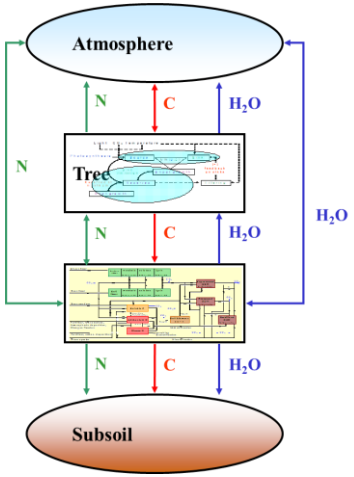
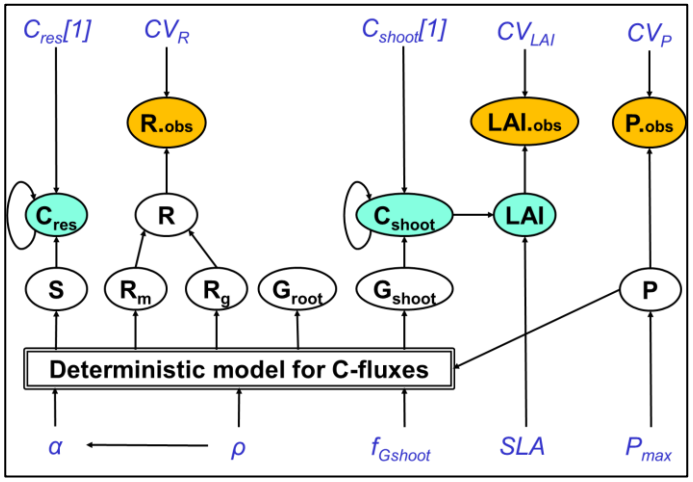
Bayesian Methods for Ecological and Environmental Modelling: MODEL COMPARISON & EVALUATION





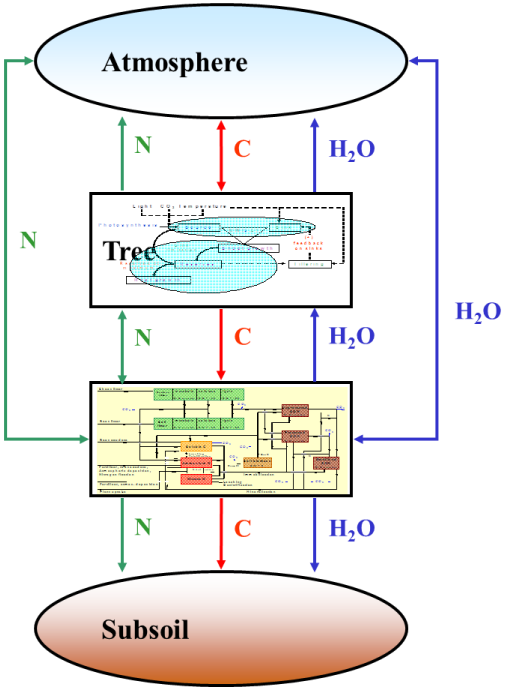
Marcel van Oijen
CEH-Edinburgh, 2019-09-12

Contents

1. Introduction

1.1 Using environmental models: Uncertainties

Uncertainty in input maps and model parameter values



Uncertainty about model structure

BC, DA, MDF

BMC, BMA

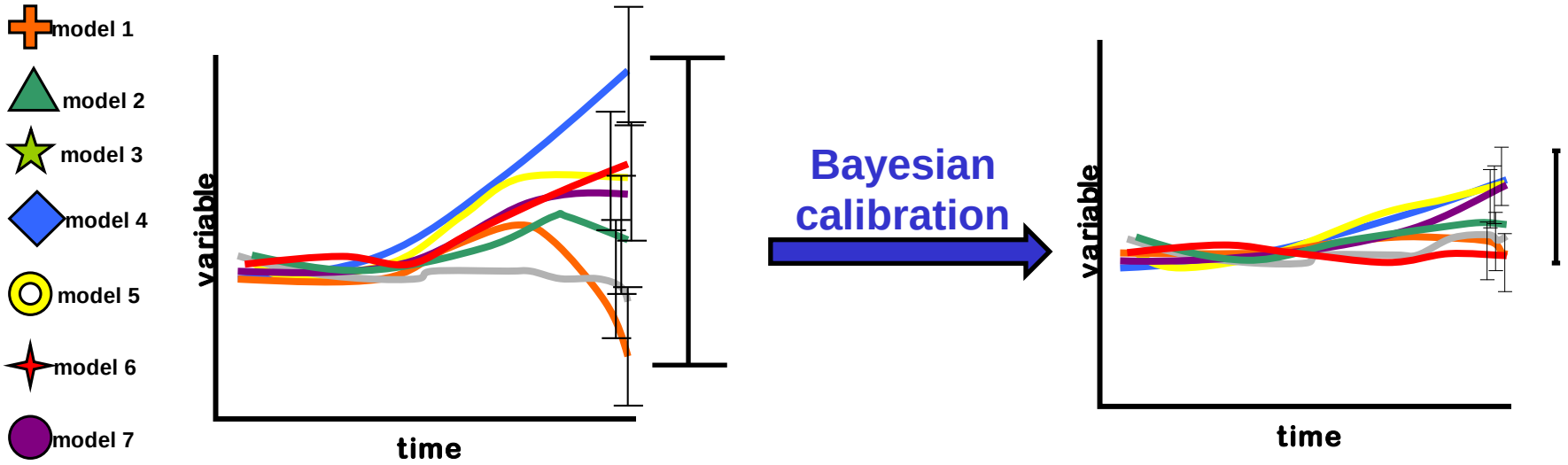
General approach to uncertainty quantification and reduction:

(1) Specify prior probability distributions

(2) Get data

(3) Determine posterior distributions

1.2 Workflow of model calibration & comparison



- ***Parameter uncertainty reduced***
- ***Model structural uncertainty highlighted***

1.3 BC & BMC

1. **Express all uncertainties probabilistically** ⇒
   Assign probability distributions to (1) data, (2) the collection of models, (3) the parameter-set of each individual model

2. **Use the rules of probability theory to transfer the information from the data to the probability distributions for models and parameters**

   Main tool from probability theory to do this: Bayes' Theorem

$$P(\alpha|D) \propto P(\alpha)\ P(D|\alpha)$$

Posterior **is proportional to** ***prior*** **times** ***likelihood***

- **α = parameter set ⇒** ***parameterisation*** **("Bayesian Calibration", BC)**
- **α = model set ⇒** ***model evaluation*** **("Bayesian Model Comparison", BMC)**

1.4 A simple algorithm for BMC

Ingredients

1. Multiple models: $M^1, \ldots, M^n$
2. For each model, a list of its parameters: $\theta^1, \ldots, \theta^n$
3. Data: D

Recipe

1. For each model M^i, take a sample of k parameter vectors $\theta^i_{j=1,k}$
2. For each parameter vector, calculate $L(\theta^i_j)=p(D|M^i,\theta^i_j)$
3. For each model calculate its ***integrated likelihood*** = $p(D|M^i)$ = average of $L(D|M^i,\theta^i_j)$
4. Assuming equal prior model probabilities $p(M^i)=1/n$, the **posterior probability for each model** = $p(M^i|D) = L(D|M^i,\theta^i_j) / \sum L$

1.5 BMC

1a. The probability distribution for the models could assume that, a priori, all models have equal probability of being correct ⇒

$$P(M) = P(\{M^1 , \dots , M^n\}) = \{1/n , \dots , 1/n\}$$

1b. Define the prior probability distributions for the parameters. Each model has its own list of parameters so there will be *n* different parameter distributions $P(\theta^1), \dots , P(\theta^n)$.

1c. Define the uncertainty about the data, e.g. by assigning Gaussians for uncertainty about measurement error.

2a. Calculate, for each model separately, its 'Integrated Likelihood'. That is the probability of finding the observed data if the model is correct but accounting for uncertainty about its parameter values:

$$P(D|M^i) = \int P(\theta^i)\, P(D|M^i,\theta^i)\, d\theta^i$$

2b. Apply Bayes' Theorem to calculate the posterior distribution for the models:

$$P(M^i|D) = P(D|M^i)\, P(M^i) \,/\, \Sigma\, P(D|M^j)\, P(M^j)$$

If we, in step 1a, assumed all $P(M_j)=1/n$, then the preceding formula simplifies:

$$P(M^i|D) = P(D|M^i) \,/\, \Sigma\, P(D|M^j)$$

1.6 Kass & Raftery 1995

Bayes Factors

Robert E. KASS and Adrian E. RAFTERY*

In a 1935 paper and in his book *Theory of Probability*, Jeffreys developed a methodology for quantifying the evidence in favor of a scientific theory. The centerpiece was a number, now called the *Bayes factor*, which is the posterior odds of the null hypothesis when the prior probability on the null is one-half. Although there has been much discussion of Bayesian hypothesis testing in the context of criticism of *P*-values, less attention has been given to the Bayes factor as a practical tool of applied statistics. In this article we review and discuss the uses of Bayes factors in the context of five scientific applications in genetics, sports, ecology, sociology, and psychology.

We emphasize the following points:

- From Jeffreys' Bayesian viewpoint, the purpose of hypothesis testing is to evaluate the evidence in favor of a scientific theory.
- Bayes factors offer a way of evaluating evidence *in favor of* a null hypothesis.
- Bayes factors provide a way of incorporating external information into the evaluation of evidence about a hypothesis.
- Bayes factors are very general and do not require alternative models to be nested.
- Several techniques are available for computing Bayes factors, including asymptotic approximations that are easy to compute using the output from standard packages that maximize likelihoods.
- In "nonstandard" statistical models that do not satisfy common regularity conditions, it can be technically simpler to calculate Bayes factors than to derive non-Bayesian significance tests.
- The Schwarz criterion (or BIC) gives a rough approximation to the logarithm of the Bayes factor, which is easy to use and does not require evaluation of prior distributions.
- When one is interested in estimation or prediction, Bayes factors may be converted to weights to be attached to various models so that a composite estimate or prediction may be obtained that takes account of structural or model uncertainty.
- Algorithms have been proposed that allow model uncertainty to be taken into account when the class of models initially considered is very large.
- Bayes factors are useful for guiding an evolutionary model-building process.
- It is important, and feasible, to assess the sensitivity of conclusions to the prior distributions used.

KEY WORDS: Bayesian hypothesis tests; BIC; Importance sampling; Laplace method; Markov chain Monte Carlo; Model selection; Monte Carlo integration; Posterior model probabilities; Posterior odds; Quadrature; Schwarz criterion; Sensitivity analysis; Strength of evidence.

1. INTRODUCTION

The Bayesian approach to hypothesis testing was developed by Jeffreys (1935, 1961) as a major part of his program for scientific inference. Although Jeffreys called his methods "significance tests," apparently borrowing the term from Fisher, this is misleading, because Jeffreys's perspective and goals were quite different. Jeffreys was concerned with the comparison of predictions made by two competing scientific theories. In his approach, statistical models are introduced to represent the probability of the data according to each of the two theories, and Bayes's theorem is used to compute the posterior probability that one of the theories is correct.

Considerable attention has been given to distinctions between the two approaches (e.g., Berger and Delampady 1987, Berger and Berry 1988, and references therein). Often lost from the controversy, however, are the practical aspects of the Bayesian methods: how conclusions may be drawn from them, how they can provide answers when non-Bayesian methods are hard to construct, what their strengths and limitations are. These concerns are the focus of this article. We will also discuss the Bayesian approach to accounting for uncertainty in the model-building process, which is closely connected to the methodology for hypothesis testing.

In Section 2 we motivate the work with several applications from the areas of genetics, sports, ecology, sociology, and psychology. These help connect hypothesis testing with model selection and introduce several problems that Bayesian methodology can solve, including the evaluation of the evidence in favor of a null hypothesis, the inclusion of other information in the weighing of evidence, the comparison of nonnested models, and accounting for uncertainty in the choice of models. In Section 3 we introduce the *Bayes factor*, which is the posterior odds of one hypothesis when the prior probabilities of the two hypotheses are equal.

Bayesian methods involve integrals and thus, often, numerical integration. Many integration techniques have been adapted to problems of Bayesian inference, including the computation of Bayes factors; this is discussed in Section 4. Bayes factors require priors on the parameters appearing in the models that represent the competing hypotheses. The choice of these priors and the extent to which Bayes factors are sensitive to this choice is discussed in Section 5.

* Robert E. Kass is Professor, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213. Adrian E. Raftery is Professor of Statistics and Sociology, Department of Statistics, University of Washington, Seattle, WA 98195. Kass's research was supported by National Science Foundation Grant DMS-9005858 and by National Institutes of Health Grant RO1-CA54852-01. Raftery's research was supported by Office of Naval Research Contract N-00014-91-J-1074, by the Ministère de la Recherche et de l'Espace, Paris, by the Université de Paris VI, and by INRIA, Rocquencourt, France. Raftery thanks the latter two institutions, Paul Deheuvels, and Gilles Celeux for hearty hospitality during his Paris sabbatical in which part of this article was written. The authors are grateful to former editor Don Guthrie for encouraging them to write this article, to David Madigan and Larry Wasserman for many helpful comments and discussions, and to Jim Albert, James Dickey, Andrew Gelman, Julia Mortera, Michael Newton, Sue Rosenkranz, Michael Sobel, Mike Titterington, the editor, the associate editor, and two anonymous referees for very helpful comments on an earlier version of the article.

Journal of the American Statistical Association
June 1995, Vol. 90, No. 430, Review Paper

B_{10}	Evidence against H_0
1 to 3.2	Not worth more than a bare mention
3.2 to 10	Substantial
10 to 100	Strong
>100	Decisive

1.7 BMC: Tuomi et al. 2007

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/ecolmodel

Heterotrophic soil respiration—Comparison of different models describing its temperature dependence

Mikko Tuomi[a,b,*], Pekka Vanhala[a], Kristiina Karhu[a], Hannu Fritze[c], Jari Liski[a]

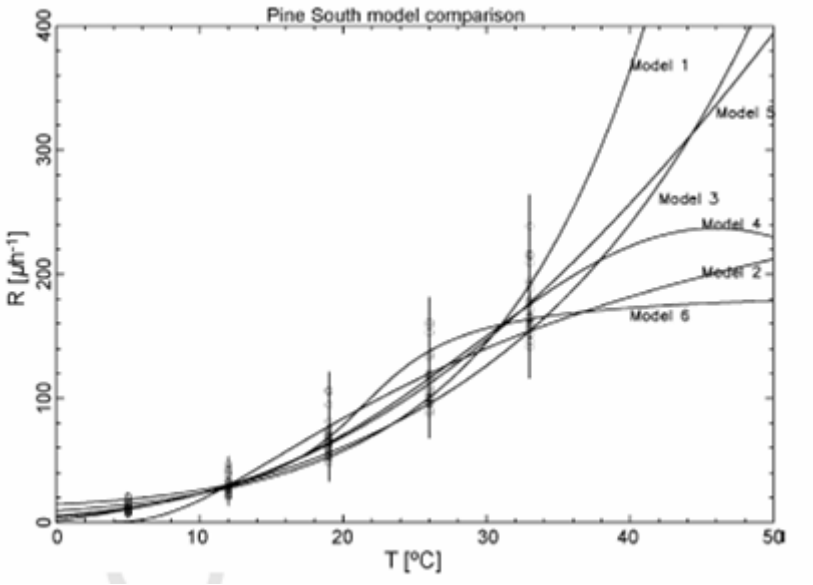


Table 2 – Conditional probabilities of different models, calculated using Bayes factors

	M1	M2	M3	M4	M5	M6
P (MG and NS)	0.015	0.001	0.276	0.398	0.217	0.093

1.8 Don't select a model but use them all: BMA

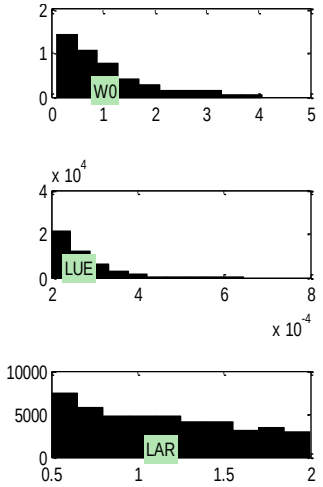
- **Bayesian Model Averaging (BMA) takes the results from multiple models and averages the probabilities of all possible output values. NOTE: BMA does not average the output values themselves.**

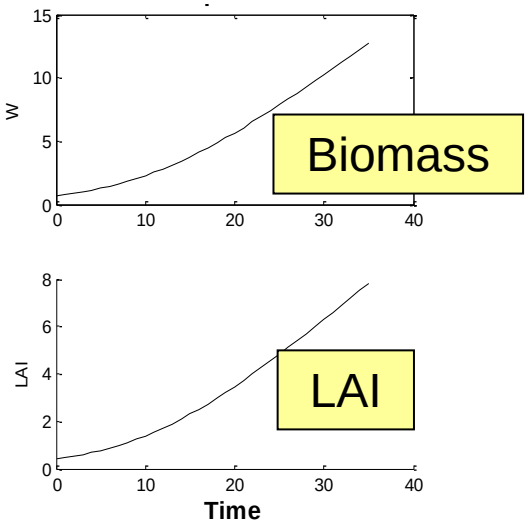
2.
From three to four parameters

2.1 Bayesian comparison of two expolinear models

SIMPLE EXPOLINEAR

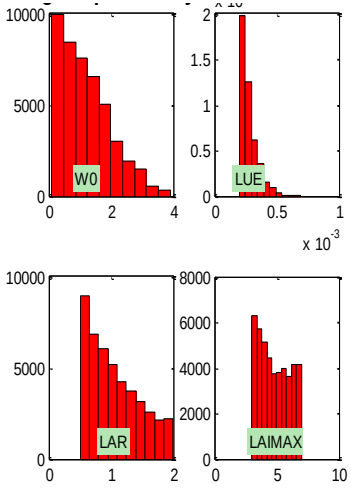
- **Constant light intensity**
- **Light int. by Beer's law**
- **Constant LUE**
- ***Constant LAR***

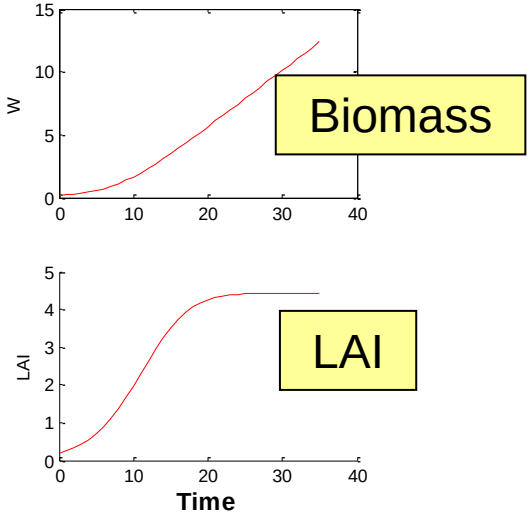




COMPLEX EXPOLINEAR

- **Constant light intensity**
- **Light int. by Beer's law**
- **Constant LUE**
- ***LAI reaches maximum***

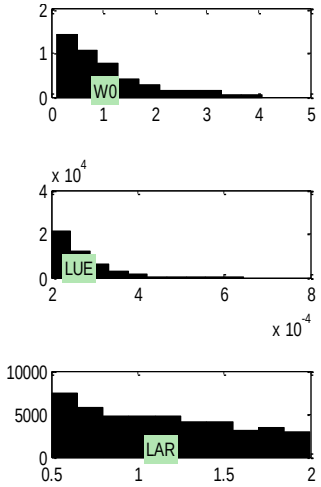




2.2 Bayesian comparison of two expolinear models

SIMPLE EXPOLINEAR

- **Constant light intensity**
- **Light int. by Beer's law**
- **Constant LUE**
- ***Constant LAR***



Bayes Theorem <u>for model probab.</u>:

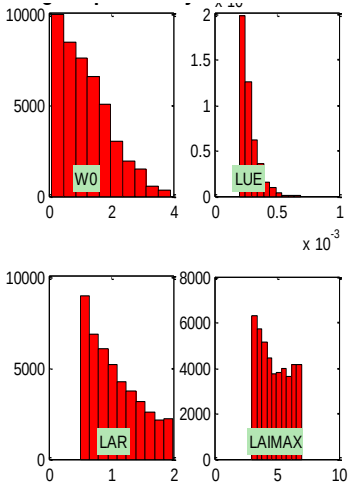
$P(M|D) = P(M)\ P(D|M) / P(D)$

If $P(M_1) = P(M_2) = ½$, then

$P(M_2|D) / P(M_1|D) = P(D|M_2) / P(D|M_1)$

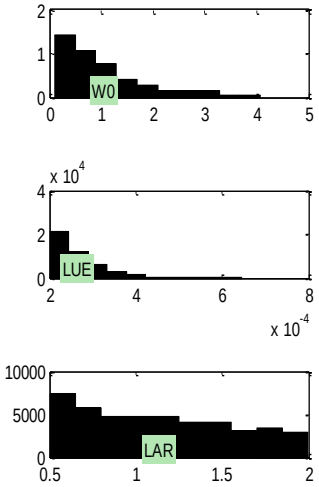
COMPLEX EXPOLINEAR

- **Constant light intensity**
- **Light int. by Beer's law**
- **Constant LUE**
- ***LAI reaches maximum***

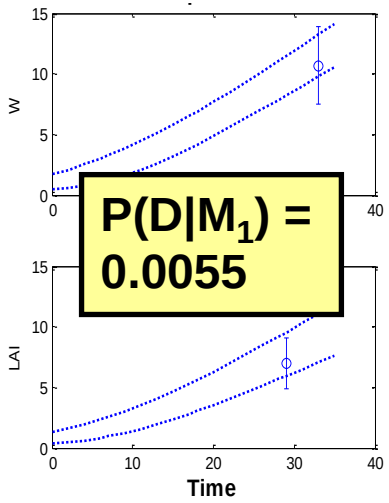


The "Bayes Factor" $P(D|M_2) / P(D|M_1)$ quantifies how the data D change the odds of M_2 over M_1

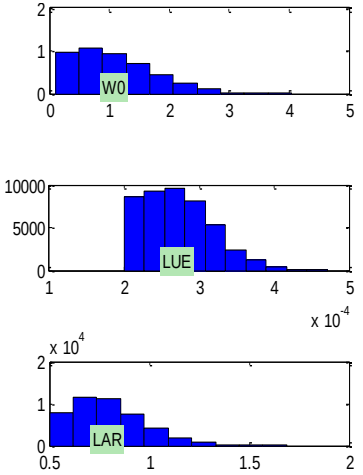
2.3 Bayesian comparison of two expolinear models



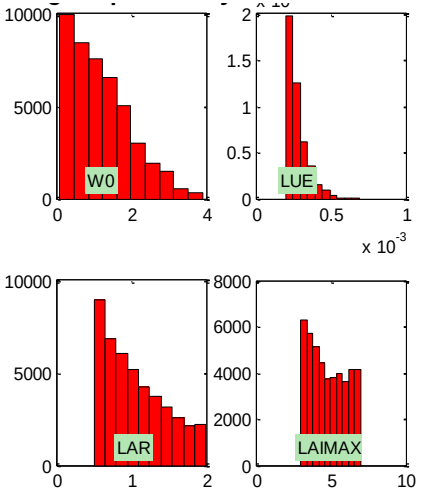
New data



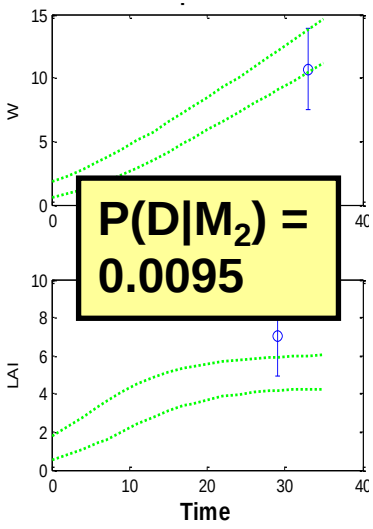
$P(D|M_1) = 0.0055$



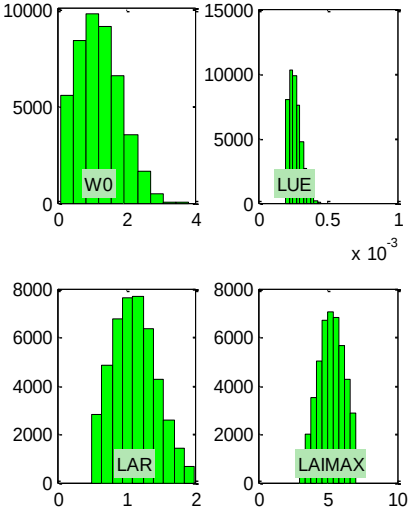
$BayesFactor_{21} = 0.0095/0.0055 = 1.74$



New data



$P(D|M_2) = 0.0095$



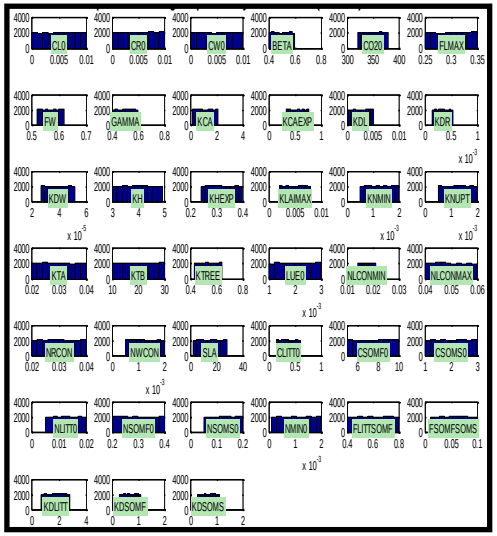
2.4 The Bayes Factor is sensitive to prior & data

	$BayesFactor_{21}$
Default priors & data:	**1.74**
LAI – data very precise (σ = 1 instead of 2.3):	**0.51**
Prior for LAIMAX "low" uniform(2,4) instead of (3,7):	**0.49**

3.
From 39 to 41 parameters

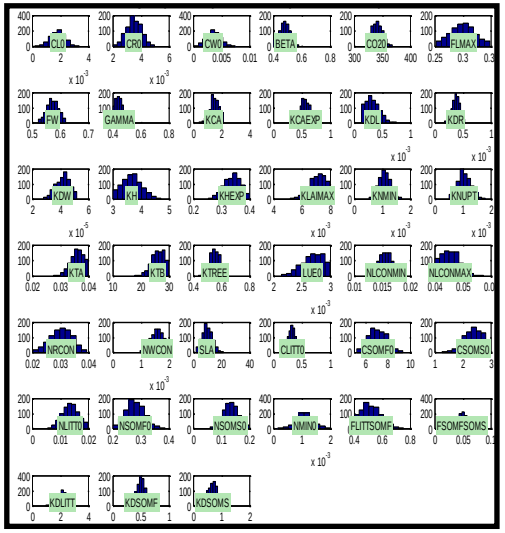
3.1 Bayes Factor for two big forest models

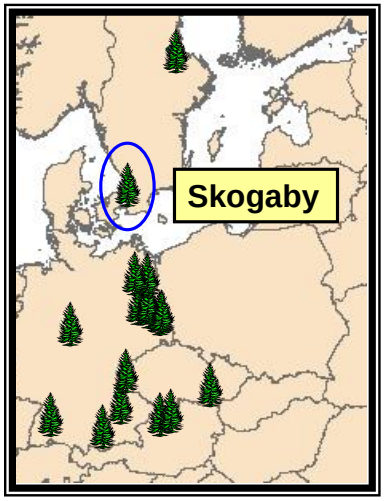
BASFOR
39 params



Calibration

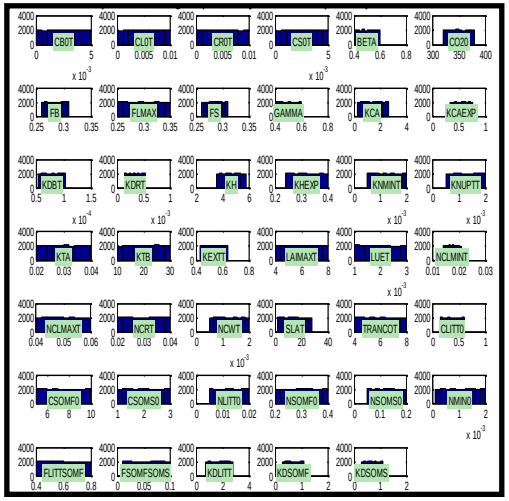
MCMC 10000 steps





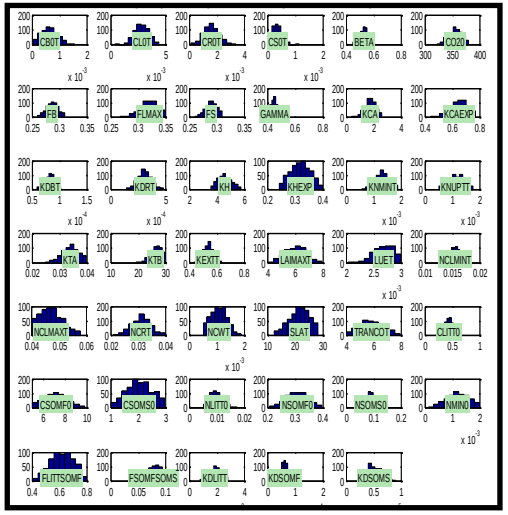
BASFOR +
41 params

(Penman eq., corrections)

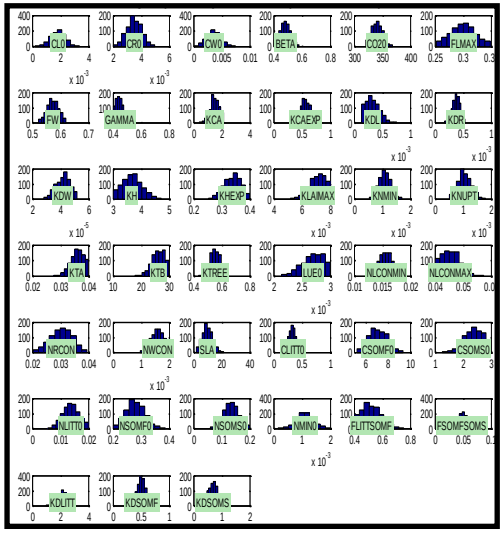


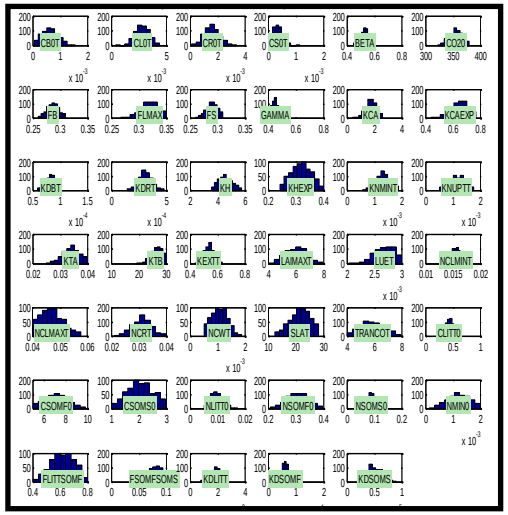
Calibration

MCMC 10000 steps



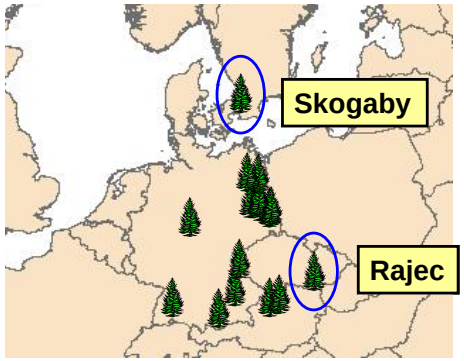
3.2 Bayes Factor for two big forest models





Calculation of P(D|BASFOR)

MCMC 5000 steps

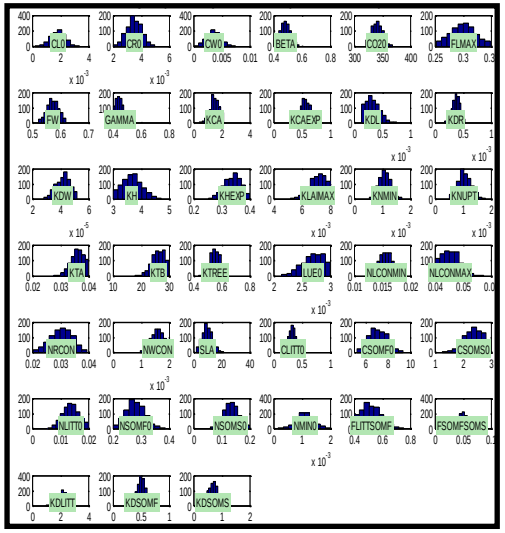


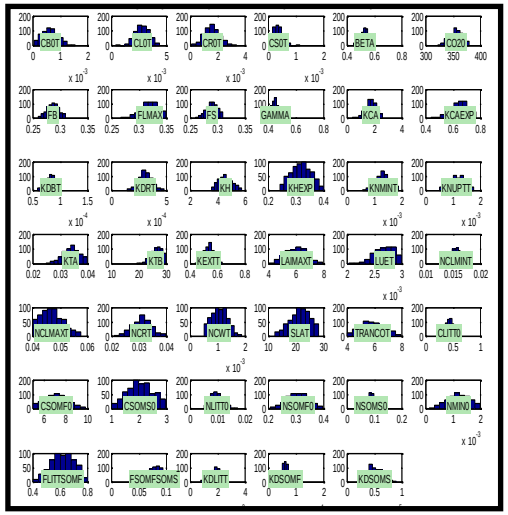
Data Rajec: Emil Klimo

Calculation of P(D|BASFOR+)

MCMC 5000 steps

3.3 Bayes Factor for two big forest models

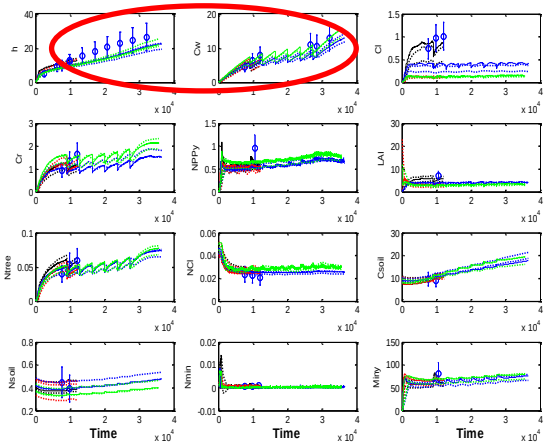




Calculation of P(D|BASFOR)

MCMC 5000 steps

$P(D|M_1)$ = 7.2e-016



Data Rajec: Emil Klimo

Bayes Factor = 7.8, so BASFOR+ supported by the data

Calculation of P(D|BASFOR+)

MCMC 5000 steps

$P(D|M_2)$ = 5.8e-15

4. Approximation methods for BMC: Information criteria

4.1 DIC introduced by Spiegelhalter et al. (2002)

- Deviance Information Criterion, DIC = 'goodness of fit' + 'complexity'
- Fit is measured via the deviance:

  $$\text{Dev}(\theta) = -2 \log p(\text{data}|\theta)$$

- Complexity measured by estimate of the 'effective number of parameters':

  $$p_D = E_{\theta|y}[\text{Dev}] - \text{Dev}(E_{\theta|y}[\theta])$$
  $$= \overline{\text{Dev}} - \text{Dev}(\bar{\theta});$$

  i.e. posterior mean deviance minus deviance evaluated at the posterior mean of the parameters

- The DIC is then defined analagously to AIC (which is just DIC with a flat prior) as

  $$\text{DIC} = \overline{\text{Dev}}(\theta) + 2p_D$$
  $$= \text{Dev} + p_D$$

- Models with smaller DIC are better supported by the data

4.2 The Information criteria zoo

- **AIC**
- **BIC**
- **DIC**
- **TIC**
- **QAIC**
- **WAIC**

Each criterion uses data to quantify the average out-of-sample prediction error for a model ... ... <u>without having new independent data</u>

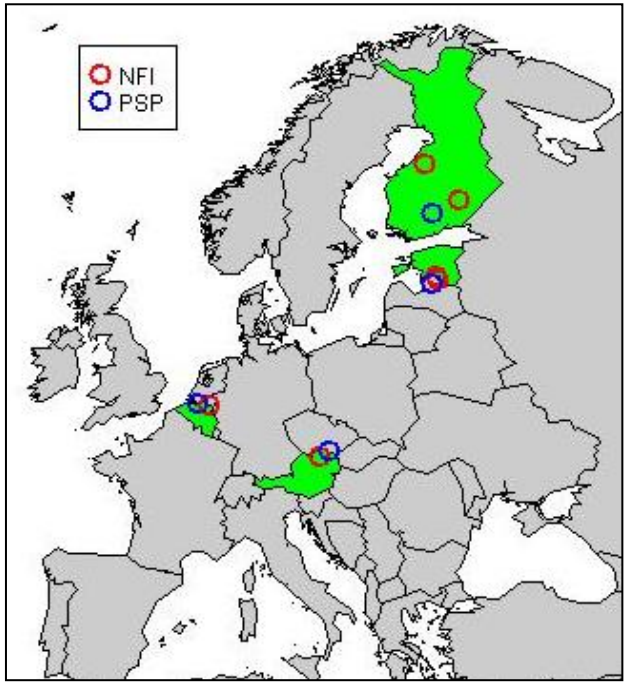
This is not possible!

But criteria can be converted into approximations of Bayesian posterior model probabilities (Burnham & Anderson 2002, Model Selection and Multimodel inference, Ch. 6.4.5)

If you have many data, the criteria as well as the Bayes Factor (Integrated Likelihood) tend to identify the same model as best, especially when the posterior is close to multivariate Gaussian

AIC $=-2 \ln[L_{max}] + 2n_{params}$; BIC $= -2 \ln[L_{max}] + \ln[n_{data}]*n_{params}$

5. CASE STUDY:
Calibration and comparison of six models for European forests



5.1 Protocol BMC COST 603 (24 Sep 2009)

Using Bayesian Model Comparison in COST 603

Marcel van Oijen, 24-9-2009

Bayesian Model Comparison (BMC) is a formal method for comparing models. It tells us how plausible different models are in the light of new data. In COST 603, we carry out a BMC of forest models. This document describes a very simple way to do a BMC, as an introduction to the method. The underlying theory is not presented, but references to introductory literature are given.

To do a BMC we need three things:

1. **A set of models M,**
2. **For each model, a probability distribution for its parameters, P(θ),**
3. **A collection of data D.**

Strictly speaking, we also need a prior probability distribution for the models themselves. But we are going to assume that initially, before using the data D, all models are equally plausible. That simplifies the BMC.

We use the following notation:

- We write the set of models as M = {M_1, ..., M_m}, where *m* is the total number of models. For example, M_1 could be a complex process-based model, M_2 a simple growth curve, and so on.
- We use the symbol θ to indicate the parameter vector of a model. So if a model has *n* different parameters, we can write θ = {θ(1), ..., θ(*n*)}. The value of *n* may be different for each model.
- We write $P_1(\theta)$ for the probability distribution for the parameters of M_1, $P_2(\theta)$ for the parameter distribution for M_2, and so on.
- Say we have *k* different points in our data set D. Each point has its own uncertainty, so we write D = {D_i, σ_i}, with i = 1…*k*.
- The aim of the BMC is to quantify the plausibility of each model in the light of the data. We denote these as P(M_j|D), or "the probability for model M_j given the data".

A relatively simple way to do the BMC, i.e. quantify the P(M_j|D), is the following. The first five steps need to be done for every model individually.

1. **Choose the uniform probability distribution as the prior distribution for the parameters. This means that for every parameter in the model we define a lower and upper bound.**
2. **Take a random sample of, say, 1000 parameter vectors from the prior.**
3. **Run the model for each of the 1000 parameter vectors.**
4. **For each of the 1000 sets of model outputs {M(θ)}, calculate the likelihood L of the data, assuming that all data points have independent measurement errors. We use the Gaussian probability distribution to describe our uncertainty about measurement errors. So:**

$$\mathbf{L} = \prod_{i=1,k} \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left[-\frac{1}{2} \left(\frac{M(\theta)_i - D_i}{\sigma_i} \right)^2 \right]$$

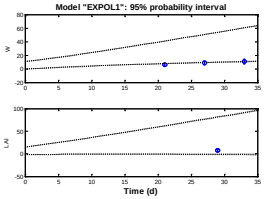
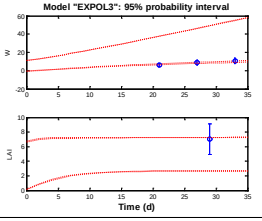
5. **Calculate the average of the 1000 likelihoods, $\overline{L}$.**
6. **After having calculated the average (also called 'integrated') likelihood $\overline{L}_j$ for each model M_j, we calculate the posterior probability for each model as:**

$$\mathbf{P(M_j|D)} = \overline{L}_j \Big/ \sum_{j=1}^{m} \overline{L}_j$$

That concludes the BMC. Note that steps 1-3 are just standard forward propagation of parameter uncertainty to model output; BMC only adds steps 4-6. The different values of P(M_j|D) now tell us how much support the data give to every model. The table on the next page shows an example of the procedure using two models. More detail can be found in Van Oijen (2008) and a nice example can be found in Tuomi et al. (2008; Table 2).

Note that the six-step procedure for BMC just described is deliberately kept simple. **A more thorough BMC would not use uniform distributions for the parameters (Step 1), and the assumption of independent measurement errors with Gaussian distributions (Step 4) is also rarely satisfied.** However, if we do make these quite common simplifying assumptions, the procedure is formally correct and may be considered a first BMC that can be improved upon in further discussions between modellers and data-providers.

An example of Bayesian Model Comparison.
Two models were used: EXPOL1 and EXPOL3.
The data consisted of four points D_i: tree biomass was measured at age 21, 27, 33 years (D_1=6.09, D_2=8.8, D_3=10.66 kg m^{-2}), and LAI was measured at age 29 years (D_4=7.2 m^2 m^{-2}). Measurement uncertainties were 30% for each of the four points.

Step	EXPOL1	EXPOL3
1	Three parameters, with different lower and upper bounds: Initial weight, W0: [0.1, 10] Light-Use Efficiency, LUE: [0.0002, 0.0010] Leaf Area Ratio, LAR: [0.5, 2]	Four parameters, with different lower and upper bounds: Initial weight, W0: [0.1, 10] Light-Use Efficiency, LUE: [0.0002, 0.0010] Leaf Area Ratio, LAR: [0.5, 2] Maximum LAI, LAIMAX: [3, 7]
2	Sample of 1000 parameter vectors {W0,LUE,LAR}: θ^1 = {9.4997,0.0007,1.739} θ^2 = {2.0101,0.0005,1.6150} … θ^{1000} = {1.9327,0.0009,0.5990}	Sample of 1000 parameter vectors {W0,LUE,LAR,LAIMAX}: θ^1 = {5.0604,0.0008,1.2761,6.4874} θ^2 = {9.6901,0.0003,1.8515,6.4899} … θ^{1000} = {3.3605,0.0009,1.5928,6.9219}
3	Model outputs for the 1000 parameter vectors: M(θ^1) = {32.5709, 39.1628, 45.7547, 71.9252} M(θ^2) = {18.1247, 22.8332, 27.5417, 39.4104} … M(θ^{1000}) = {28.2581, 36.7326, 45.2078, 23.6950} 	Model outputs for the 1000 parameter vectors: M(θ^1) = {30.3278, 37.5675, 44.8071, 6.4874} M(θ^2) = {19.1925, 21.9075, 24.6225, 6.4899} … M(θ^{1000}) = {31.9731, 40.1823, 48.3914, 6.9219} 
4	Likelihoods for each of the 1000 sets of model output: L(θ^1) = exp(-746.8184) L(θ^2) = exp(-180.6454) … L(θ^{1000}) = exp(-199.4562)	Likelihoods for each of the 1000 sets of model output: L(θ^1) = exp(-198.9252) L(θ^2) = exp(-51.6460) … L(θ^{1000}) = exp(-253.4870)
5	Integrated Likelihood: $\overline{L}$ = 1.4981e-006	Integrated Likelihood: $\overline{L}$ = 1.0335e-005
6	P(M_1|D) = 1.4981e-006 / (1.4981e-006 + 1.0335e-005) = 0.1266	P(M_2|D) = 1.0335e-005 / (1.4981e-006 + 1.0335e-005) = 0.8734

Conclusion: given the currently available information (data D), model 2 (EXPOL3) has a seven times higher probability of being correct than model 1 (EXPOL1).

References

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

Van Oijen (2008). Bayesian Calibration (BC) and Bayesian Model Comparison (BMC) of process-based models: Theory, implementation and guidelines. NERC/Centre for Ecology & Hydrology, 16 pp. http://nora.nerc.ac.uk/6087/

5.2 Collaborators on the case-study

Forest Ecology and Management 289 (2013) 255–268

Contents lists available at SciVerse ScienceDirect

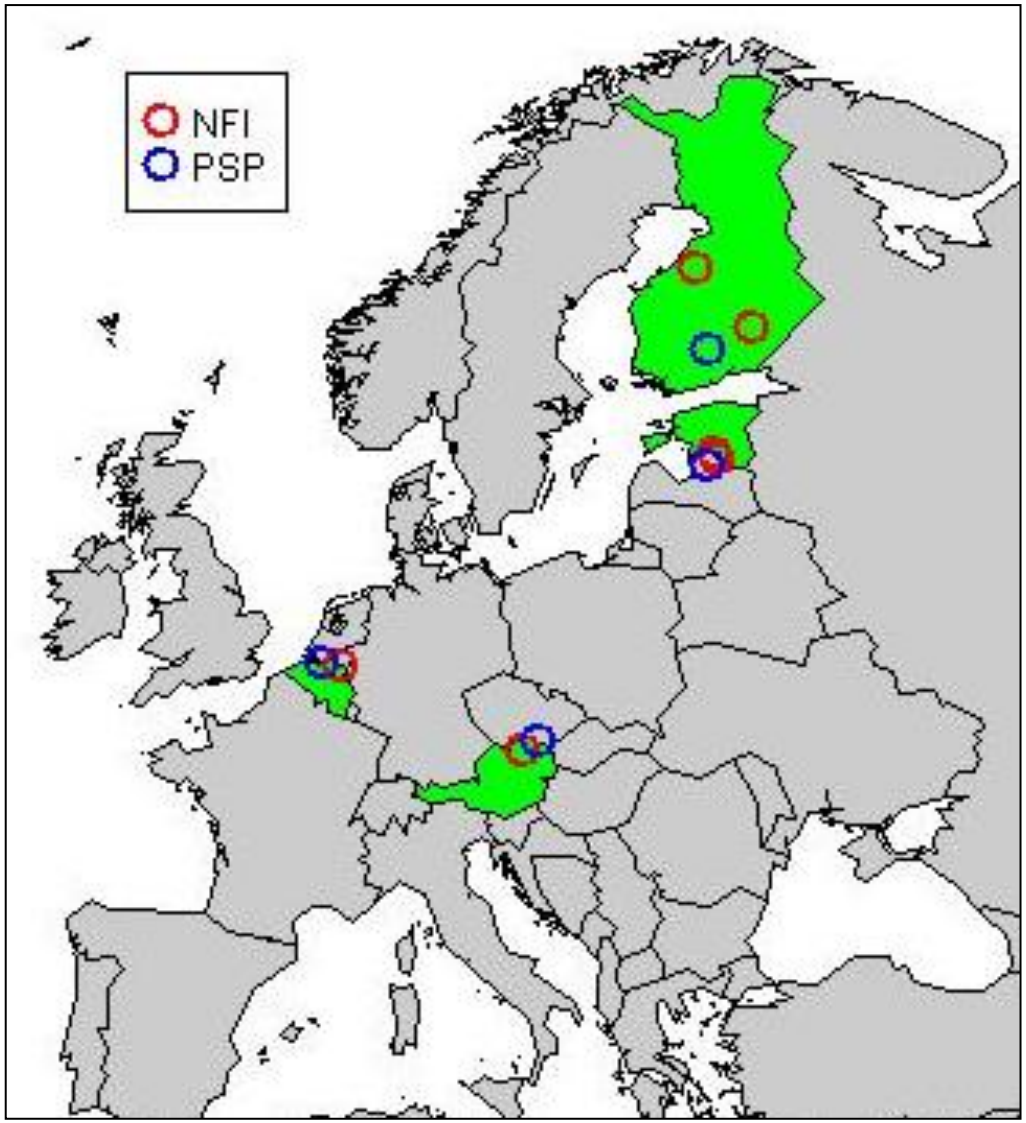
Forest Ecology and Management

journal homepage: www.elsevier.com/locate/foreco

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

M. van Oijen [a,*], C. Reyer [b], F.J. Bohn [c], D.R. Cameron [a], G. Deckmyn [d], M. Flechsig [b], S. Härkönen [e], F. Hartig [c], A. Huth [c], A. Kiviste [f], P. Lasch [b], A. Mäkelä [g], T. Mette [h], F. Minunno [i], W. Rammer [j]

5.3 Data from12 Scots pine plots



MEASUREMENT SITES:

- 8 NFI
- 4 PSP

ENVIRONMENTAL DATA:

- Soil
- Climate
- Atmospheric N-deposition
- Forest management

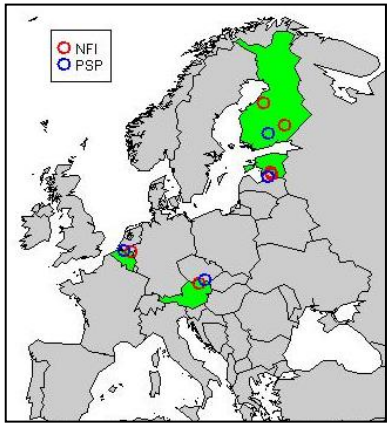
TREE DATA:

- Tree height
- Stem diameter

5.4 Six forest models

Model	# parameters In calibration	Initialisation
3PG	48	Planting
4C	43	First measurement
ANAFORE	138	First measurement
BASFOR	41	Planting
BRIDGING	13	First measurement
FORMIND	4	First measurement

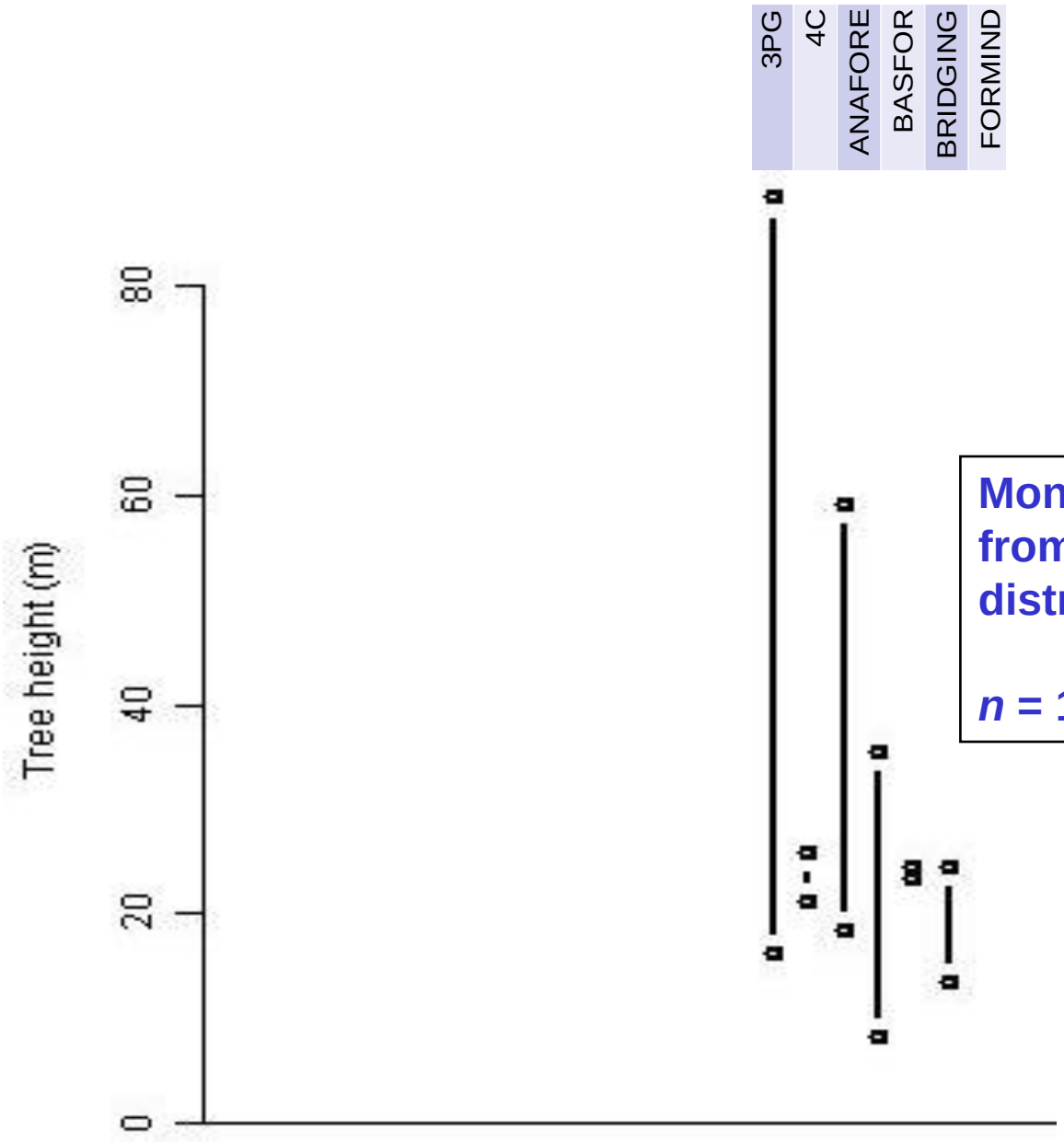
5.5 Data + Models



Model	# parameters In calibration	Initialisation
3PG	48	Planting
4C	43	First measurement
ANAFORE	138	First measurement
BASFOR	41	Planting
BRIDGING	13	First measurement
FORMIND	4	First measurement

- Can NFI-data reduce uncertainty about model parameters and outputs?
- Can models be calibrated generically, i.e. for Europe-wide application? Or is country-specific calibration better?
- Can NFI- and PSP-data help identify the most plausible forest model?
- What makes models good predictors?

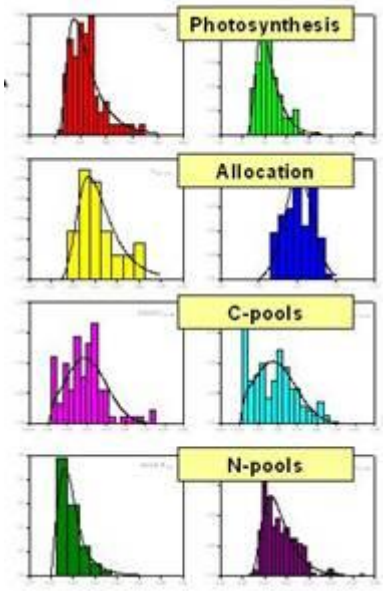
5.6 Prior prediction of final tree height in PSP-Estonia

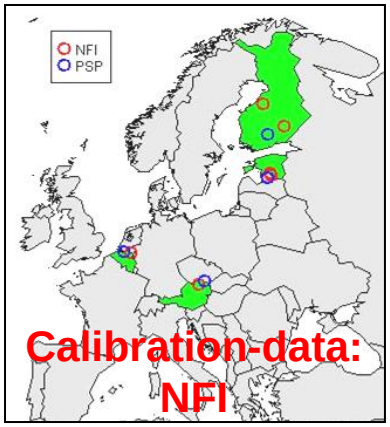


Monte Carlo sampling from prior parameter distributions

***n* = 1001**

5.7 Reducing parameter uncertainty by BC





Prior pdf for the parameters

Likelihood of the data

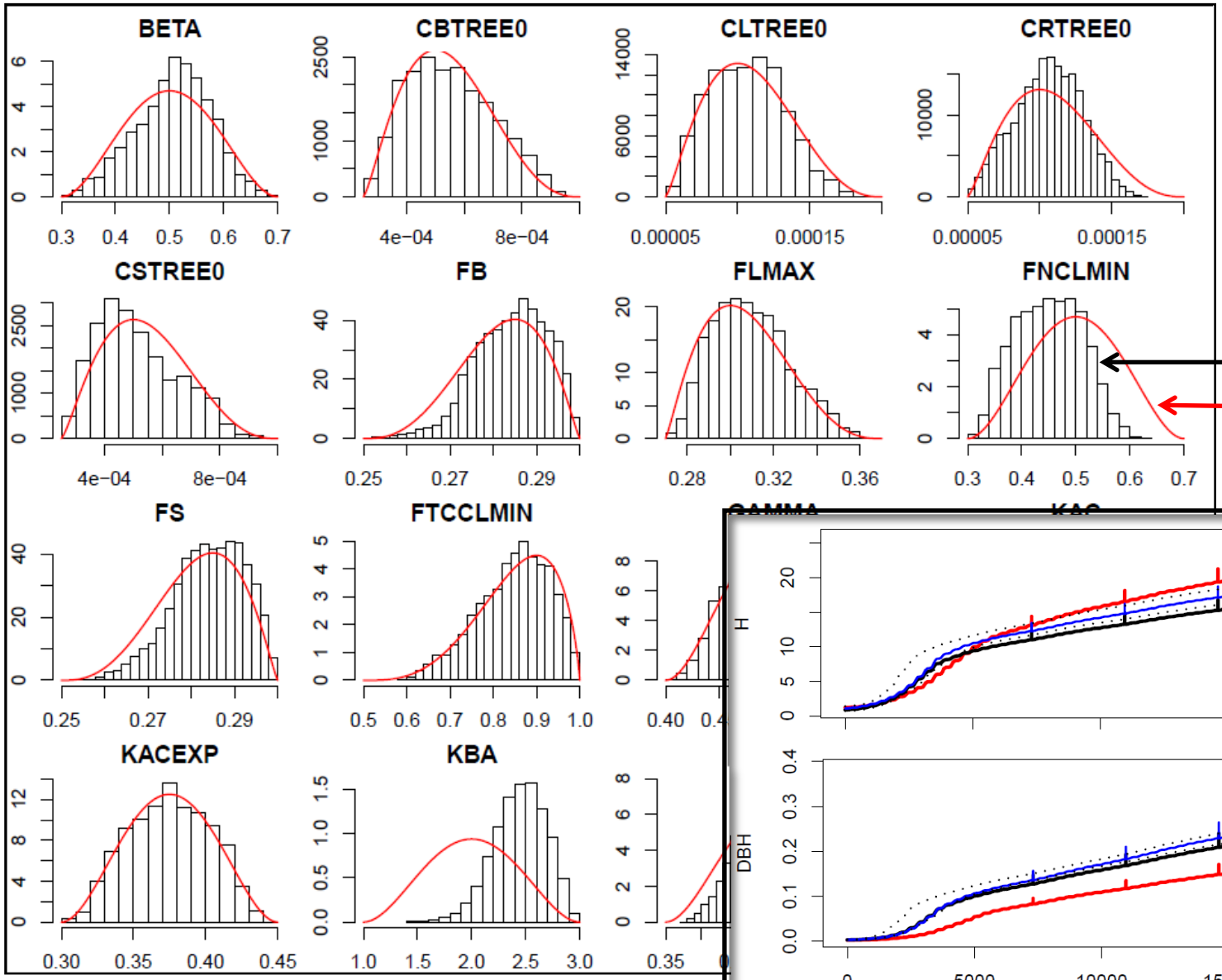
Bayes' Theorem

$$P(\vartheta|D) = P(\vartheta)\, P(D|\vartheta) / P(D)$$

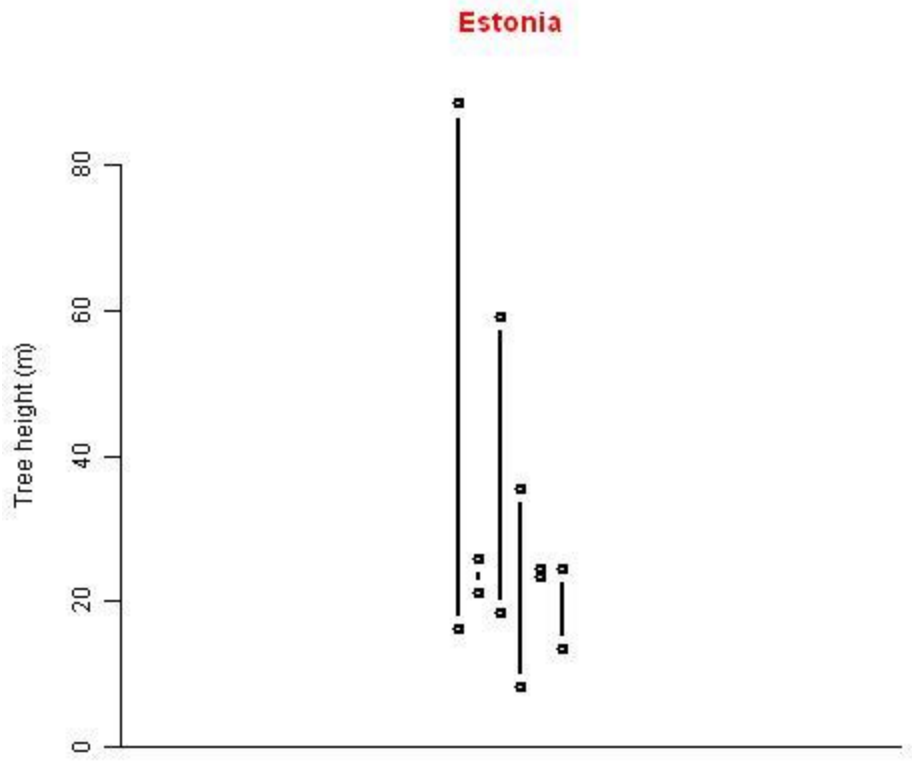
Posterior pdf for the parameters

Integrated likelihood
(= $\int P(\vartheta)\, P(D|\vartheta)\, d\vartheta$)

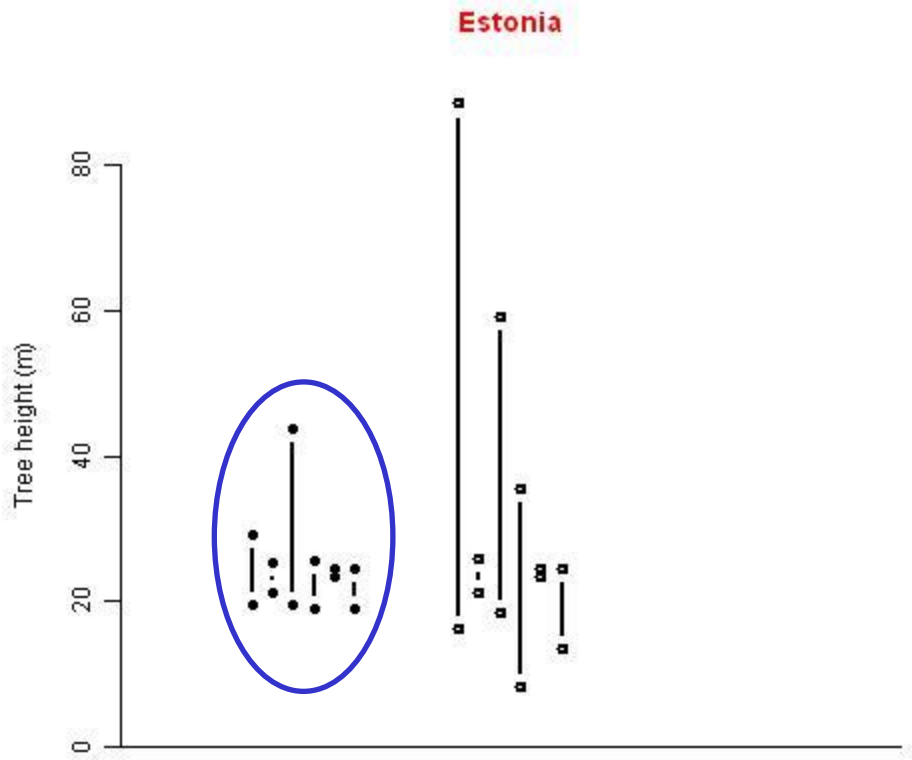
5.8 BC of model BASFOR on NFI-data



5.9 Uncertainty: Prior

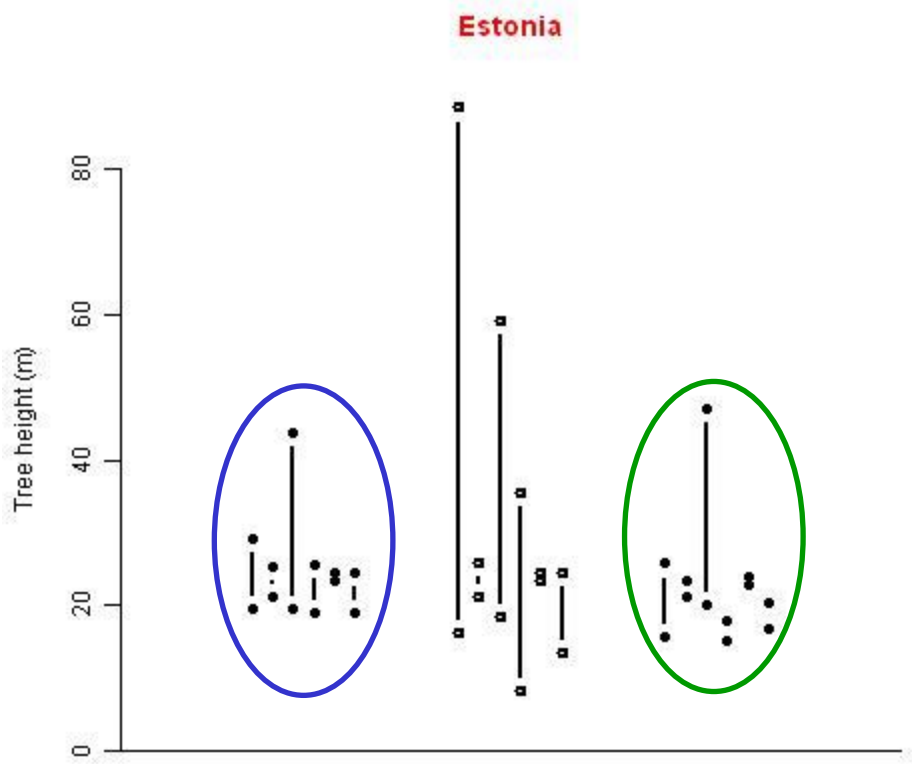


5.10 Uncertainty: Posterior$_{country}$ | Prior



After country-specific Bayesian calibration on NFI-data

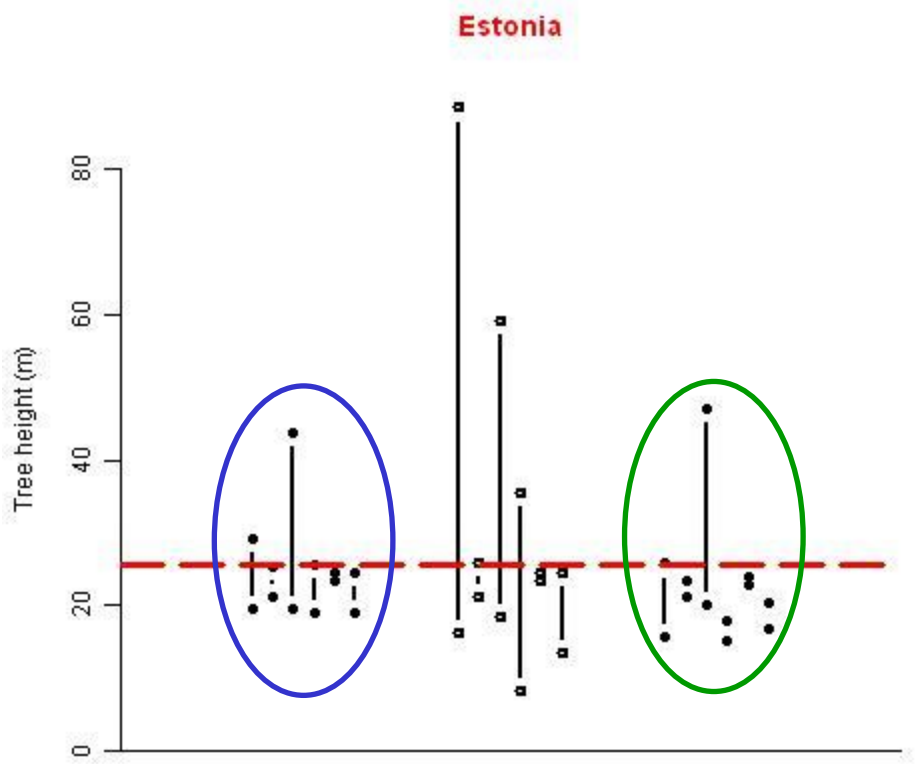
5.11 Uncertainty: Posterior$_{country}$ | Prior | Posterior$_{generic}$



After country-specific Bayesian calibration on NFI-data

After generic Bayesian calibration on NFI-data

5.11 Uncertainty: Posterior$_{country}$ | Prior | Posterior$_{generic}$

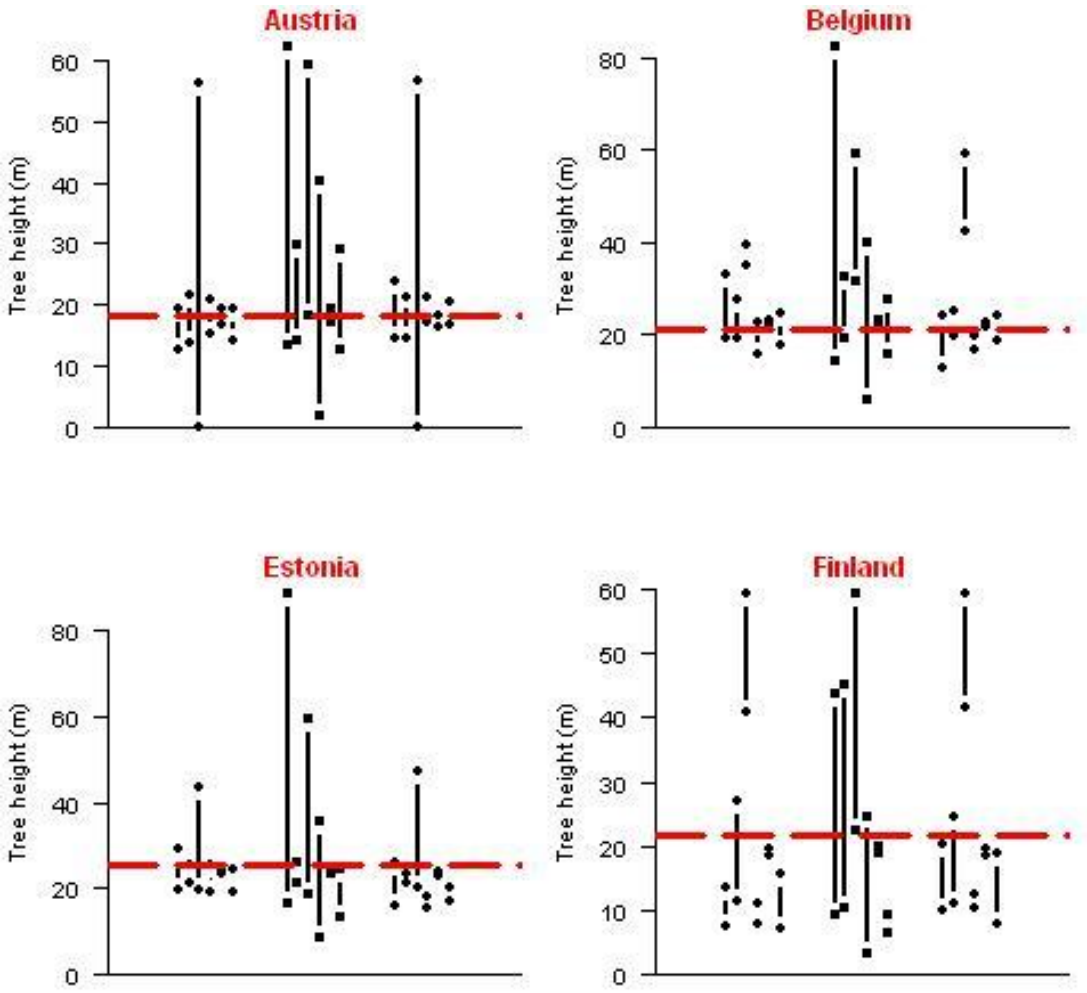


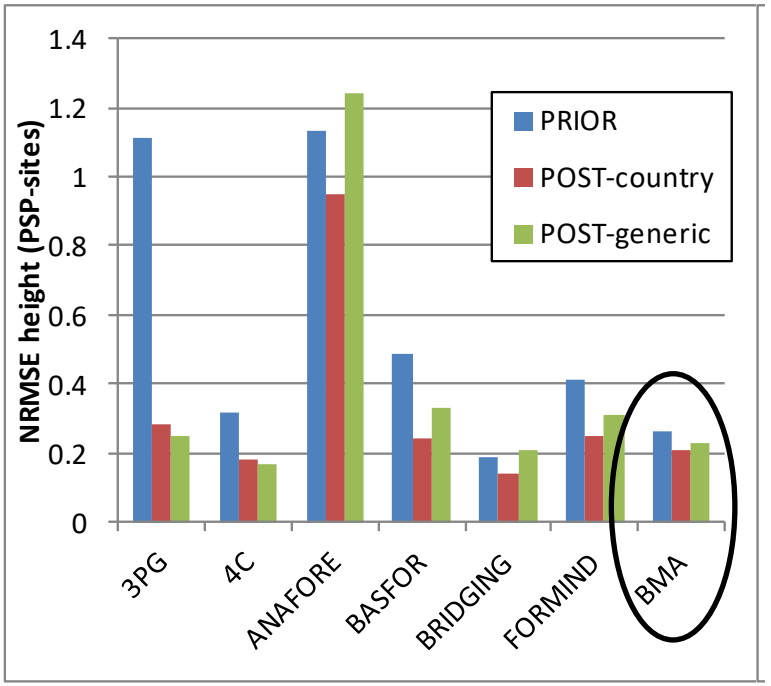
Measured value at PSP (not used in calibration)

After country-specific Bayesian calibration on NFI-data

After generic Bayesian calibration on NFI-data

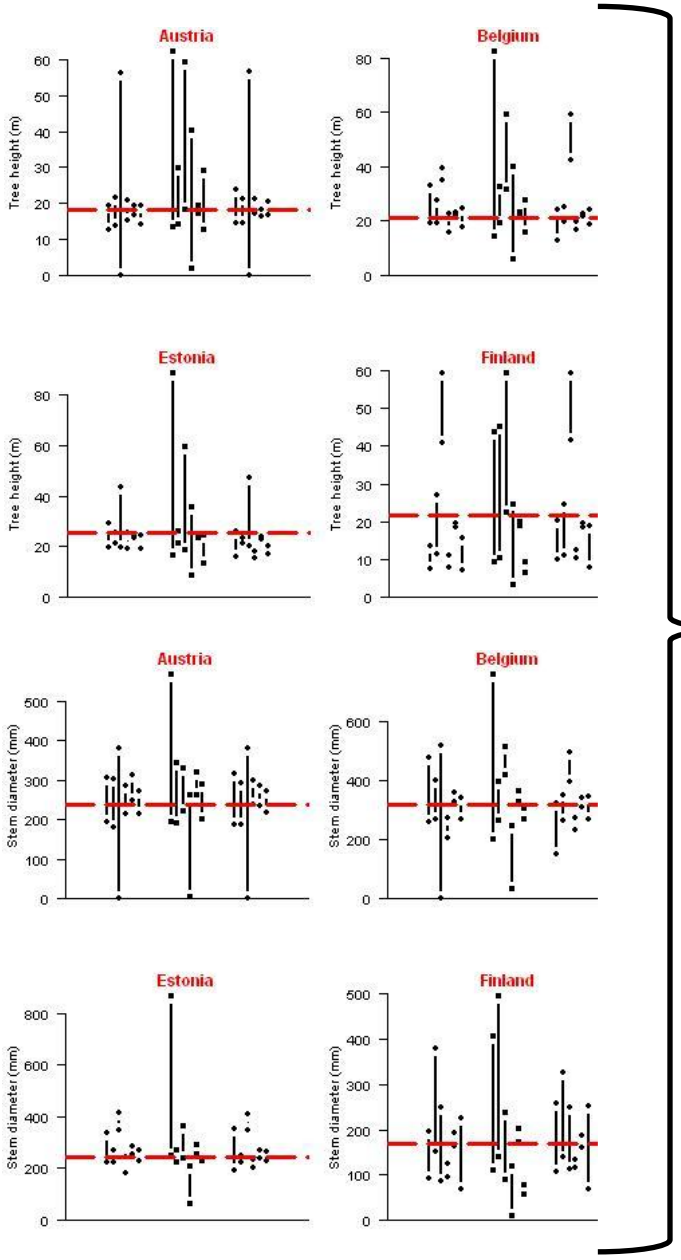
5.12 Posterior-Prior-Posterior: HEIGHT at PSP's



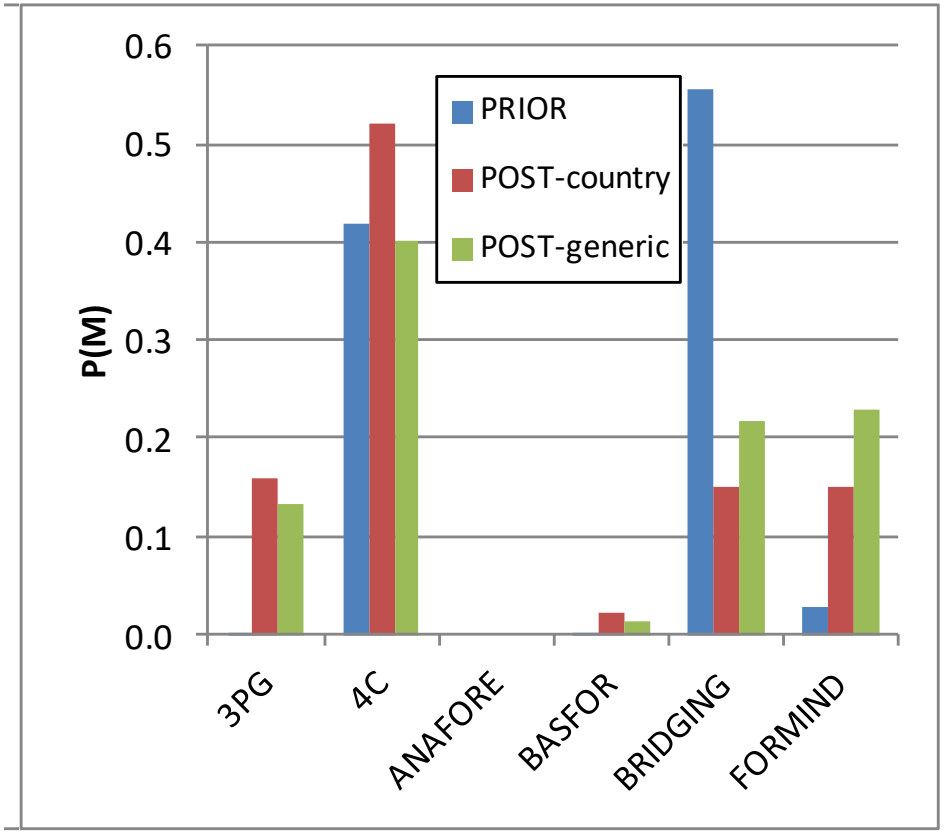


Bayesian Model Averaging (BMA)

5.13 Bayesian model comparison (BMC)



Model probabilities



BMA:

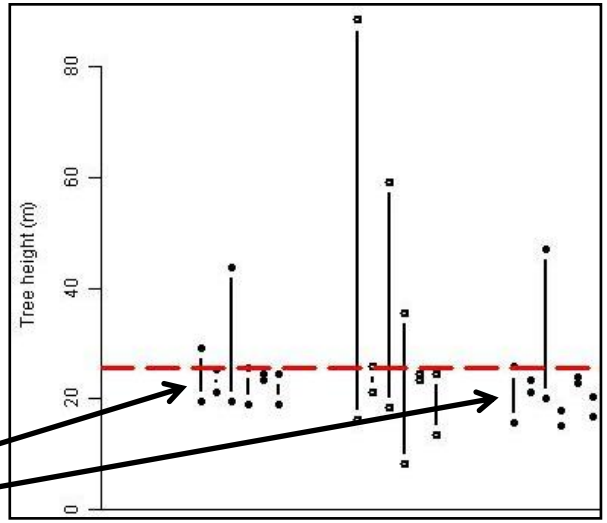
$$P(y) = \sum P(M)\, P(y|M)$$

$$= \sum P(M) \int P(\theta_M)\, P(y|M,\theta_M)\, d\theta_M$$

5.14 Conclusions from the case study

- Can NFI-data reduce uncertainty about model parameters and outputs?

YES

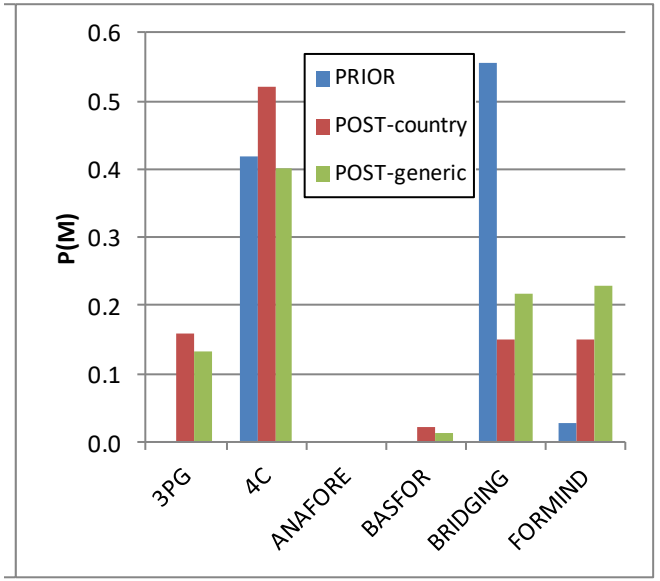


- Can models be calibrated generically, i.e. for Europe-wide application? Or is country-specific calibration better?

SIMILAR

- Can NFI- and PSP-data help identify the most plausible forest model?

YES



- What makes models good predictors?

Model	# parameters In calibration	Initialisation
3PG	48	Planting
4C	43	First measurement
ANAFORE	138	First measurement
BASFOR	41	Planting
BRIDGING	13	First measurement
FORMIND	4	First measurement

FEW PARAMETERS, LATE-INITIALISATION

Forest Ecol. Man. 289 (2013): 255-268

6. BMC: General discussion

6.1 What can BC & BMC not do?

• BC tells us about our parameters: what their values probably are
• BMC tells us about the structure of our models: which model is more plausible than others.

But …

• BC does not tell us why the most probable parameter values sometimes look strange
• BMC does not tell us whether the most plausible model could be improved, or how.

6.2 Evaluating the state of the art (1/2)

- Models cannot be derived from or evaluated against first principles.
- Model performance depends on the spatiotemporal scale. Every model has its own 'believable scale' outside of which it works poorly.
  - Small-scale processes are nonlinear and input variables are spatially heterogeneous, so a model structure that functions well at point-support cannot function well at block-support, and vice versa.
- Model performance cannot be evaluated without specifying our knowledge about parameter values.
  - There is always uncertainty about a model's parameters (θ).
  - So we can only evaluate the combination of model + probability distribution for the parameters, $p(\theta)$.

6.3 Evaluating the state of the art (2/2)

- Models for explanation can be simpler than models for prediction, because only the latter need to work well under both existing and new conditions.
- But simpler models are easier to parameterise and more robust than complex models.
- The repertoire of dynamical behaviour that a model produces depends less on the precise form of its equations than on its feedback-structure.
  - Feedbacks have different time constants. Some feedbacks operate slowly (e.g. soil-mediated ones), others fast (e.g. plant physiology). The slow ones are key to long-term prediction but cannot be measured readily.
- Model evaluation must be against a combination of data + competing models. Without competing models, evaluation is pointless. Without data, evaluation is academic.
  - Uncertainty about model structure requires the use of model ensembles.
  - Model ensembles share biases which need to be estimated probabilistically.