissions in crops by Lehuger et al. (2009).

We ask the following questions:

- How effective are local stand data in reducing uncertainties about forest model parameters in a Bayesian framework?
- Are the considered dynamic models for Scots pine sufficiently general to allow a generic calibration to data from across Europe, or should models be calibrated on a country-by-country basis?
- How effective is Bayesian model comparison in identifying plausible predictive models, and what are the main distinguishing characteristics of forest models that are selected?
- Does Bayesian model averaging lead to improved predictions compared to individually calibrated models?

Although these questions, as well as the models and data used, are focused on forestry in Europe, our methodology is unrestrictedly general. BC, BMC and BMA, and the contrasts made between within- and cross-country applications, can be applied to any other combination of data sets and models in the environmental sciences.

2. Materials and Methods

2.1 *Overview of methodology*

Our study used 6 models and 12 data sets which originated from forest measurements in four European countries (Table 1). The data were from National Forest Inventory (NFI) sites and from sites with Permanent Sample Plots (PSP). From all sites we retrieved environmental data (weather, soil, management) and tree growth data (height, diameter). These data were used by all models to the extent of each model's input data requirements (Table 2). Fig. 1 is a flow chart that shows how the data were used in the consecutive stages of the study. The environmental data from the NFI-sites were used as drivers for model application to those sites. Each model was run multiple times for each NFI-site, to assess the impact of parameter uncertainty on model outputs. We refer to this step as 'prior uncertainty quantification' (prior UQ) because no data of tree growth had been used at this point for improvement of parameter values. The distributions of model outputs generated by this prior UQ were used in a Bayesian model comparison (prior BMC) to quantify the relative plausibility of each model before

calibration. These differences in model plausibility were then used as weights in Bayesian model averaging (BMA), thus producing an averaged prediction to which all six models contributed differently. Next, the NFI-data were used for Bayesian calibration of the parameters of the different models. The calibration was carried out both per country and generically using data from all NFI-sites. The calibrated models were then applied to the PSP-sites using local environmental data. At this stage, we again carried out uncertainty quantification, now termed 'posterior UQ' because the model parameter distributions were already informed by the NFI-data. Finally, the results from the posterior UQ were compared with measurements from the PSP-sites for a posterior Bayesian model comparison, again accompanied by BMA. In the rest of this section, we describe data, models and statistical methods in more detail.

[FIG. 1 HERE]

2.2 *Data*

Data of twelve even-aged *P. sylvestris* stands were assembled from four European countries (Table 1). From each country, two NFI sites and one PSP-site were selected for this study. An exception was Estonia, for which NFI-data were not available and three PSPs were used. For ease of reference, we used a site-code for each site consisting of the first letter of the country's name, followed by 1 or 2 for the NFI-sites and 3 for the PSP-site (Table 1), except for Estonia where the numbers refer to the three PSPs. For model calibration, we only used data from the sites coded 1 or 2, whereas for model comparison and averaging the data from sites with code number 3 were used. The data used were for mean tree height and stem diameter at 1.3 m above ground, which were available from all sites. Data on stem number and tree age were used as uncalibrated inputs. All sites provided several measurements for the different variables (between 2 and 7), separated by intervals of at least 5 years (Fig. 2). We now briefly describe the sites in each country.

[FIG. 2 HERE]

2.2.1 Austria

The NFI-plots A1 and A2 are part of the Austrian Forest Inventory grid consisting of ~10000 points. The plots are 100% *P. sylvestris* and the soils are classified as Semipodsol and Cambisol with soil depths exceeding 0.3 m and field capacity around

36%. They are located at different altitudes in the "Waldviertel", a region in Lower Austria north of the Danube. A1 lies about 300 m higher than A2 and is cooler and drier. On both sites, measurements were taken in two years (1987 & 2000 and 1989 & 2002). The sample consisted for each plot of a combined angle count measurement (for trees > 10.5 cm diameter) and a circle with a fixed radius (for trees < 10.5 cm). Height measurements were done for a subset of trees of the angle count measurement; the other heights were calculated. Nothing is known about management history or planting time, except that no management occurred during the period of measurements.

The selected PSP-site, A3, was established in 1970 and measured every five years. The site is maintained by the Austrian Federal Forest Office BWF (http://bfw.ac.at/) and is located near A2 with similar soil properties. It is a pure *P. sylvestris* stand with a size of 1500 m² and a stem number of 790 ha^{-1} in 1980.

Climate data for the NFI- and PSP -sites were provided from nearby weather stations of the Austrian weather service ZAMG (Central Institute for Meteorology and Geodynamics).

All three stands reached heights of about 18 m at an age of about 60 years. However, they differ significantly in diameter (207-324 mm), with lower values at high stem number.

2.2.2 Belgium

The Belgian plots B1 and B2 are NFI's of the ANB (Agentschap Natuur en Bos, 'Forest and Nature Agency'), situated in the Campine region of north-eastern Belgium, were established in 1937 and 1942 respectively and regularly thinned since then from the original 12500 trees ha^{-1}. B1 is situated on loamy sand, and data from 2000 and 2004 were available; thinning during this period reduced stem number from 400 to 380 ha^{-1}. B2 is situated on sandy soil close to B1 and data from 2000 and 2008 were available. Thinning during this period reduced stem number from 520 to 393 ha^{-1}. The data were obtained from 40 x 25 m sample plots.

The PSP-site, B3, "De Inslag", is a mixed patchy coniferous/deciduous forest located in Brasschaat also in the Belgian Campine region. The site is part of the European Carboeurope-IP network and is a level-II observation plot of the European network program (ICP-II forests) for intensive monitoring of forest ecosystems (EC-UN/ECE, 1996), managed by the Flemish Research Institute for Nature and Forest (INBO). Here we only focus on one particular even-aged Scots pine stand planted in 1929 and described by Curiel Yuste et al. (2005). In this experimental stand, stem number was 556 ha^{-1} in 1997. In November 1999, a thinning was performed reducing

the stem number to 377 ha^{-1} and further thinned to 362 ha^{-1} in 2002. The soil is loamy sand, moderately wet, with a distinct humus and iron B-horizon (Baeyens et al., 1993) and is classified as Umbric Regosol. Although the Belgian plots are on relatively sandy soils, soil water table is quite high (0.7-1.1 m) and soil fertility is high due to high nitrogen deposition (30-40 kg N ha^{-1} $year^{-1}$).

Despite similar age (66-67 years) and stem number (380-390 ha^{-1}), the two NFI-plots had quite different heights (18.4, 23.2 m) and diameter (271, 293 mm) indicating differences in site quality. The PSP-site was older and had lower tree number; height was intermediate but diameter was greater than at the NFI-plots.

2.2.3 Estonia

The Estonian plots E1, E2 and E3 belong to the Estonian Forest Research Plots Network which consists of more than 700 PSP and are maintained by the Estonian University of Life Sciences (Sims et al., 2009). These plots were established at the observation sites of the European network programme ICP Forest Level I plots. The plots, established in 2000, are circular with radii of 25, 20 and 25 meter, respectively and were re-measured in 2005 and 2010. The plots have not been thinned during that period, but earlier management history is unknown. On each plot, the diameter at breast height was assessed for each tree. Tree height and height to crown base were measured in every fifth tree. All three plots are dominated by Scots pine (more than 90% of total volume), but there is a small mixture of Silver birch (*Betula pendula*) and Norway spruce (*Picea abies*). The plots are located in southern Estonia where mean effective temperature sum is about 1650 degree days. The plots are on sandy soils on glaciofluvial deposits with sufficient water availability belonging to WRB 2006 soil units Gleyic Podzol, Histic Podzol and Albic Podzol respecticely. The vegetation types of the plots are Rhodococcum, drained Polytrichum-Nyrtillus, and Rhodococcum. The basal area of the plots reached 24.8, 33.7, and 31.8 m^2 ha^{-1} at stand ages 70, 67, and 73 years, with average heights of 25.2, 24.7, and 25.6 m and volumes of 285, 384, and 374 m^3 ha^{-1}. Differences in diameter (237-274 mm) were larger than height differences, with largest values reached at the lowest stem number.

2.2.4 Finland

The Finnish plots F1 and F2 are permanent NFI sample plots located in Southern Finland established by the Finnish Forest Research Institute. They have been measured in 1985 and 1995. The plots have not been thinned during that period. The earlier treatment history is unknown. The plot size varied according to the stem diameter at breast height, being 100 m^2 when the diameter was under 10.5 cm, and otherwise 300 m^2. The trees with diameter smaller than 4.5 cm were measured only if

they were expected to survive until the next measuring date. Diameter at breast height and tree species were recorded from all the tally trees. Heights, crown base heights and crown widths were measured from the sample trees, which include the trees that were located in a circular area around the sample plot mid-point, where the circle radius is half of the original sample plot radius.

The Finnish plot F3 is a control plot with no thinnings in a permanent thinning experiment of the Forest Research Institute at Vesijako in southern Finland. The experiment was established in 1948 in a pine stand sown in 1918, and it was followed until 1997. The site is fairly fertile with adequate moisture for pine. The plot has a small mixture of birch (*Betula* spp.), less than 10% of basal area. Plot size was 1000 m^2, and all trees were numbered on this plot and measured for breast height diameter in a total of seven measurements. For height (and crown base height in the two most recent measurements), 21-67 trees were chosen as sample trees. The final heights of 17.8 m (75 yrs, NFI 1), 10.1 m (55 yrs, NFI 2) and 21.8 m (79 yrs, PSP) indicate that despite the age difference, the site conditions at NFI 2 were probably less favourable (cf. Fig. 2a). The comparatively low stem number and the high diameter, and the fact that no mortality occurred, suggest that the NFI plots were thinned at some point before the surveys. In contrast, at the PSP-site only self-thinning occurred leading to high stem numbers and low diameters.

2.3 *Models*

We used six different forest models in the assessment, ranging from simple semi-empirical models to parameter-rich process-based models (Table 2). All models are able to predict mean tree height and mean stem diameter. Some of the models are able to simulate variation between individual trees as well, but the corresponding predictions were not tested against data. Four of the models are initialised at the first measurement date, i.e. they require the earliest observed values of mean tree height and/or diameter to quantify the model's initial constants (Table 2). This reduces the number of data available for Bayesian calibration. The remaining two models, 3PG and BASFOR, include state variables that are difficult to estimate from mean height and stem diameter only, such as nitrogen pools in soil and trees, and it was therefore decided to initialise them from planting. These two models therefore have more data available for calibration, but their predictions of forest growth may already start deviating from observations before the first measurement date. We shall now briefly describe each model, referring to earlier publications for more detail. Each model description finishes with an account of how the prior probability distribution for the model's parameters was set by the

respective modellers. The role of these probability distributions in uncertainty quantification and Bayesian calibration is explained in §§ 2.4-5.

2.3.1 3PG

3PG calculates the dynamics of biomass in different organs (foliage, roots and stem) and simulates the soil water balance and variables of interest to forest managers, such as stand timber volume, mean diameter at breast height, stand basal area and mean annual growth increment. Gross primary production (GPP) is calculated by multiplying photosynthetically active radiation absorbed by the stand with a light-use efficiency that changes with environmental conditions. Light absorption is calculated using Beer's law, while the light-use efficiency varies in dependence of atmospheric vapour pressure deficit, air temperature, the presence of frost, soil water balance, tree age and site fertility. Net primary productivity (NPP) is calculated as a constant fraction of GPP (Law et al., 2000; Waring et al., 1998). Carbon allocation is based on allometric equations, applied on a single-tree basis. The fraction of NPP allocated below-ground decreases with soil fertility. Site fertility is expressed through a site specific reduction factor (FR) that varies between 0 (for the least fertile sites) and 1 (for sites that do not have nutrient limitations). The remaining NPP is partitioned between the aboveground organs as a function of stem diameter at breast height. The diameter at breast height and the average stand height are calculated through allometric functions of average aboveground biomass per tree. 3PG has been applied to various different species and sites and is widely used in research as well as by companies to assess forest growth and site productivity. Detailed descriptions of 3PG were provided by Landsberg and Waring (1997) and Sands and Landsberg (2002).

Before this study, Landsberg et al. (2005) tested the performance of 3PG for Scots pine in Finland, using a modified carbon allocation routine. Xenakis et al. (2008) coupled 3PG with ICBM/2N (Introductory Carbon Balance Model (Andren and Katterer, 1997)) a soil matter decomposition model. The new model, 3PGN, was calibrated and tested for Scots pine plantations in Scotland. The information from these two previous studies was utilised to construct the prior, using truncated Gaussian distributions. For each parameter, the prior mean was set to the average of the values used in Landsberg et al. (2005) and Xenakis et al. (2008). The bounds of the prior were set at ±30% of the mean value. The site fertility parameters were also included in the BCs and BMCs; the FRs ranged between 0 and 1, while the prior mean was 0.5. For all parameters, the prior was kept quite uninformative (i.e. high variance and wide ranges), reflecting the fact that the 3PG-modeller in the current study did not have previous experience with Scots pine.

2.3.2 4C

The forest model 4C (FORESEE –FORESt Ecosystems in a changing Environment) has been developed to simulate the impact of changing environmental conditions on forest ecosystems. It is climate sensitive and calculates physiological processes on the tree and stand level depending on the process in question in daily to yearly time steps (Bugmann et al., 1997; Suckow et al., 2001). Establishment, growth and mortality of tree cohorts are explicitly modelled at the patch scale on which horizontal homogeneity is assumed. Cohorts of trees compete for light, water and nutrients (Bugmann et al., 1997). Every cohort develops specific values for fine root, foliage, stem biomass etc. and species-specific parameters steer the physiological processes for each species. Photosynthetic rate is calculated after Haxeltine & Prentice (1996) and a constant fraction of GPP is lost to respiration (Landsberg & Waring 1997). The resulting NPP thus depends on environmental conditions and is allocated according to the principles of the pipe model (Shinozaki et al. 1964) and of the functional balance (Davidson 1969) and organ-specific, constant senescence rates. In this allocation model, height growth is decoupled from diameter growth, with high degrees of intra-canopy shading leading to extra height growth. Nitrogen limitation has been calculated dynamically. When the tree water demand of a cohort exceeds the plant available water in the soil, the canopy conductance and ultimately NPP of that cohort is reduced. 4C requires daily meteorological variables, a soil description including physical and chemical parameters as well as a forest stand description.For further details of model processes and recent model applications, see Suckow et al. (2001), Lasch et al. (2005), Seidl et al. (2008) and Reyer et al. (2010).

The prior distribution for all parameters of 4C was uniform with boundaries at $\pm$50% of the initial (standard 4C) value, reflecting large uncertainty about parameter values. The selection of the parameters to be calibrated was restricted to species-specific parameters that could be informed by Scots Pine data, giving a total of 43 parameters amenable to calibration.

2.3.3 ANAFORE

ANAFORE (ANAlysing FORest Ecosystems) is a stand-scale, mechanistic forest model that dynamically simulates the fluxes of carbon, water and nitrogen through the ecosystem (Deckmyn et al., 2008). The forest stand is described as consisting of trees of different size cohorts (e.g. dominant, co-dominant and suppressed trees), either of the same or of different species (deciduous or coniferous). Half-hourly carbon and water fluxes are modelled at the leaf, tree and stand level from half-hourly, daily or monthly climate data. In addition to total growth and yield, the model

simulates allocation changes in crown size, DBH-height ratio, root-shoot ratio and even the daily evolution of tracheid or vessel biomass and radius, parenchyma and branch development. From these data, early and late wood biomass, wood tissue composition and density are calculated to allow wood quality estimation. Simulation of the labile carbon stored in the living tissues allows for simulation of trans-seasonal and trans-yearly effects, and simulation of the long-term effects of environmental stresses on growth. A detailed soil model including fungal, bacterial and mycorrhizal effects on SOM degradation and aggregate formation is included (Deckmyn et al., 2009). Model initialisation was at the first measuring point. Because ANAFORE needs a detailed tree description – not available for most sites - allocation as observed at the Belgian sites was used throughout (% heartwood, branch biomass, crown length). Crown width was set to fill the site.

The prior distribution for the parameters was uniform with boundaries at ±10% of the initial value, reflecting measured data (mainly on the Belgian Brasschaat site) and data from literature as described in Deckmyn et al. (2008). Although ANAFORE was calibrated for Scots pine before this study, this was only for Belgian stands and the uncertainty concerning parameterisation across Europe is large, so the same prior was used.

2.3.4 BASFOR

The BASic FORest simulator, BASFOR, is a deterministic daily time step forest model used for simulating coniferous or deciduous forests. The model simulates carbon and nitrogen cycling in trees, soil organic matter and litter. It simulates the response of trees and soil to radiation, temperature, precipitation, humidity, wind speed, atmospheric CO_2 and N-deposition, and thinning regime. The model has 14 state variables, representing carbon and nitrogen pools in trees and soil, and 48 parameters which include the initial constants of the state variables. Besides time series for the state variables, output may be produced of NPP, tree height, stem diameter, ground cover, LAI, N-mineralisation and other tree and soil variables. BASFOR is built from well known process representations. Light absorption is calculated by Beer's law. GPP is calculated as light absorption times a light-use efficiency (LUE). NPP is calculated as a fixed ratio of GPP. LUE is temperature-, CO_2- and water-dependent and may be reduced if insufficient nitrogen is taken up by the plants. Potential nitrogen uptake scales with root system surface area. Actual nitrogen uptake is the minimum of demand, determined by tissue N-concentration, and potential uptake. Allocation of assimilates follows allometric rules, but water

stress may limit leaf area index (LAI). Turnover of tree and soil components proceeds at temperature-dependent relative rates.

The model structure was described by Van Oijen et al. (2005), more recent model applications are reported by Van Oijen & Thomson (2010) and Van Oijen et al. (2011), and the model is now also in use as the tree component of an agroforestry model (Van Oijen et al., 2010). The prior for BASFOR was constructed from beta-distributions for the individual parameters, with ranges and modes based on literature as described before (Levy et al. 2004; Van Oijen et al. 2005, 2011).

2.3.5 BRIDGING

The BRIDGING model (Valentine and Mäkelä, 2005) was developed to bridge the gap between process-based and empirical approaches to modelling tree growth by formulating a process-based model that can be fitted and applied in an empirical mode. Tree growth in the model is based on carbon balance, and its allocation is consistent with pipe model theory and an optimal control model of crown development (Mäkelä and Sievanen, 1992). These provide a framework for expressing the components of tree biomass in terms of tree height, crown height and stem cross-sectional area, the growth of which is regulated by photosynthesis and respiration. The parameters of the model comprise physiological rates and morphological ratios and can be estimated from lower-level process models or direct measurements. In the empirical mode, the original parameters are combined into a set of fewer, aggregate parameters which can be estimated from inventory type data using statistical procedures. Here, we calculate the photosynthesis and respiration parameters from lower-level models of stand productivity (Mäkelä et al., 2008) and canopy structure (Duursma and Mäkelä, 2007) using a procedure proposed by Härkönen et al. (2010). The productivity model is driven by daily data of global radiation, vapour pressure deficit and air temperature, while field data on inventory variables (stand-level mean values of height, diameter, crown base height and crown width, stocking density or basal area, and site fertility) are used for parameterising canopy structure. These parameters are given fixed, deterministic values. The parameters related to growth of tree height and basal area are employed in their aggregate form and estimated using the Bayesian approach with the given inventory data.

The Bridging model has 38 different parameters, of which the 13 parameters relating to the dynamic growth of tree height and basal area were used in the calibration. Uniform distributions were used throughout. Parameters left out of the calibration included structural relationships, which were calculated directly based on

the measured stand data, biomass estimates, and light-use efficiency estimates. The uniform distributions were mainly quantified based on earlier pipe model studies (Mäkelä 1997, Mäkelä and Vanninen 2001, Vanninen and Mäkelä 2005, Valentine and Mäkelä 2005, Palmroth et al. 1999, Duursma and Mäkelä 2007).

2.3.6 FORMIND

FORMIND is an individual-based, spatially semi-explicit gap-type model (Köhler and Huth, 1998; Ruger et al., 2007). Spatially semi-explicit means that the modelled plot (in this case 1 ha) is divided into 20 x 20 m gaps. Tree individuals are assigned to one of these gaps, but do not have an explicit position within gaps. As in classical gap models, tree crowns are assumed to cover the gap uniformly in horizontal direction at a certain height, depending on the size of the trees. The vertical stratification through the different crown heights of the trees and the differences in light climate that result from that for each individual tree are important determinants of the predicted community dynamics. NPP is calculated as the difference between GPP and respiration. GPP of each individual tree depends on the available light at crown top, temperature and soil water content. The temperature dependence follows a hump shape. A reduction due to insufficient soil water occurs below a threshold and GPP is completely reduced if soil water content falls below the permanent wilting point. Additionally, maintenance respiration has a temperature dependence following the Q10-approach (Gutiérrez and Huth, 2012). The model was initialised for each site at the first recorded year with the observed number of trees, all of the same observed average diameter, randomly distributed over the modelled area of one hectare.

The marginal prior probability distributions for FORMIND were all uniform. Parameters were excluded from the calibration that were either unrelated to those model outputs that were compared to calibration data, or for which there were other parameters already under calibration that acted on the model outputs in a similar way. Based on this premise, four parameters were selected for calibration. These included the two parameters that determine the diameter-height relationship, the main growth parameter that determines the maximum growth rate under full light, and the wilting point, which is the determinant of how strongly the plants react to water stress. The other parameters were fixed according to literature data. For each of the calibration parameters, flat and relatively wide priors were chosen reflecting large uncertainty about parameter values.

2.4 *Uncertainty quantification (UQ)*

Predictive uncertainty (i.e. uncertainty regarding model outputs) was quantified for each model at three stages in our study: before any parameter calibration had been carried out (prior UQ), and after country-specific and generic calibration (posterior UQ) (Fig. 1). In each case, the UQ consisted of running the model 1001 times, using a sample of that length from the parameter distribution for the model.

For each model, the prior parameter uncertainty - before any of the NFI- or PSP-data had been used for calibration – was expressed in the form of a probability distribution. This was done by each modelling group separately, no standardisation of priors being attempted (see §2.3). To derive from that the prior predictive uncertainty, we used a sample consisting of the mode of this parameter distribution plus 1000 other parameter vectors sampled from the prior distribution using Latin Hypercube Sampling to ensure good coverage of parameter space. This prior UQ was carried out for all 12 sites.

To assess the posterior predictive uncertainty, i.e. the uncertainty resulting from the reduced parameter uncertainty after country-specific or generic Bayesian calibration (see below), we used the mode of the posterior parameter distribution, i.e. the Maximum A Posteriori (MAP) parameter vector, and again 1000 other parameter vectors that were selected by equidistant subsampling from the parameter chains generated in the calibration. Posterior UQ was carried out only for PSP-sites because the data from those sites had not been used in the calibration.

2.5 *Bayesian calibration (BC)*

Bayesian calibration was carried out as documented in other recent forest model studies (Van Oijen et al., 2011; Van Oijen et al., 2005) and we shall give only a brief outline here. The method starts by expressing uncertainty about the model's parameter values in a so-called prior parameter distribution, $P(\theta)$. In this notation, θ represents the full parameter vector of a model, so $P(\theta)$ is a multivariate distribution. All modellers in this study assigned prior distributions without any correlations between different parameters, so $P(\theta)$ could be written as the product of independent distributions for the individual parameters. By comparing model predictions with NFI-data, D, we can derive a likelihood value $P(D|\theta)$ for each possible parameter value (see below), which can be interpreted as a relative "goodness-of-fit" measure for this parameter (Hartig et al., 2012). Bayes' formula then allows us to combine both pieces of information (prior and likelihood) into one posterior parameter distribution. The formula states that

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

i.e. that posterior probability is proportional to prior times likelihood P($D|\theta$). To derive a likelihood function, we made the assumption, for all models and measurements, that measurement errors were normally distributed with a coefficient of variation of 20%. The fairly high value of 20% was chosen to account for multiple factors affecting the measurements, including instrument error, demographic stochasticity of the tree populations, and environmental heterogeneity. No correlations between measurement errors were assumed, so our likelihood function could be written as the product of independent Gaussian functions of the difference between data D and model output $M(\theta)$:

$$\mathrm{P}(D|\theta) = \text{Probability of measurement error equal to } D\text{-}M(\theta)$$

$$= \prod_{i=1}^{n} \varphi(D_i - M_i(\theta); 0, (0.2D_i)^2),$$

where the i-subscripts index the n data points and the corresponding model outputs, and where φ denotes a Gaussian probability density function with given mean and variance.

To estimate the posterior distributions, we used a Markov Chain Monte Carlo (MCMC) algorithm (Metropolis et al., 1953; Van Oijen et al., 2005). Convergence of the MCMC was verified both visually – by inspection of the parameter trace plots – and by calculation of the Gelman-Rubin statistic (Gelman and Rubin, 1992).

2.6 *Bayesian model comparison (BMC) and calculation of NRMSE*

Bayesian model comparison relies on the same probabilistic ideas as BC, but now the probability distribution to be informed by the data is not that for the parameters but for the models themselves (Kass and Raftery, 1995). A key strength of BMC is that it evaluates models not at one single parameter vector value but takes into account parameter uncertainty (Tuomi et al., 2008). The formal need for this coverage of parameter uncertainty is seen when we write out Bayes' Theorem as applied to model comparison:

$$\mathrm{P}(M|D) \propto \mathrm{P}(M)\ \mathrm{P}(D|M),$$

where, following the law of total probability:

$$\mathrm{P}(D|M) = \int P(D \mid M(\theta))\ \mathrm{P}(\theta)\ \mathrm{d}\vartheta.$$

So each model's parameter uncertainty, and not only the best value, determines how much support a model receives. Among other things, this provides a natural safeguard against

overfitting using overly flexible models. P($D|M$) is referred to as the 'integrated likelihood', or also the 'marginal likelihood' as it is calculated by marginalizing out the uncertain influence of the model's parameters. We assumed that each model had the same prior probability of 1/6 before any data were used. Application of the models to the NFI-sites, in the prior UQ, provided 1000 model results which were used to derive each model's integrated likelihood for those data. The posterior probability for each model was then calculated as the model's integrated likelihood divided by the sum of the integrated likelihoods for all models (Kass and Raftery, 1995). A similar procedure was applied at the next applications of BMC, where the integrated likelihoods of the models were calculated for the PSP-data after the models had been calibrated on the NFI-data. These posterior BMC's were carried out after both country-specific and generic BC.

Additionally, we calculated a standard goodness-of-fit measure, the normalised root mean squared error (NRMSE), for model predictions at PSP-sites. This was done for both the prior and posterior parameter distributions. In contrast to the calculation of the integrated likelihood, the NRMSE had to be calculated separately for height and diameter as its calculation involves a normalisation by the average of the measurements:

$$\text{NRMSE} = \frac{1}{\overline{D}}\sqrt{\frac{1}{n_c \times 1000}\sum_{c=1}^{n_c}\sum_{\theta=1}^{1000}(M_c(\theta) - D_c)^2}$$

Where n_c is the number of countries from which PSP-data were used, D_c are the measured values, $\overline{D}$ is the average of the measurements across the n_c countries, θ indexes the 1000 parameter vectors sampled from prior or posterior distribution and $M_c(\theta)$ is model prediction for country c using parameter vector θ. In the case of the prior and generic posterior parameter distribution, the calculation of NRMSE uses n_c = 4, but in the case of country-specific posteriors, NRMSE is calculated first per country (n_c = 1) followed by averaging of the four errors to arrive at an estimate of overall NRMSE.

2.7 *Bayesian model averaging (BMA)*

Bayesian model averaging uses the different model probabilities P(M), derived in preceding BMC, to calculate a weighted probability distribution for model outputs (Hoeting et al., 1999; Kass and Raftery, 1995):

$$\text{P}(y) = \sum_{m=1}^{6} P(M^{(m)})P(y \mid M^{(m)})$$

Where P(y) is the averaged output distribution, P($M^{(m)}$) is the probability for model m as derived from the BMC, and P ($y|M^{(m)}$) is the output distribution for model $M^{(m)}$. Expanding the last term gives:

$$P(y) = \sum_{m=1}^{6} P(M^{(m)}) \int P(y \mid M^{(m)}, \theta^{(m)}) P(\theta^{(m)}) d\theta^{(m)},$$

which shows that the BMA accounts for both overall model structural uncertainty, P($M^{(m)}$), and each individual model's parameter uncertainty, P($\theta^{(m)}$). In this study, BMA was applied after both prior and posterior BMC, with P($\theta^{(m)}$) representing prior and posterior parameter uncertainty, respectively. The same model output samples used in BMC were used for BMA as well, but subsampled with sample size proportional to P($M^{(m)}$). The BMA-forecasts thus produced were compared against the measurements at the PSP-sites. Note that in this procedure only the prior BMA was subjected to a fully out-of-sample test of predictive capacity of the model averaging.

[FIG. 3 HERE]

3. Results

3.1 *Uncertainty quantification before and after Bayesian calibration*

The first quantity calculated was the prior predictive uncertainty, that is, the model uncertainty before any data were used for calibration. Table 3 shows summary statistics of the prior predictive distributions for the NFI-sites: the value of mode of the prior plus the 5% and 95% quantiles. Figs 3 and 4 depict the ranges between the 5% and 95% quantiles for the PSP-sites. The prior output ranges – delimited by the 5% and 95% quantiles – were generally widest for the three most parameter-rich models, i.e. ANAFORE, BASFOR and 3PG.

Bayesian calibration (BC) was carried out both per individual country and generically, so samples from five different posterior parameter distributions were produced for each model. Our results show that generic Bayesian calibration reduced parameter uncertainty in all models except ANAFORE, with average reductions in the standard deviation of marginal parameter distributions (i.e. for individual parameters) ranging from 1 to 13%. These averages were invariably the result of a majority of parameters being hardly affected by the BC and a small number with strongly reduced uncertainty, with maximum

reductions in standard deviation for individual parameters ranging from 6 to 83% across all models (data not shown). The results of country-specific BC were similar but with generally lower reductions in uncertainty.

Figures 3 and 4 show predictive uncertainty after calibration for mean height and diameter. With respect to output uncertainty, measured as the distance between the 5% and 95% quantiles, the results for country-specific and generic BC were quite similar (Table 3; Figs 3, 4). BC reduced tree height uncertainty in all models, but most in 3PG and BASFOR and least in BRIDGING. For stem diameter, 3PG and BASFOR again saw large uncertainty reductions but otherwise the results differed markedly from those for tree height, with ANAFORE and BRIDGING seeing no clear reductions in predictive uncertainty and FORMIND even becoming worse at B3, E3 and F3.

[FIG. 4 HERE]

3.2 *Bayesian model comparison before and after calibration*

The predictions of the uncalibrated models for the NFI-sites, generated as part of the prior UQ reported in the previous paragraph, were compared against the corresponding NFI-data in a prior Bayesian model comparison (BMC) (Fig. 5). Despite the fact that the data tended to fall between the 5% and 95% quantiles of each model's prior uncertainty ranges (Table 3), the Bayesian model comparison still assigned very different prior probabilities to the different models. The most parameter-rich model, ANAFORE, and the two models initialised at planting, 3PG and BASFOR, had prior probabilities orders of magnitude lower than the other three models. BRIDGING and, to slightly lesser extent, 4C achieved the highest integrated likelihoods (Fig. 5).

The posterior BMC, in which models outputs after calibration were compared with measurements at PSP-sites, showed smaller differences between model probabilities and slightly altered the ranking of the models (Fig. 5). The posterior BMC assigned the highest probability to 4C, followed by BRIDGING and FORMIND with 3PG thereafter.

Similar ranking can be observed in the values of NRMSE (Fig. 6), which like the integrated likelihoods of the models were calculated as averages for the whole parameter distribution. For all models except ANAFORE, the values of NRMSE for mean height and diameter were markedly reduced by BC but with little difference between country-specific and generic BC.

[FIG. 5 HERE]

[FIG. 6 HERE]

3.2 *Bayesian model averaging before and after calibration*

The weighted average predictions of the models for the PSP-sites, using prior and posterior model probabilities as weights, are included in Figures 3, 4 and 6. The prior BMA, which was based on model probabilities derived from NFI-data without any model calibration, showed robust out-of-sample predictive capacity for the PSP-sites, as shown by low NRMSE-values for both output variables (Fig. 6). In the case of tree height, only the BRIDGING model had lower NRMSE, whereas for stem diameter only 4C had clearly lower error. Also, predictive uncertainty from the prior BMA was moderate, with at least half of the models showing larger uncertainty ranges for all combinations of variable and site except stem diameter at F3.

Predictions from posterior BMA were also compared against the measurements at PSP-sites (Figs 3, 4, 6). In contrast to the tests of prior BMA, and despite the fact that only NFI-data were used in model calibration, these were in-sample tests of predictive capacity because PSP-data had been used to calculate the model probabilities. Prediction using posterior BMA was less of an improvement compared to most individual models than was the case for prior BMA (Figs 3, 4, 6).

4. Discussion

4.1 *Model performance before and after Bayesian calibration on NFI-data*

If forest models are to be useful in management, their predictions must be sufficiently accurate and precise. A quantification of model accuracy for growth is given in Table 3, where the predictions for the modes of prior parameter distributions can be compared against measurements. The same table also provides information about predictive uncertainty, in the form of the 5% and 95% quantiles of model predictions. The results show that only the BRIDGING model had high *a priori* predictive accuracy for mean tree height with low accompanying uncertainty at all sites except F3. For stem diameter, none of the uncalibrated models was very precise – BRIDGING, 4C and FORMIND did best – and only BRIDGING and FORMIND had low uncertainties throughout. The balance of accuracy and precision for

the NFI-sites was such that the prior Bayesian model comparison assigned 55% prior probability to BRIDGING and 42% to 4C.

One reason for the prior success of BRIDGING and 4C, and to lesser extent FORMIND, was that these models were initialised for each site at the first date of measurement. The models were thus started off with values of mean tree height and stem diameter correct for the site, and with fewer years of growth remaining to be predicted than what was asked from models initialised at planting, such as 3PG and BASFOR. The advantage of late model initialisation – having less time to deviate from true on-site growth patterns - apparently weighed heavier than that of 3PG and BASFOR being able to process more detailed information about the site conditions. Furthermore, information about the early management history of sites, such as the tree thinning regime, tends to be less reliable than information for the measurement periods. Late initialisation, however, does not always improve predictive performance, as demonstrated by the results for ANAFORE. In the case of ANAFORE, a highly detailed model, there was a large suite of other state variables besides mean height and diameter that needed to be initialised, and for which no good information was available for most sites so default model settings could not be adjusted. While some models may be designed to run with stand-level information such as typically provided by NFIs, other models may perform better if more detailed initialisation data are available. In this study, the most complex model, ANAFORE was clearly overparameterized in relation to the very limited data. We also note that BRIDGING and 4C might have been rated best if initialisation values would have been estimated rather than being set a priori – but that was not investigated in this study.

These comparisons of the prior performance of the different models were inevitably also affected by how the prior parameter distributions were defined. Different methods for quantifying prior parameter distribution of a process-based forest model, PnET-II, were discussed by Radtke et al. (2001). The prior distributions in our study were set independently by each modelling group, using the information available to them from literature and from previous experience with their model. This partly explains why some models, such as 3PG, showed wider prior output ranges than other models.

To restrict the influence of subjective prior parameterisation, it is therefore important to compare differences in model performance after all models have been calibrated for the tree species under study. Both country-specific and generic Bayesian calibration on NFI-data markedly increased the accuracy and precision of prediction for the PSP-sites by all models except the most complex and parameter-rich model, ANAFORE (Figs 3, 4). After these general improvements, the 4C model performed best (Fig. 5), but note that the differences in model initialisation method again affected the results, and that the strength of the data was

probably still not sufficient to completely overrule the effect of prior choice after calibration. Also note that the assessments of model performance and plausibility in this study are restricted to predictions for mean tree height and stem diameter. If data from other variables, such as above-and belowground biomass and wood quality, had been used, model evaluation would likely have yielded different results.

4.2 *Spatial differences in model performance*

All models had the poorest predictions of mean tree height for the Finnish PSP-site. That site, F3, had an atypically high stem number (Table 1), which may have contributed to comparatively strong height growth at relatively small diameter despite advanced age (Fig. 2). Most models apparently struggled to simulate this growth pattern, irrespective of model complexity. The problems with this site largely persisted after calibration.

Sites within a single country are likely to be more similar in tree provenance, soil type and climate than sites in different parts of Europe. Therefore, the performance of models at a given PSP-site was expected to be best after calibration exclusively on the two NFI-sites from the same country, as opposed to model performance after generic calibration on all NFI-sites. However, the two types of calibration led to predictions of similar integrated likelihood and NRMSE (Figs 5 and 6). It should be noted that this somewhat surprising result is partly explained by the fact that we had fewer data available per country, so the likely greater relevance of data used in within-country calibration was offset by the low weight of evidence from using data from 2 NFI-sites as compared to 8 in generic BC. Still, it can be conjectured that the considered models are sufficiently general to provide a useful generic parameterisation for Scots pine in Europe, although a future study with larger numbers of NFI-sites per country would be needed to test this hypothesis rigorously. The extra sites should be chosen to cover spatial variation in tree genotypes and geographical conditions. Such increased spatial coverage would also be needed if we want to move from assessing model predictive capacity at site-level to country-wide upscaling.

4.3 *Quantifying and reducing uncertainties*

The extent to which Bayesian calibration can reduce parameter uncertainties of a model depends both on the structure of the model and on the prior distribution assigned by the modeller. In the present study, Bayesian calibration reduced parameter and output uncertainty of all models except the parameter-richest one, ANAFORE. Likewise, the Bayesian model comparison was able to identify which models were most plausible by calculating the integrated likelihood for each model at different stages in the study. The integrated likelihood accounts for parameter uncertainty (by integrating over its distribution) and is a natural way

of combining diverse measurements in one model comparison criterion. This is in contrast to the commonly used NRMSE, which has to be calculated for every variable separately. Another potential advantage of the integrated likelihood over other measures, such as NRMSE and squared correlation coefficient, r^2, is that the integrated likelihood can account for different levels of uncertainty about measurement error for different data points. However, that did not play a role in the present study because all height and diameter data were assumed to have the same degree of uncertainty.

4.4 *Impact of the choices of prior distribution*

As discussed in §§ 4.2-4, the choices made to set the prior probability distributions for the parameters of the different models affected our results to some degree, in particular in the early stages of the analysis where the prior predictive performance of the models was quantified and compared. Because prior distributions for structurally different models cannot be set in a standardised way, and were based on the expertise of the responsible modellers, this introduced a subjective element in the study. This included model-specific choices about parameter-screening, i.e. which of a model's parameters to include in the Bayesian calibration. This subjectivity concerning the prior parameter distribution is unavoidable, to some extent, in any application of Bayesian methodology. However, the procedure we applied here, where all models were calibrated on the same data (NFI) and were subsequently compared against the same independent data (PSP) removed much of the effect of the choice of prior (Figs 3, 4). We therefore suggest that Bayesian model comparisons are most useful after such standardisation.

4.5 *On the use of multiple models*

The use of BMC is formally conditional on one of the models being 'correct' – which is never truly the case in environmental modelling – so we should use the results from the BMC as a guide towards finding the most plausible model in the set of six rather than as formal model probabilities. The results suggest that the 4C model should be recommended as the model of choice for a forest manager who wants to select a single model to help estimate future productivity out of the six models in this study. We believe that for the forest scientist the results are less clear-cut because the Bayesian probabilities do not by themselves explain what makes one model structure more plausible than another. The Bayesian model comparison largely treats the models as black boxes characterised by their input-output relationships. In a previous Bayesian forest model comparison (Van Oijen et al., 2011) it was therefore recommended that after the BC of all models, and their BMC, a detailed analysis should be carried out of the model-data mismatch remaining after calibration. It was recommended in

particular to decompose likelihoods into terms for individual output variables and to decompose mean squared errors (MSE) into terms for bias, variance mismatch and phase-shift (Kobayashi and Salam, 2000). However, in our study with only two output variables and extremely short time-series, these decompositions are not informative. To allow such detailed study of model-data mismatch – and therefore to help explain the results presented here – we would need more detailed data sets, e.g. long time-series of annual data.

Another natural follow-up to BMC, and one that was carried out in this study, is calculating forecasts using Bayesian model averaging (BMA; e.g. Kass and Raftery, 1995). In BMA, no single model is selected for making predictions; instead the probability distributions for the individual model predictions are averaged using as weights the model probabilities determined by the BMC. Because BMA integrates parameter and model structural uncertainty, it is less prone to underestimation of predictive uncertainty than the common practice of selecting and using only a single 'best' model. In the present study, the out-of-sample predictive capacity of BMA was very good, as shown by the NRMSE-values for both output variables in the prior BMA. This is not exceptional; BMA has been reported to have higher forecasting skill than each individual model in other fields, such as medical prognosis (Hoeting et al., 1999) and climate prediction (Min and Hense, 2006). We found that the predictive performance of posterior BMA was only average. However, this was a partly within-sample test - with model probabilities (but not parameters) informed by the PSP-data - so this should be repeated with independent data.

5. Conclusions

- Bayesian calibration successfully reduced uncertainties in parameters and predictions of five out of six forest models.
- Calibrating models separately for each country did not clearly improve within-country predictive capacity compared to generic calibration. This might change when more data become available per country.
- Bayesian model comparison using NFI- and PSP-data identified the 4C model, which is of moderate complexity but mechanistic, as the most plausible forest model after calibration.
- The main caveat to the results is the issue of model initialisation: how it is carried out and which data are available for it. This study suggests that models are favoured that are initialised using on-site measurements of tree growth, unless model complexity requires more data for such initialisation than are available. But model ranking might have been

different if more data, or data from other variables than mean tree height and stem diameter, would have been available for use.

- For a detailed analysis of model-data mismatch, NFI-data are insufficient, but information from PSPs not used in this study, such as single tree data, could be used.
- BMA afforded good out-of-sample forecasts of forest productivity and may be a promising tool for forest management, of sufficient accuracy and precision whilst not underestimating uncertainties.

6. Acknowledgements

We thank the EU for support of all participants through COST Action FP603 and for support of M.v.O. in IP Carbo-Extreme (FP7, GA 226701),. We also thank the national forestry services in Austria, Belgium and Finland for providing the NFI- and PSP-data. The Estonian Meteorological and Hydrological Institute provided climate data and the Estonian Environment Information Centre provided soil data. F.H. acknowledges support from ERC advanced grant 233066.

1

Table 1. Data. Each row represents one of the twelve measurement sites. If multiple values of stem number are shown, they refer to changes over the period of measurement. The rightmost column gives the total number of data points at the site, for tree height and diameter combined.

Country	Site name	Site code	Site type	Lat. (º)	Long. (º)	Plot size (m^2)	Mean temp. (º C)	Mean precip. ($mm\ y^{-1}$)	Age at last obs. (y)	Stem number (ha^{-1})	# Data
Austria	Point 1	A1	NFI	48.31°	14.79°	1200	7.6	855	~64	554-526	4
	Point 2	A2	NFI	48.51°	15.70°	1200	9.2	466	~66	1772-1363	4
	PSP	A3	PSP	48.51°	15.70°	1500	9.2	466	59	790-690	4
Belgium	Hechtel	B1	NFI	51°17'	5°31'	1000	9.9	812	67	400-380	4
	Pijnven	B2	NFI	51°17'	5°31'	1000	9.9	819	66	520-393	4
	Brasschaat	B3	PSP	51°18'	4°31'	20000	9.9	811	79	538-362	6
Estonia	EST-1	E1	PSP	57°51'	25°55'	1963	5.4	629	70	428-402	6
	EST-2	E2	PSP	57°59'	25°38'	1257	5.4	632	67	796-692	6
	EST-3	E3	PSP	57°35'	25°17'	1963	5.3	625	73	652-667	6
Finland	NFI-1	F1	NFI	61° 58'	27° 40'	100-300	2.8	534	75	899	4
	NFI-2	F2	NFI	63° 50'	24° 39'	100-300	2.2	442	55	1067	4
	Vesijako	F3	PSP	61° 20'	25° 2'	1000	3.5	521	79	8700-1710	14

2
3
4
5

Table 2. Models. Each row represents one of the six models. The weather variables driving the models include radiation, temperature, precipitation, wind speed and atmospheric humidity (BASFOR), or a subset of those (3PG, 4C, ANAFORE, BRIDGING, FORMIND). The rightmost column shows whether models simulated forest growth from planting or were initialised using the earliest measurements at each site. IBM = Individual-Based Model requiring specification of size and position of each tree.

Model	Time step	Environmental variables	Number of state variables	Number of parameters (# in calibration)	Initialisation
3PG	Monthly	Weather	9	51 (48)	Planting date
4C	Daily-Yearly	Weather, Soil conditions, N-deposition, CO_2	15	46 (43)	First measurement
ANAFORE	Half-hourly	Weather, Soil conditions, N-deposition, CO_2	26	146 (138)	First measurement
BASFOR	Daily	Weather, N-deposition, CO_2, Soil conditions	14	48 (41)	Planting date
BRIDGING	Yearly	Weather	5	38 (13)	First measurement
FORMIND	Yearly	Weather	IBM	42 (4)	First measurement

6
7

1
2

Table 3. Prior predictions by six models of final tree height (m) and stem diameter (mm) on twelve sites. Site-codes (A1, A2, etc.) are explained in Table 1. For each combination of model and variable, the first row shows the predictions using the mode of the prior parameter distribution, and the second gives the range (5%-95% quantiles). The upper two rows show the measured values for comparison.

Source	Variable	A1	A2	A3	B1	B2	B3	E1	E2	E3	F1	F2	F3
Data	**Height**	**18.5**	**17.7**	**18.1**	**18.4**	**23.2**	**21.3**	**25.0**	**24.9**	**25.6**	**17.8**	**10.1**	**21.8**
	Diameter	**324**	**207**	**239**	**271**	**293**	**319**	**274**	**237**	**245**	**191**	**146**	**170**
3PG	Height	52.4	21.0	28.4	28.6	28.8	32.8	40.7	32.7	36.0	30.2	23.5	19.5
		21.3-145	*10.7-45.0*	*13.5-62.1*	*13.1-66.9*	*13.5-67.6*	*14.3-82.2*	*17.7-102*	*15.4-78.9*	*16.3-88.5*	*14.1-68.0*	*11.5-47.6*	*9.3-43.6*
	Diameter	622	211	303	301	305	356	462	357	400	325	241	194
		337-1476	*140-403*	*195-568*	*178-607*	*188-599*	*201-760*	*287-960*	*227-749*	*248-865*	*205-646*	*156-430*	*110-407*
4C	Height	21.6	20.9	20.7	19.6	23.1	24.5	22.5	20.7	21.8	16.7	12.5	26.0
		15.9-29.1	*15.6-27.2*	*14.3-29.9*	*17.8-25.0*	*20.0-30.1*	*19.2-32.6*	*20.0-29.3*	*19.0-25.4*	*21.3-26.0*	*14.4-22.2*	*7.6-20.9*	*10.2-45.3*
	Diameter	381	267	284	287	297	352	288	254	244	205	161	340
		291-430	*191-298*	*191-344*	*267-305*	*250-322*	*263-398*	*243-320*	*211-271*	*224-271*	*170-233*	*120-201*	*139-495*
ANAFORE	Height	30.2	27.6	28.5	19.4	25.4	46.9	29.0	28.7	24.7	26.7	20.5	48.0
		23.9-59.2	*17.4-59.1*	*18.3-59.2*	*18.9-23.1*	*23.3-33.6*	*31.4-59.0*	*18.8-52.0*	*20.5-51.6*	*18.5-59.2*	*20.3-49.5*	*10.0-46.6*	*22.4-59.3*
	Diameter	457	185	330	309	323	457	471	355	376	280	238	219
		335-481	*182-195*	*222-331*	*299-323*	*303-344*	*417-516*	*277-426*	*210-326*	*241-364*	*245-314*	*206-436*	*89-237*
BASFOR	Height	25.9	14.6	18.9	22.5	18.9	21.2	18.0	17.9	19.0	16.4	14.6	13.1
		12.6-48.1	*1.4-36.2*	*1.7-40.2*	*10.8-41.6*	*1.4-36.9*	*5.8-39.9*	*7.8-33.9*	*7.8-33.4*	*8.3-35.6*	*2.5-31.1*	*2.2-27.9*	*3.1-24.7*
	Diameter	229	98	144	186	144	170	133	132	145	115	97	82
		131-319	*3-221*	*3-261*	*103-259*	*3-220*	*31-244*	*52-190*	*49-189*	*62-208*	*6-170*	*4-143*	*9-119*
BRIDGING	Height	18.2	17.5	18.2	19.2	21.8	22.6	22.7	21.4	23.9	17.5	11.5	12.9
		17.5-18.8	*17.0-18.1*	*17.0-19.4*	*18.9-19.6*	*21.5-22.2*	*22.0-23.2*	*22.1-23.3*	*20.9-22.0*	*23.3-24.5*	*16.6-18.4*	*10.0-13.0*	*12.1-16.8*
	Diameter	423	261	305	312	331	353	320	271	279	226	210	265
		375-442	*229-273*	*261-321*	*296-321*	*302-349*	*327-363*	*290-334*	*245-282*	*255-289*	*200-237*	*175-225*	*233-388*
FORMIND	Height	26.6	21.0	22.1	22.0	20.9	22.1	20.9	18.5	19.8	16.0	11.0	8.0
		16.0-32.4	*12.0-26.3*	*12.5-29.1*	*14.8-26.4*	*15.1-26.0*	*16.0-27.6*	*14.3-25.9*	*13.0-22.7*	*13.5-24.5*	*11.2-19.6*	*8.2-13.1*	*6.3-9.1*
	Diameter	352	251	270	268	250	270	250	210	230	170	100	63
		302-362	*190-264*	*201-288*	*260-273*	*250-273*	*270-305*	*250-251*	*210-212*	*230-232*	*170-170*	*100-102*	*56-78*

3

4

References

Andren, O. and Katterer, T., 1997. ICBM: The introductory carbon balance model for exploration of soil carbon balances. Ecological Applications, 7(4): 1226-1236.

Baeyens, L., van Slycken, J. and Stevens, D., 1993. Description of the soil profile in Brasschaat, Institute of Forestry and Game Management, Geraardsbergen, Belgium.

Bugmann, H., Grote, R., Lasch, P., Lindner, M. and Suckow, F., 1997. A new forest gap model to study the effects of environmental change on forest structure and functioning, Impacts of Global Change on Tree Physiology and Forest Ecosystems. Forestry Sciences. Springer, Dordrecht, pp. 255-261.

Curiel Yuste, J., Konopka, B., Janssens, I.A., Coenen, K., Xiao, C.W. and Ceulemans, R., 2005. Contrasting net primary productivity and carbon distribution between neighboring stands of Quercus robur and Pinus sylvestris. Tree Physiology, 25(6): 701-12.

Davidson, R.L., 1969. Effect of root/leaf temperature differentials on root/shoot ratios in some pasture grasses and clover. Annalsof Botany 33: 561-569.Deckmyn, G., Mali, B., Kraigher, H., Torelli, N., Op de Beeck, M. and Ceulemans, R., 2009. Using the process-based stand model ANAFORE including Bayesian optimisation to predict wood quality and quantity and their uncertainty in Slovenian Beech. Silva Fennica, 43(3): 523-534.

Deckmyn, G., Verbeeck, H., Op de Beeck, M., Vansteenkiste, D., Steppe, K. and Ceulemans, R., 2008. ANAFORE: A stand-scale process-based forest model that includes wood tissue development and labile carbon storage in trees. Ecological Modelling, 215(4): 345-368.

Duursma, R.A. and Makela, A., 2007. Summary models for light interception and light-use efficiency of non-homogeneous canopies. Tree Physiology, 27(6): 859-870.

Fontes, L., Bontemps, J.D., Bugmann, H., Van Oijen, M., Gracia, C., Kramer, K., Lindner, M., Rotzer, T. and Skovsgaard, J.P., 2010. Models for supporting forest management in a changing environment. Forest Systems, 19: 8-29.

Fu, Y.H., Campioli, M., Van Oijen, M., Deckmyn, G. and Janssens, I., 2012. Bayesian comparison of six different temperature-based budburst models for four temperate tree species. Ecological Modelling, 230: 92-100.

Gelman, A. and Rubin, D.B., 1992. Inference from iterative simulation using multiple sequences. Statistical Science, 7: 457-511.

Green, E.J., MacFarlane, D.W., Valentine, H.T. and Strawderman, W.E., 1999. Assessing uncertainty in a stand growth model by Bayesian synthesis. Forest Science, 45(4): 528-538.

Gutiérrez, A.G. and Huth, A., 2012. Successional stages of primary temperate rainforests of Chiloé Island, Chile. Perspectives in Plant Ecology, Evolution and Systematics, in press.

Härkönen, S., Pulkkinen, M., Duursma, R. and Mäkelä, A., 2010. Estimating annual GPP, NPP and stem growth in Finland using summary models. Forest Ecology and Management, 259(3): 524-533.

Hartig, F., Dyke, J., Hickler, T., Higgins, S., O'Hara, R.B., Scheiter, S. and Huth, A., 2012. Connecting dynamic vegetation models to data - an inverse perspective. Journal of Biogeography, accepted.

Haxeltine, A., Prentice, I.C., 1996. A general model for the light-use efficiency of primary production. Functional Ecology 10: 551-561.Hoeting, J.A., Madigan, D., Raftery, A.E. and Volinsky, C.T., 1999. Bayesian model averaging: A tutorial. Statistical Science, 14(4): 382-401.

Kass, R.E. and Raftery, A.E., 1995. Bayes Factors. Journal of the American Statistical Association, 90(430): 773-795.

Kobayashi, K. and Salam, M.U., 2000. Comparing simulated and measured values using mean squared deviation and its components. Agronomy Journal, 92(2): 345-352.

Köhler, P. and Huth, A., 1998. The effects of tree species grouping in tropical rainforest modelling: Simulations with the individual-based model FORMIND. Ecological Modelling, 109(3): 301-321.

Landsberg, J., Makela, A., Sievanen, R. and Kukkola, M., 2005. Analysis of biomass accumulation and stem size distributions over long periods in managed stands of *Pinus sylvestris* in Finland using the 3-PG model. Tree Physiology, 25(7): 781-792.

Landsberg, J.J. and Waring, R.H., 1997. A generalised model of forest productivity using simplified concepts of radiation-use efficiency, carbon balance and partitioning. Forest Ecology and Management, 95(3): 209-228.

Lasch, P., Badeck, F.W., Suckow, F., Lindner, M. and Mohr, P., 2005. Model-based analysis of management alternatives at stand and regional level in Brandenburg (Germany). Forest Ecology and Management, 207(1-2): 59-74.

Law, B.E., Waring, R.H., Anthoni, P.M. and Aber, J.D., 2000. Measurements of gross and net ecosystem productivity and water vapour exchange of a *Pinus ponderosa* ecosystem, and an evaluation of two generalized models. Global Change Biology, 6(2): 155-168.

Lehuger, S., Gabrielle, B., Van Oijen, M., Makowski, D., Germon, J.C., Morvan, T. and Hénault, C., 2009. Bayesian calibration of the nitrous oxide emission module of an agro-ecosystem model. Agriculture, Ecosystems & Environment, 133(3-4): 208-222.

Levy, P.E., Wendler, R., Van Oijen, M., Cannell, M.G.R. and Millard, P., 2004. The effects of nitrogen enrichment on the carbon sink in coniferous forests: uncertainty and sensitivity analyses of three ecosystem models. Water, Air and Soil Pollution: Focus, 4: 67-74.

Luo, Y., Weng, E., Wu, X., Gao, C., Zhou, X. and Zhang, L., 2009. Parameter identifiability, constraint, and equifinality in data assimilation with ecosystem models. Ecological Applications, 19(3): 571-574.

Mäkelä, A., 1997. A Carbon Balance Model of Growth and Self-Pruning in Trees Based on Structural Relationships. Forest Science 43: 7-24.

Mäkelä, A., del Río, M., Hynynen, J., Hawkins, M.J., Reyer, C., Soares, P., Van Oijen, M. and Tomé, M., 2012. Using forest growth models for estimating indicators of sustainable forest management. Forest Ecology and Management 285: 164-178.

Mäkelä, A., Pulkkinen, M., Kolari, P., Lagergren, F., Berbigier, P., Lindroth, A., Loustau, D., Nikinmaa, E., Vesala, T. and Hari, P., 2008. Developing an empirical model of stand GPP with the LUE approach: analysis of eddy covariance data at five contrasting conifer sites in Europe. Global Change Biology, 14(1): 92-108.

Mäkelä, A. and Sievanen, R., 1992. Height growth strategies in open-grown trees. Journal of Theoretical Biology, 159(4): 443-467.

Mäkelä A. and Vanninen P., 2001. Vertical structure of Scots pine crowns in different age and size classes. Trees 15: 385-392.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E., 1953. Equation of state calculations by fast computing machines. Journal of Chemical Physics, 21: 1087-1092.

Min, S.-K. and Hense, A., 2006. A Bayesian approach to climate model evaluation and multi-model averaging with an application to global mean surface temperatures from IPCC AR4 coupled climate models. Geophys. Res. Lett., 33(8): L08708.

Palmroth, S., Berninger, F., Nikinmaa, E., Lloyd, J., Pulkkinen, P., Hari, P., 1999. Structural adaptation rather than water conservation was observed in Scots pine over a range of wet to dry climates. Oecologia. 121: 302-309.Radtke, P.J., Burk, T.E. and Bolstad, P.V., 2001. Estimates of the distributions of forest ecosystem model inputs for deciduous forests of eastern North America. Tree Physiology, 21(8): 505-512.

Reinds, G.J., van Oijen, M., Heuvelink, G.B.M. and Kros, H., 2008. Bayesian calibration of the VSD soil acidification model using European forest monitoring data. Geoderma, 146(3-4): 475-488.

Reyer, C., Lasch, P., Mohren, G.M.J. and Sterck, F.J., 2010. Inter-specific competition in mixed forests of Douglas fir (*Pseudotsuga menziesii*) and common beech (*Fagus sylvatica*) under climate change - a model-based analysis. Annals of Forest Science, 67(8): 11.

Ruger, N., Gutierrez, A.G., Kissling, W.D., Armesto, J.J. and Huth, A., 2007. Ecological impacts of different harvesting scenarios for temperate evergreen rain forest in southern Chile - A simulation experiment. Forest Ecology and Management, 252(1-3): 52-66.

Sands, P.J. and Landsberg, J.J., 2002. Parameterisation of 3-PG for plantation grown *Eucalyptus globulus*. Forest Ecology and Management, 163(1-3): 273-292.

Seidl, R., Rammer, W., Lasch, P., Badeck, F.W. and Lexer, M.J., 2008. Does conversion of even-aged, secondary coniferous forests affect carbon sequestration? A simulation study under changing environmental conditions. Silva Fennica, 42(3): 369-386.

Shinozaki, K., Yoda, K., Hozumi, K., Kira, T., 1964. A quantitative analysis of plant form - the pipe model theory. I. Basic analysis. Japanese Journal of Ecology 14: 97-105.

Sims, A., Kiviste, A., Hordo, M., Laarmann, D. and von Gadow, K., 2009. Estimating tree survival: a study based on the Estonian Forest Research Plots Network. Annales Botanici Fennici, 46(4): 336-352.

Suckow, F., Badeck, F.W., Lasch, P. and Schaber, J., 2001. Nutzung von Level-II-Beobachtungen für Test und Anwendungen des Sukzessionsmodells FORESEE. Beiträge für Forstwirtschaft und Landschaftsökologie, 35: 84-87.

Tuomi, M., Vanhala, P., Karhu, K., Fritze, H. and Liski, J., 2008. Heterotrophic soil respiration--Comparison of different models describing its temperature dependence. Ecological Modelling, 211(1-2): 182-190.

Valentine, H.T. and Mäkelä, A., 2005. Bridging process-based and empirical approaches to modeling tree growth. Tree Physiology, 25(7): 769-779.

Vanninen, P. and Mäkelä, A. 2005. Carbon budget for Scots pine trees: Effect of size, competition and site fertility on growth allocation and production. Tree Physiology 25: 17-30.

Van Oijen, M., Cameron, D.R., Butterbach-Bahl, K., Farahbakhshazad, N., Jansson, P.E., Kiese, R., Rahn, K.H., Werner, C. and 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.

Van Oijen, M., Dauzat, J., Harmand, J.-M., Lawson, G. and Vaast, P., 2010. 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. 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.

Waring, R.H., Landsberg, J.J. and Williams, M., 1998. Net primary production of forests: a constant fraction of gross primary production? Tree Physiology, 18(2): 129-134.

Xenakis, G., Ray, D. and Mencuccini, M., 2008. Sensitivity and uncertainty analysis from a coupled 3-PG and soil organic matter decomposition model. Ecological Modelling, 219(1-2): 1-16.

Figure

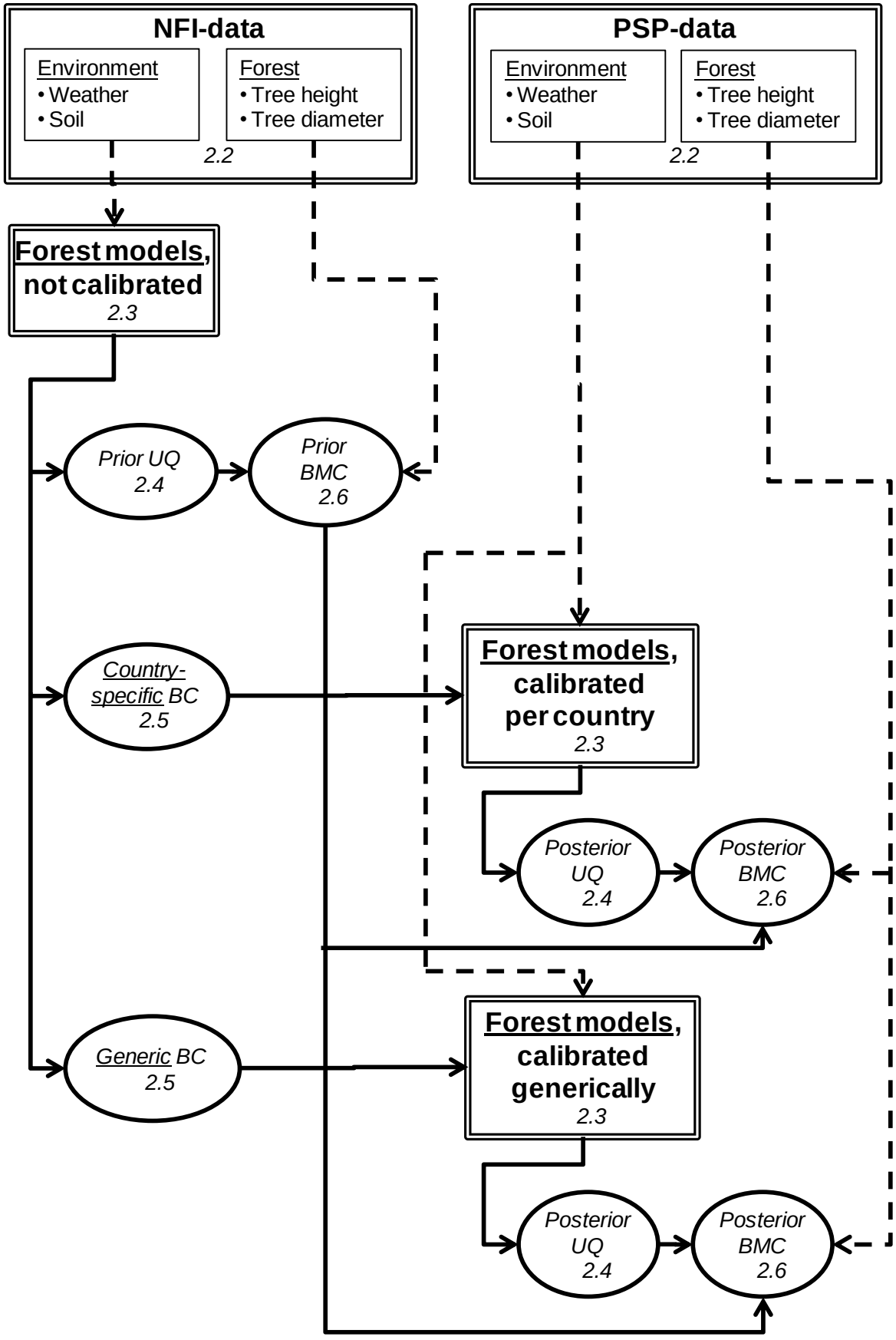

Figure 1. Flow chart of the study. The numbers within icons (2.2-2.6) indicate in which paragraph of the Materials and Methods further explanation of can be found.

Figure

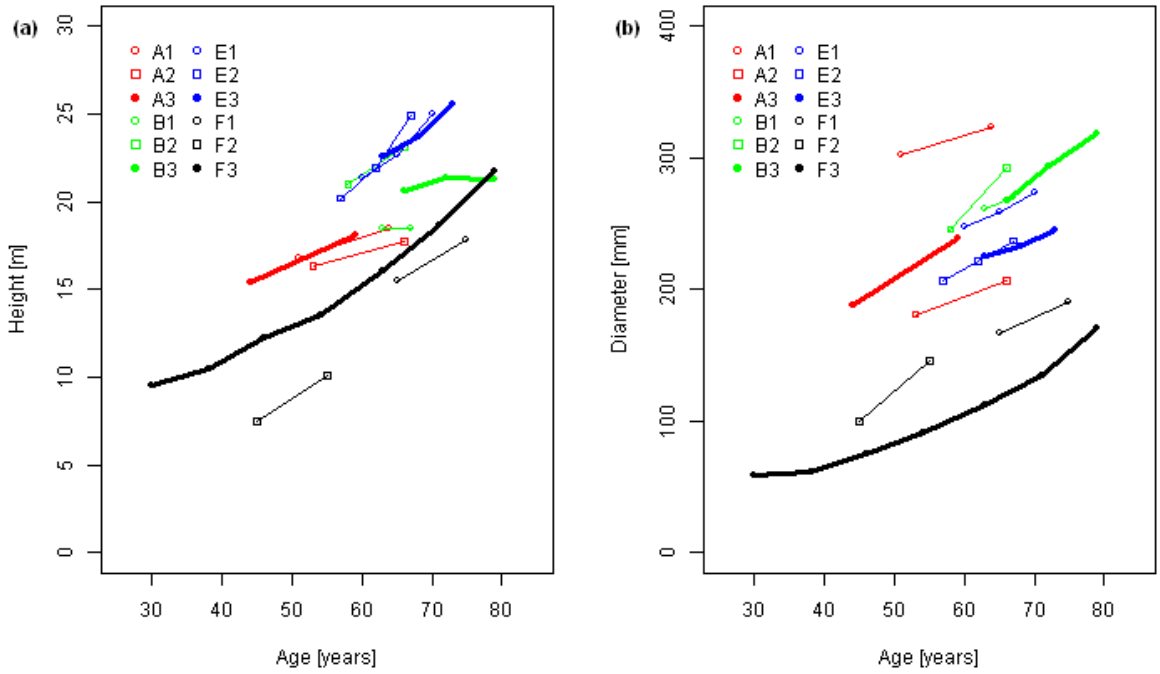

Figure 2. (a) Mean tree height vs. stand age as observed at the twelve forest sites. (b) Idem for stem diameter. Site-codes (A1 .. F3) are explained in Table 1.

Figure

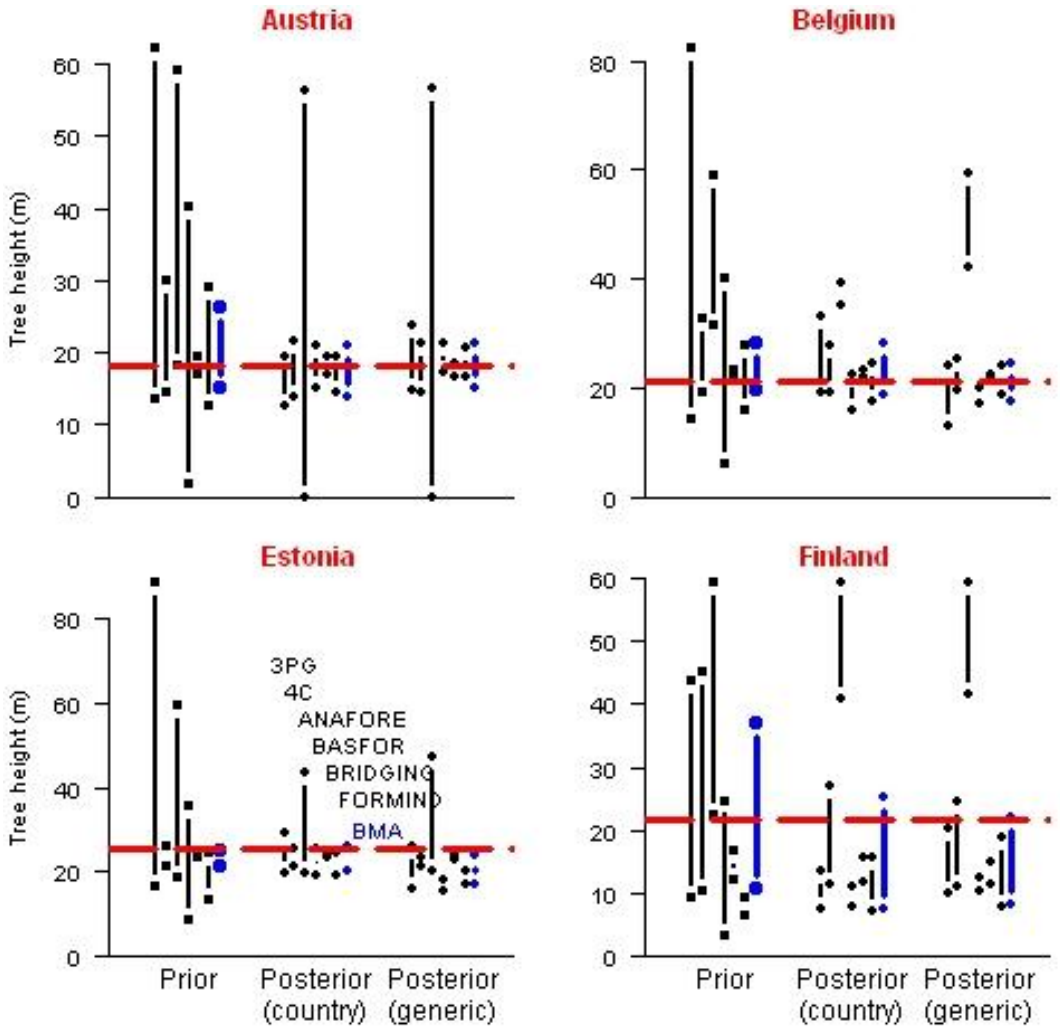

Figure 3. Model output uncertainty for final mean tree height at the PSP-sites A3, B3, E3, F3. Vertical bars show the central 90% of distributions. For each country, the three clusters of bars show prior and posterior (country-specific, generic) predictions. The seven bars in each cluster are for the six models plus the Bayesian Model Averaging result, in the order indicated in the bottom-left panel. The dashed horizontal lines indicate observed values, which were not used for model calibration.

Figure

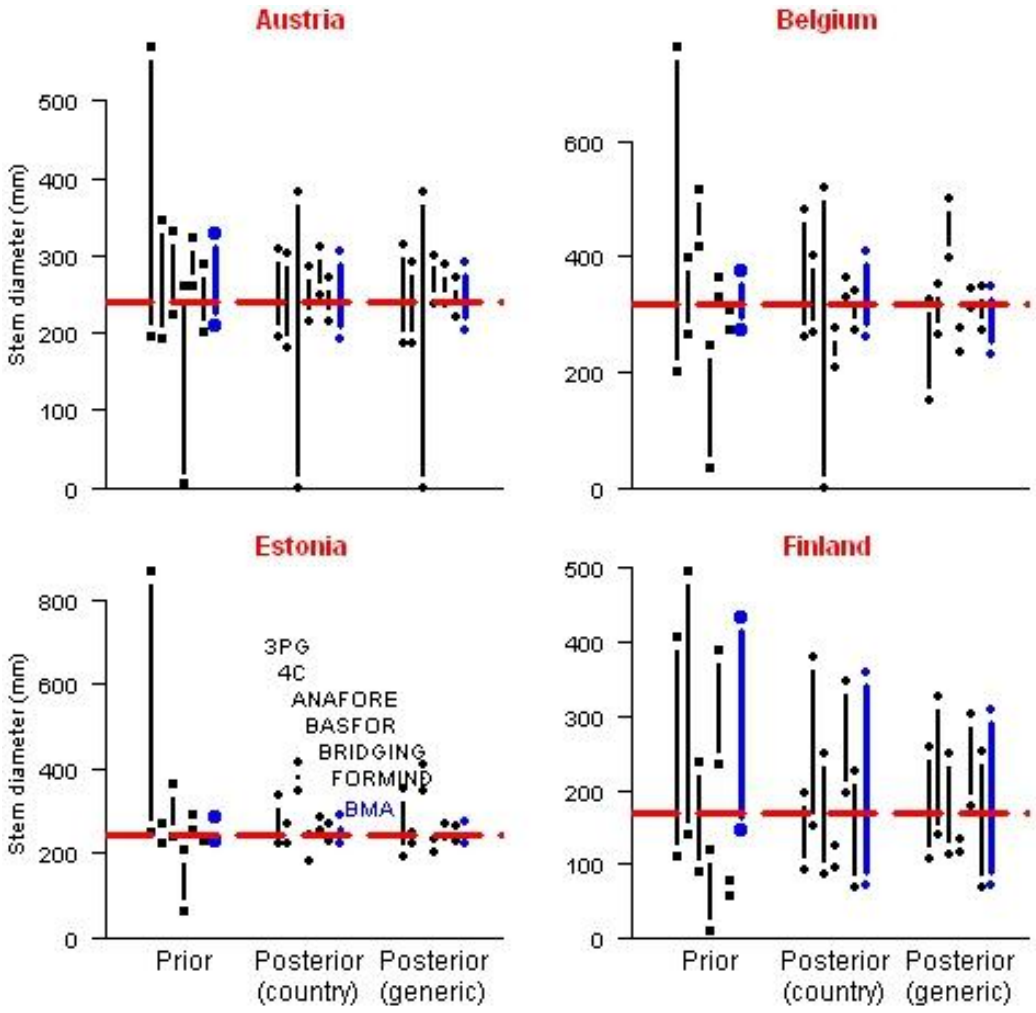

Figure 4. Model output uncertainty for final mean stem diameter at the PSP-sites A3, B3, E3, F3. The lay-out of the figure is the same as for Fig. 3.

Figure

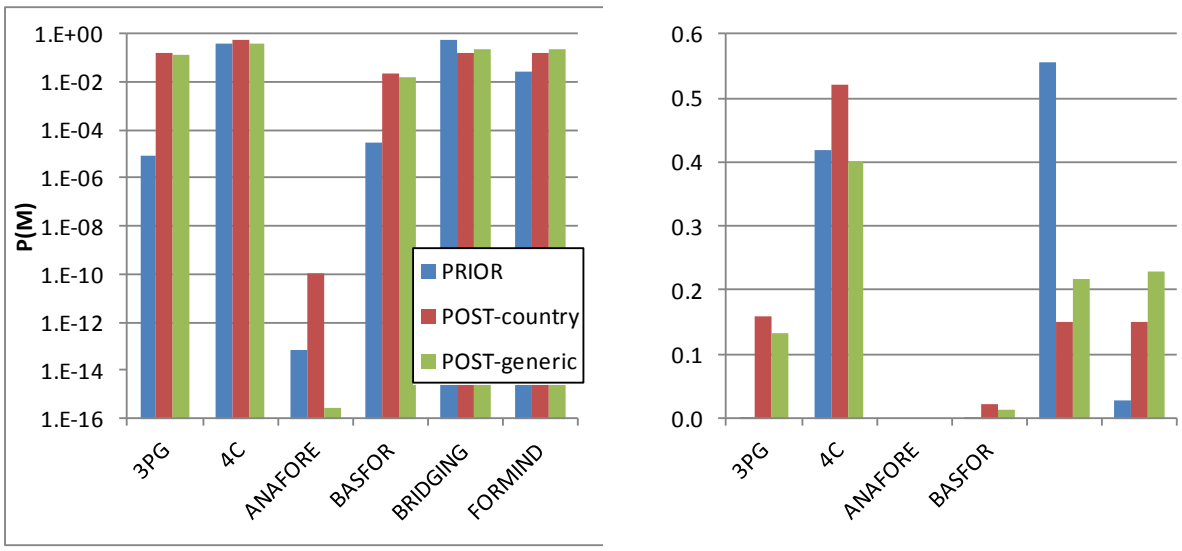

Figure 5. Prior and posterior model probabilities, derived from the integrated likelihoods of NFI and PSP-measurements. Left: logarithmic scale; Right: absolute scale.

Figure

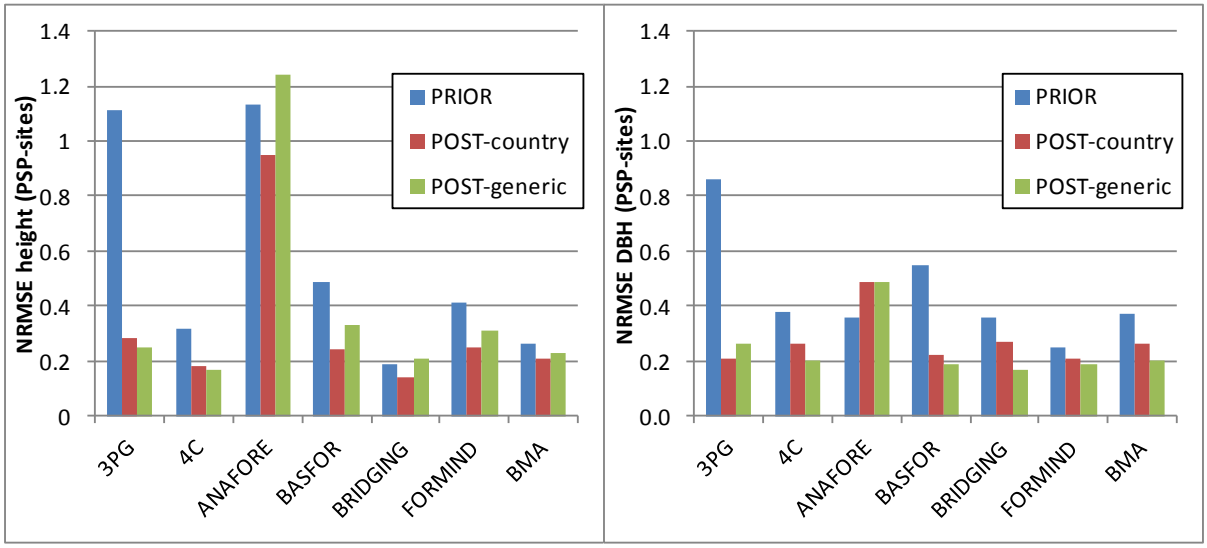

Figure 6. Normalised RMSE, derived from simulations at PSP-sites using samples from prior and posterior parameter distributions. Left: tree height, right: diameter at breast height. The rightmost three bars in both panels are the result of Bayesian Model Averaging (BMA).

Highlights

- Bayesian calibration successfully reduced uncertainties in parameters and predictions of five out of six forest models.
- Calibrating models separately for each country did not clearly improve within-country predictive capacity compared to generic calibration. This might change when more data become available per country.
- Bayesian model comparison using NFI- and PSP-data identified the 4C model, which is of moderate complexity but mechanistic, as the most plausible forest model after calibration.
- The main caveat to the results is the issue of model initialisation: how it is carried out and which data are available for it. This study suggests that models are favoured that are initialised using on-site measurements of tree growth, unless model complexity requires more data for such initialisation than are available. But model ranking might have been different if more data, or data from other variables than mean tree height and stem diameter, would have been available for use.
- For a detailed analysis of model-data mismatch, NFI-data are insufficient, but information from PSPs not used in this study, such as single tree data, could be used.
- BMA afforded good out-of-sample forecasts of forest productivity and may be a promising tool for forest management, of sufficient accuracy and precision whilst not underestimating uncertainties.