Application of Bayesian calibration for propagation of uncertainty in dynamic models

Anvendelse av Bayesiansk kalibrering for forplantning av usikkerhet i dynamiske modeller

Philosophiae Doctor (PhD) Thesis

Anne-Grete Roer

Dept. of Chemistry, Biotechnology and Food Science
Norwegian University of Life Sciences

Ås 2010

Thesis number 2010: 8
ISSN 1503-1667
ISBN 978-82-575-0920-0

To my family: Inge and Cornelia

Preface

The work present in this thesis is submitted to the Norwegian University of Life Sciences (UMB). The thesis is based on work funded by the Norwegian Institute of Agricultural and Environmental Research (Bioforsk).

Initially I will express my gratitude to my supervisors who have supported me during the whole study, Dr. Trygve Almøy at UMB and Dr. Trond Rafoss at Bioforsk. You have always been receptive for questions and discussions, and your guidance and comments on the written papers have been of great value.

I will thank Dr. Marcel van Oijen at Centre for Ecology & Hydrology in Scotland for being real helpful and for answering questions. I am deeply thankful for your experiences and advices within the Bayesian framework. I will thank Dr. Solve Sæbø at UMB for all valuable help on statistical challenges, both within the classical frequentistic and the Bayesian approach.

I will thank my fellow PhD-student, Stig Morten Thorsen for good cooperation and interesting discussions during the work.

I want to express my gratitude of having had the opportunity to work with statistical issues within plant science. I am thankful for all knowledge and experiences it has given me.

Anne-Grete Roer

Ås, January 2010

Contents

Individual papers I-IV

Abstract

The thesis is about quantification of uncertainties in complex models. Models are built to describe, explain or predict a real world outcome. It is well known that models are related with uncertainty, and that uncertainties are related to how close the simulation is to the real world outcome. Still, uncertainties are rarely quantified in dynamic models. We have focused on parameter uncertainty and output uncertainty derived from the parameters. Uncertainty originated from the empirical data is integrated into the posterior parameter distributions through the likelihood functions. Additionally, uncertainty related to the representativeness of the collected data to the population has been focused.

The Bayesian statistical framework, with the Markov chain Monte Carlo algorithm random walk Metropolis was used for model calibration in the four papers. The algorithm was found simple in idea and implementation into the computer program Matlab, but challenges emerged when the method was used at complex models. In this work these challenges have been pursued together with searching for efficiency improvements in order to make as few model evaluations as possible.

Paper I: explores the challenges emerging when applying Bayesian calibration to a complex deterministic dynamic model of snow depth. How prior information and new data affect the calibration process, the parameter estimates and model outputs were demonstrated. Parameter uncertainty and model uncertainty derived from the parameters were quantified, visualized and assessed. The random walk Metropolis algorithm was used and in order to reach convergence more effectively, informative priors, Sivias' likelihood, reflection at the prior boundaries and updating the proposal distribution with parts of the data gave successful results. Methods for objective and correct determination of Markov chain convergence were studied, and the use of multiple chains and the Gelman-Rubin method was found useful.

Paper II: presents a dynamic model for snow cover, soil frost and surface ice. The Bayesian approach was used for model calibration and sensitivity analysis identified the non-important parameters.

Paper III*:* shows the importance of splitting the data several times in two for model development and assessment/selection, for the model to fit well to novel data from the system and not only to the specific data at hand. Different models of ascospore maturity of *Venturia*

inaequalis were further developed and compared by the deviance information criterion and root mean square error of prediction to show model improvements, and the analysis of variance was used to show significance of the improvements.

Paper IV: examines the potential effects of selection of likelihood function when calibration a model. Since the likelihood function is rarely known for certain, but gives a reasonable quantification of how probable the data are given model outcome, it is of great importance to quantify the effect of using different likelihood functions on parameter uncertainty and on model output uncertainty derived from the parameters.

Sammendrag

(Norwegian summary)

Denne avhandlingen omhandler kvantifisering av usikkerhet i komplekse modeller. Modeller bygges for å beskrive, forklare og predikere virkelige systemer. Selv om det er kjent at usikkerhet er knyttet til modeller og dermed er relatert til hvor lik den simulerte og den virkelige verdien er, blir usikkerhet sjeldent kvantifisert i dynamiske modeller. Det har i denne avhandlingen blitt lagt vekt på parameterusikkerhet og usikkerhet i utgangsdata fra modeller med opprinnelse i parametrene og empiriske data. I tillegg har det blitt fokusert på usikkerhet i forhold til hvor godt de innsamlede observasjonene representerer populasjonen.

Bayesiansk statistikk har blitt anvendt til å kalibrere modellene i de fire påfølgende artiklene ved hjelp av Markov chain Monte Carlo algoritmen random walk Metropolis. Algoritmen er enkel å forstå og å implementere i dataprogrammet Matlab, men utfordringer oppstod ved anvendelse på komplekse modeller. Disse utfordringene og effektivitetsforbedringer ved å minimere antall modellevalueringer har blitt vektlagt.

Artikkel I: utforsker utfordringer ved bruk av Bayesiansk kalibrering på en kompleks deterministisk dynamisk modell for snødybde. Fokus er lagt på hvordan den opprinnelige usikkerheten, a'priori usikkerheten, sammen med nye innsamlede data gjennom rimelighetsfunksjonen påvirker kalibreringsprosessen, parameter estimatene og modellens utgangsdata. Parameterusikkerhet og modellusikkerhet grunnet parametrene ble kvantifisert, vist og vurdert. Random walk Metropolis algoritmen ble anvendt, og for å oppnå konvergens raskere ble informative fordelinger på parametrene, Sivias' rimelighetsfunksjon, refleksjon og å oppdatere forslagsfordelingen med deler av dataene testet med gode resultater. Det ble dessuten lagt vekt på viktigheten av en metode for å avgjøre både objektivt og korrekt når kjedene konvergerte, hvor parallelle kjeder og Gelman-Rubins metode ble funnet nyttig.

Artikkel II: presenterer en dynamisk modell for snødybde, frostdybde og overflate-is. Bayesiansk rammeverk ble anvendt for å kalibrere modellen og sensitivitetsanalyse identifiserte de mindre viktige parametrene.

Artikkel III: viser hvor viktig det er å dele data flere ganger i to for modell utvikling og modell validering for at modellen ikke kun skal passe de spesifikke dataene, men også nye data fra det samme systemet. Ulike modeller for sporemodning av *Venturia inaequalus* ble

videreutviklet og sammenlignet ved bruk av kriteriene devianse informasjons kriteri (DIC) og prediksjonsfeil (RMSEP) for å vise modellforbedringer. Variansanalyse ble anvendt for å angi statistisk signifikans til forbedringene.

Artikkel IV: undersøker effekten av rimelighetsfunksjonen på en snødybdemodell. Siden rimelighetsfunksjonen sjelden er kjent, men kun gir en fornuftig kvantifisering av hvor sannsynlig data er gitt modellens utdata, er det viktig å kvantifisere effekten av å anvende ulike rimelighetsfunksjoner på parameterusikkerhet og på usikkerheten relatert til modellens utgangsdata med opprinnelse i apriori parameterusikkerhet og empiriske data.

List of papers

I. Roer, A. G., Thorsen, S. M., Rafoss, T., van Oijen, M. and Almøy, T: Fine-tuning Bayesian calibration for complex systems with application to a snow depth model (manuscript)

II. Thorsen, S. M., Roer, A. G. and van Oijen, M (2010) Modelling the dynamics of snow cover, soil frost and surface ice in Norwegian grassland (Polar Research 29:110-126)

III. Roer, A. G., Eikemo, H., Stensvand, A., Almøy, T., Creemers, P. and Rafoss, T: Modelling spore maturation in a Bayesian framework with application to spore release of *Venturia Inaequalis* (manuscrript)

IV. Roer, A. G., Almøy, T. and Rafoss, T: The influence of the likelihood function in Bayesian calibration to a snow depth model (manuscript)

Introduction

Background

Computer models are abstract representations of real world systems. They are made in different complexities and are constructed in place of conducting experiments either because it is found more convenient, more efficient or because it can not at all be done in reality. They usually take a set of state variables and unknown or uncertain values (parameters) as inputs, to generate simulations of the real world outcome. The models are built to describe, explain or predict a real world system, to generate new insight in the system and about the response. Unfortunately, there will always be a difference between the prediction and the observation. This error is translated to uncertainty in conclusions drawn from the model. According to Goldstein and Rougier (2006), there are three sources of uncertainty

1. Models often contain parameters whose values are not known
2. Models contain simplifications of the real world system
3. Data involved are induced by measurement error

Measurement error includes the possibility that the specific data at hand do not represent the population (Bøvelstad et al. 2007).

Models are generated to make statements about their underlying real world system. In some cases model prediction will not be tested until several decades or a century ahead (e.g. models related to climate change). Since uncertainties are related to the models, the question of how to learn about the actual physical system through the model becomes important (e.i. provide predictions that are transparent with respect to uncertainty (Thyer et al. 2009)). Still uncertainties are rarely quantified and conclusions are usually made conditional on the model being correct.

In situations with substantial uncertainties, it is natural to adapt a Bayesian approach (Kennedy and O' Hagan 2001). The Bayesian framework automatically includes uncertainty quantification and makes conclusions conditional on the collected data. The use of the Bayesian framework to environmental models has been increasing in recent years (e.g. Reinds et al. 2008, Lehuger et al. 2008), and it has also been introduced to models within the agronomy science, such as plant pathology and epidemiology (Dunson 2001, Mila et al.

2003), where decision making models e.g. made to alert the farmer about when and where to spray fungicide on the crop has become more popular.

Main objectives

The aim of this work has been quantification of uncertainty and its propagation by use of the Bayesian framework. Different models related to plant protection were used as case studies. Specifically, the objectives were:

- Application of Bayesian calibration (learning about the model parameters using data from the system) to complex models in order to explore practical problems with the calibration as well as work out solutions.
- Uncertainty quantification including propagation of uncertainty, and assessment of whether the predictive uncertainty is over or under estimated.
- Model development and model selection in order to obtain a robust and reliable model.

Modelling

A model is a simplified abstract view of the complex reality of a system. In computer science, a simulator is the software program to model a real-life situation on a computer so that it can be studied to see how the system behaves. The simulator can typically be divided into three parts (Goldstein and Rougier 2009)

$$\text{Simulator} = \text{Model} + \text{Treatment} + \text{Solver}$$

Where the model usually evolves in time and space, and describes fundamental laws and equations of the state that variables exist. The treatment contains initial conditions that make the model applicable to a particular instance and the solver turns the model and the treatment into calculations that approximates the evolution of the state vector. Computer simulations have become a useful part of mathematical modeling of many natural systems in e.g. physics, astrophysics, chemistry and biology.

Computer models in biology are commonly classified into empirical or process based models. The empirical models are statistical oriented while the process based models are explicitly represented (mechanistically) by the known underlying processes (Radtke and Robinsom 2006).

To validate models, the best approach is to divide the dataset in two parts: a training set and a test set (Hastie et al. 2001). The training set is used for model development and test set used for model validation. We have generally used about 2/3 of the data for model development and 1/3 for model validation.

The nature is complex, and models searches to describe and capture the main interacting factors by simplifications of the reality. They are built in all kinds of complexity, and as Figure 1 shows, there is an optimal model complexity that gives minimum prediction error (the expected error over an independent test sample). The figure also shows that the training error (the average loss over the training samples) is not a good estimate of the prediction error. It will consistently decrease with model complexity and typically converge to zero if the model complexity is increased enough. Also, there is a bias-variance tradeoff in choosing the appropriate complexity of the model, where bias refers to squared directional error in an estimator and variance refers to squared random error.

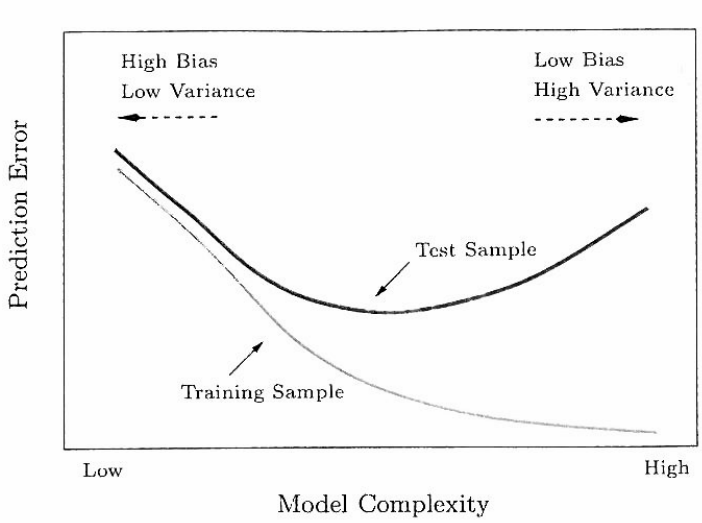


Figure 1: Behavior of test sample and training sample error as the model complexity is varied. The figure is taken from Hastie et al. (2001).

Statistical Issues

Classical versus Bayesian framework

Statistical science can be divided into two different philosophical directions; the classical approach by Fisher, and the Bayesian approach named after Thomas Bayes (Berger 1985) the founder of Bayes theorem. In the classical statistical approach, collected data are considered to be the only available source of information, and calibration is e.g. done using the maximum likelihood method or least sqare, giving point estimates for the parameters. Uncertainty may only be established through large sample arguments and they are addressed by putting confidence limits on the unknown parameters accordingly, from studying the frequency behavior of the parameter estimates (Frey and Burmaster 1999). The alternative Bayesian framework regards parameters as random variables that follow some probability distribution. Uncertainties are automatically included as probability distributions, and prior information allowed and formally incorporated through the prior distribution. Bayes theorem (Equation 1) is the building block and states that the posterior parameter distribution; the probability distribution of the parameters given the collected data ($\pi(\boldsymbol{\theta}|\boldsymbol{D})$, when $\boldsymbol{\theta}$ is the parameter vector and $\boldsymbol{D}$ data) which we seek in calibration, is a combination of prior parameter knowledge before data is collected, determined by the prior probability distribution ($\pi(\boldsymbol{\theta})$) and currently incorporated information through the likelihood function; the conditional probability density function of the collected data given the parameters, regarded as a function of the parameters ($L_D(\boldsymbol{\theta})$) of the collected data.

$$\pi(\boldsymbol{\theta}|\boldsymbol{D}) \propto \pi(\boldsymbol{\theta}) \cdot L_{\boldsymbol{D}}(\boldsymbol{\theta}) \quad (1)$$

The prior distribution

The prior parameter distribution reflects our parameter knowledge before the study takes place (Gelman 2002). It can be based on earlier studies, literature review or expert opinions, and if no prior information is available, uninformative priors can be used (Jeffrey 1961). Prior independence is assumed, and the joint distribution found as the product of the marginal distributions. In the following thesis, uniformly and beta distributed priors based on literature or expert opinions were used.

The Likelihood function

The likelihood function reflects the new incorporated information through collected data, and is rarely known for certain. It should then be a reasonable quantification of how probable the data are given the model output. For complex models the likelihood can be determined as the product of the likelihood of the individual data points, determined by the distribution of the errors (Rougier 2007). Both the Gaussian likelihood and a fat tailed Gaussian likelihood function (Sivia 2006) were used in this thesis. For binary outputs, the binomial distribution (Gelman et al. 1996b) was used.

The posterior distribution

The posterior distribution reflects out parameter uncertainty after updating our prior knowledge with new incorporated information through Bayes theorem. The prior will have larger effect on the posterior if the sample size is small or if the available data only provide indirect information about the parameters (Gelman 2002).

Calculations

There are four different techniques that can be used to find the posterior distribution in Bayesian calibration; exact calculation, analytical approximation, numerical integration and Monte Carlo simulation. Integration problems, especially when the parameter space is high dimensional, often make exact calculation impossible and it is therefore not used here.

In the following papers, the Markov chain Monte Carlo (MCMC) algorithm random walk Metropolis was used. It is an iterative algorithm, starting with a guessed parameter set $\boldsymbol{\theta}^0 = (\theta_1^0, \theta_2^0, ..., \theta_m^0)$ where m is the total number of unknown parameters. Then for each iteration step $i=1:N$

1. Draw a candidate parameter set $\boldsymbol{\theta}'$, from a spherically symmetric and independent distribution for different i, centered at the current state.
2. Compute the ratio

$$\alpha = \frac{\pi(\boldsymbol{\theta}'|\boldsymbol{D})}{\pi(\boldsymbol{\theta}^i|\boldsymbol{D})} = \frac{\pi(\boldsymbol{\theta}') \cdot L_{\boldsymbol{D}}(\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}^i) \cdot L_{\boldsymbol{D}}(\boldsymbol{\theta}^i)}$$

3. Draw a random number u from the standard uniform distribution and set

$$\boldsymbol{\theta}^{i+1} = \begin{cases} \boldsymbol{\theta}' & \text{if } u \leq \alpha \\ \boldsymbol{\theta}^i & \text{otherwise} \end{cases}$$

This results in a Markov chain of parameter sets, $\boldsymbol{\theta}^0, \boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^N$, and the idea is that the chain will in the long run converge to the posterior parameter distribution (Liu 2001). The stationary distribution of the Markov chain can then be regarded as samples from the posterior parameter distribution. To monitor convergence (burn-in) correctly is important, and if parts of the chain not converged is treated as samples from the posterior distribution, false conclusions may be produced. In the following papers, chains were run in parallel and burn-in detected by using the Gelman-Rubin diagnostic (Gelman and Rubin 1992) that compares the variability between and within the sequences.

Model assessment

Model selection in the Bayesian framework is often done using Bayes factor (BF) (Kass and Raftery 1995). The method determines which of a set of models is most probable in view of the data and prior information, and how strong it is supported relative to the alternative models in the set. It is a pairwise model selection criterion that is not effective when a high number of models are to be compared, and it is therefore not used in this thesis.

The Deviance information criterion (DIC) (Spiegelhalter et al. 2002) is a model comparison method, combining model fit and penalty on model complexity using the training data. It is defined by

$$DIC = \overline{D} + q_D \quad (2)$$

where $\overline{D}$ is the posterior mean of the deviance (quality of fit, calculated as -2 times the log-likelihood ratio of the reduced model compared to the full model, Agresti (2007)) and q_D is the estimated model complexity. The model with the lowest DIC is preferred.

Root mean square error of prediction (RMSEP) (Hastie et al. 2001) compares models and estimates the prediction errors.

$$RMSEP = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(D_t - M_t)^2} \quad (3)$$

Where T is the number of test samples, D_t is the observed value at time t and M_t is predicted model output at time t.

Uncertainty assessment

There are many ways to visualize and assess uncertainty in models and parameters. Three methods will be present here.

The 100(1-α) % highest posterior density (HPD) credible set (Berger 1985) is a measure for posterior parameter uncertainty. It is the subset C of the parameter space Θ of the form

$$C = \{\theta \in \Theta: \pi(\theta|x) \geq k(\alpha)\}, \quad (4)$$

where $k(\alpha)$ is the largest constant such that $P(C|x) \geq 1-\alpha$.

Root mean square deviation (RMSD) (Iizumi et al. 2009) calculates the model output uncertainty derived from the parameters, and is defined by

$$RMSD = \sqrt{\frac{1}{TE}\sum_{t=1}^{T}\sum_{e=1}^{E}(M_{et} - M_{\cdot t})^2} \quad (5)$$

where T is the number of test samples, E is the number of parameter sets used (ensembles), M_{et} is model output at time t using parameter set number e, and $\overline{M_{\cdot t}}$ is the mean model estimate at time t.

Predictive QQ plot (Dawid 1984, Thyer et al. 2009) assess whether the predictive uncertainty is consistent with the observed data. By comparing the empirical cumulative distribution function (cdf) of the sample of p-values with the cdf of the standard uniform distribution, the QQ plot shows whether the predictive uncertainty is over or under estimated. The interpretation of the QQ plot is showed in Figure 2.

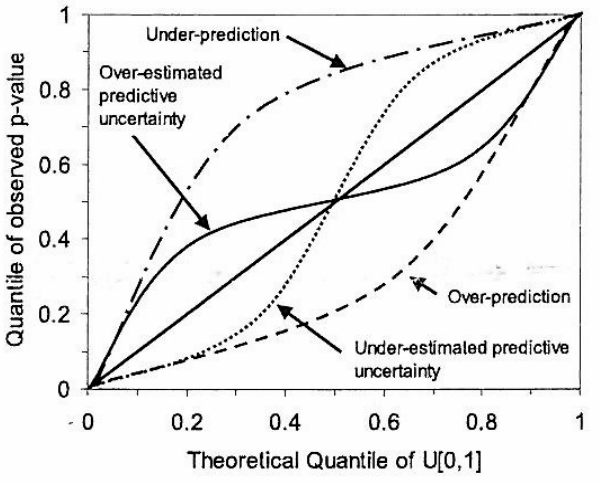


Figure 2: Interpretation of the predictive QQ plot. The figure is taken from Thyer et al. (2009).

Results and discussion

The Bayesian framework was used to calibrate models by updating the prior parameter knowledge with new incorporated information through the study, and Markov chains of samples from the updated posterior parameter distribution were generated through Monte Carlo simulation. I addition to reflect parameter uncertainties and proportion of the uncertainties, this study explored and worked out solutions to challenges that emerged when the method was applied to complex models.

Optimizing Bayesian calibration

Model evaluations were computationally expensive for the complex models of snow cover, soil frost and surface ice *(papers I, II and IV)*. Model complexity in combination with high dimensional parameter spaces and larger amounts of training data made convergence hard to reach *(paper I)*. The effectiveness of the random walk Metropolis algorithm was controlled by the proposal distribution. While small step lengths generating proposed parameter sets close to the current state gave high probability of acceptance but also slow convergence caused by

the small length of each movement, large step lengths rejected a too high proportion of the proposed movements. In the following papers, a multivariate Gaussian distribution (compared to the uniform distribution in *paper I*), with mean at its current state, zero covariance and variances found as the square of a proportion of the prior range was used. The proportion was found individually for each parameter by trial and error to give an efficient acceptance rate between 0.15 and 0.5 (Roberts et al. 1996). Some parameters were more sensitive and controlled most of the acceptance procedure, and made it impossible for Markov chains of the less sensitive parameters to converge if not weighted down by a low proportion *(paper I)*. While larger step lengths reach the position of the posterior chain within a limited amount of time they would not accept new proposed parameters when the position was reached. Some parameters were more challenging and their belonging proportion was tuned once or twice during burn-in to reach convergence within a limited amount of time. All parameter chains were studied during the calibration, and tuned in order to give convergence for all parameters. To overcome the time requiring period of trial and error to find the variances, the Adaptive Metropolis algorithm (Andrieu and Thoms 2008, Smith and Marshall 2008), which solves the problem of finding the step length by using the chain history in order to continually tune the proposal distribution can be used. It was tested during the early work of *paper III*, but only found to be efficient for a small number of parameters, and therefore not included. Also, the adaptation were successfully used informally by using the information obtained by calibrating the model for only a fraction of the data to form a new covariance matrix for the proposal by use of both the proposal distribution used and the correlation matrix calculated from the resulting parameter chains *(paper I).* Additionally the component-wise random walk Metropolis (Ntzoufras 2009) and Importance sampling (Liu 2001) were tested in an early stage of the papers, but non of them were found to be better alternatives then random walk Metropolis.

Models are imperfect representations of the real world systems, and consequently the physical interpretations of the parameters are not exactly correct. The calibration process may therefore prefer parameter values with an unrealistic interpretation. When prior intervals, as the uniform or beta distribution are used to reflect prior knowledge, the resulting posterior distribution will certainly lie within the prior range. When a Markov chain converges at its prior boundary *(papers I - IV)*, it may indicate that the prior limits the exploration of the posterior distribution. The choice of using widened prior intervals which presumably will result in parameter estimates with unrealistic physical interpretation that fit the training data better,

instead of estimates with realistic physical interpretation but worse fit comes up. During the work of *paper III*, widened priors were tested for the parameters that converged to their boundaries. The widened priors gave unrealistic interpretation of the parameters, better fit to the training data and worse fit to the test data. This was according to average RMSEP over ten different splits of the data, and should therefore not reflect a training and a test set not coming from the same population. It may indicate over fitting of the training data (Figure 1).

A Markov chain that converges to the prior boundary *(papers I - IV)* will cause a high rejection rate. To avoid generating proposal parameters outside the prior boundaries, that will regardless be rejected, reflection (Yang 2006) can be used. When having the uniform prior interval $[\theta_{min}, \theta_{max}]$, a proposed parameter $\theta' > \theta_{max}$ will be reset to $\theta_{max}-(\theta'-\theta_{max})$, while a proposed parameter $\theta' < \theta_{min}$ will be reset to $\theta_{min}+(\theta_{min}-\theta')$. When using a multivariate Gaussian proposal distribution, the reflection method only allowes for a diagonal covariance matrix (else the method is not symmetric and random walk Metropolis will be incorrect).

The use of the Bayesian framework for scientific decision making has been debated based on the use of prior knowledge and whether the prior knowledge affects the integrity of the study (Dennis 1996 and Ellison 2004). Both uniformly distributed priors and more informative priors were tested in *paper I*, where the uniform distribution gave the best fit of the model to the data, while the more informative priors permitted more meaningful physically interpretation of the parameter values. The integrity of the collected data and the belonging likelihood function is more rarely debated *(Papers III and IV)*. The likelihood function is rarely known for certain, but gives a reasonable quantification of how probable the data are given the model outcome. The choice of likelihood function and covariance matrix did have an effect on model output *(paper IV)*.

Fixing parameters makes convergence of the chains much easier to reach *(papers I and IV)*. But it has the problem of underestimating uncertainty when the parameters to be fixed are not known accurately (Gelman 1996a).

Propagation of uncertainty

The importance of visualizing how much model outputs are to be trusted, has made uncertainty quantification including proportion of uncertainty an important part of models.

Through Bayesian calibration, uncertainties are quantified by use of probability theory. The obtained Markov chains from the calibration, regarded as samples from the posterior parameter distributions are summarized in 95 % HPD credible intervals *(papers I and III)* and in frequency plots or histograms *(papers I-IV)* to show posterior parameter uncertainty. According to the idea of Bayesian calibration, knowledge will either increase or remain the same through the study. The decrease in parameter uncertainty caused by the study is showed by comparing the credibility intervals with the prior parameter intervals *(papers I and III)*, and by plotting the prior together with the posterior in the frequency plot *(papers I - III)*. The effect on the posterior parameter knowledge by adding data is visualized by making three dimensional frequency plots of the Markov chains as a function of the amount of data *(papers I and III)*. Uncertainties in model output derived from the parameters are showed by calculated RMSD based on 10.000 ensembles of parameter sets randomly drawn from the resulting Markov chains *(paper IV)* and as plots of mean model output ± 1 std calculated over respectively 10.000 and 20 ensembles *(paper I and II)*. Also, output uncertainty derived from the parameters were showed by three dimensional frequency plots of the model output derived from the 10.000 ensembles of parameter sets *(paper I)*. The different predictions derived from different choices of likelihood function and covariance matrix was plotted and the dispersion of the predictions indicated the uncertainty of the estimated outputs derived from the choice of likelihood function and covariance matrix *(paper IV)*. Finally, predictive qq-plot *(paper I)* showed over or under estimation of predictive uncertainty.

Model assessment

Model outputs were plotted together with observations *(papers I-IV)* using the MAP parameter estimate of men ± 1 std. To show model performance a scatterplot of measured and estimated values was made *(paper II)*.

Both the model selection criteria used (DIC and RMSEP) and different splits of the data may be crucial for which model is selected as the best *(paper III)*. To make more general conclusions, both criteria were used and the differences between the different models selected assessed. Also, the data were split ten times in two and the mean result over the splits used to obtain a model fitting well to more general novel data from the same population. Analysis of variance (Montgomery 2005) including the F-test and Tukeys-test were used to show

significance of the model improvements. RMSEP were additionally used to compare different calibration approaches as different prior distributions *(paper I)* and different likelihood functions *(papers IV)*. Frequency plots of the posterior parameter distribution showed how the uncertainty changed while adding data to the study *(papers I and III)*. Not only did the uncertainty decrease, but also the position of the distribution changed. This indicated that new information was added through the new data that was not included at the current data set. Before the position of the distribution stabilized, the conclusions would not be general for novel data from the same population, but apply only to the specific data at hand. In classical statistic, it is of great importance to have large enough sample sizes, in order to make general conclusions and not only conclusions for the data at hand. In Bayesian statistics, prior knowledge makes it possible to run the calibration process for even fewer samples, if having informative reliable priors. When more data are collected, it is possible to update the posterior with the new data, by treating the old posterior distribution as the prior.

Further work

There are clearly many more important and exciting aspects to study related to the use of Bayesian calibration methodology, propagation of uncertainty and to the use of dynamic models beyond what this thesis was able to deal with.

Firstly, the dynamic model for snow cover, soil frost and surface ice *(papers I, II and IV)* was difficult to calibrate. A sensitivity analysis could therefore be run in front of the calibration for fixing the least important parameters in order to gain convergence more easily for the chains.

Secondly, generalized linear models with logit link function did not fit the model generated to estimate maturity of ascospores of *Venturia inaequalis (paper III)*, and asymmetrical models as the Gompertz function (Vieira and Hoffman 1977) could be tested to improve the model.

Thirdly, the comparison of the effect of using different likelihood functions combined with different covariance matrixes *(paper IV)* could be repeated on different studies to make the conclusions of more general value.

Fourthly, predictive uncertainty was underestimated *(paper I)* and the source of this underestimation should be explored further.

References

Agresti, A. (2007). An introduction to categorical data analysis. Wiley

Andrieu, C., Thoms J. (2008). A tutorial on adaptive MCMC. Stat. Comput., 18:343-373

Berger, O. J. (1985) Statistical Decision Theory and Bayesian Analysis. New York, Springer-Verlag

Bøvelstad HM, Nygård S, Størvold HL, Aldrin M, Borgan Ø, Frigessi A, Lindgjære OC. (2007). Predicting survival from microarray data – a comparative study. Bioinformatics 23:2080-2087

Dawid, A. P. (1984). Statistical theory: The prequential Approach. J.R. Stat. Soc., Ser. A 147:278-292

Dennis, D. B. (2001) Discussion: Should Ecologists become Bayesian? Ecological Applications 6(4)1095-1103

Dunson, D. B. (2001). Commentary:Practical Advantages of Bayesian Analysis of Epidemiologic Data. American Journal of Epidemiology 153:1222-1226

Frey, H. C., and Burmaster, D. E. (1999). Methods for Characterizing Variability and Uncertainty: Comparison of Bootstrap Simulation and Likelihood-Based Approaches. Risk Analysis 19:109-130

Gelman, A., Rubin, D. B. (1992). Inference form Iterative Simulation Using Multiple Sequences. Statistical Science 7:457-511

Gelman, A. (2002). Prior distribution. Encyclopedia of Environmetrics. 3:1634-1637

Gelman, A., Bois, F., Jiang, J. (1996a) Physiologival pharmacokinetic analysis using population modeling and informative prior distributions. Journal of the American Statistical Association 91:1400-1412

Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B. (1996b). Bayesian Data Analysis. Chapman & Hall

Goldstein, M., Rougier, J. (2006). Bayesian Linear Calibrated Prediction for Complex Systems. Journal of the American Statistical Association 101:1131-1143

Goldstein, M., Rougier, J. (2009). Refied Bayesian Modelling and Inference for Physical Systems. Journal of Statistical Planning and Inference 139:1221-1239

Hastie, T., Tibshirani, R., Friedman, J. (2001). The elements of Statistical Learning. Data Mining, Inference, and Prediction. Springer

Iizumi, T., Yokozawa, M., Nishimori, M. (2009). Parameter estimation and uncertainty analysis of a large scale crop model for paddy rice: Application of a Bayesian approach. Agricultural and Forest Meteorology 149:333-348|

Kass, R. E., Raftery, A. E. (2005). Bayes Factor. Journal of the American Statistical Assosiation 90:773-795

Kennedy, M. C., O'Hagan, A. (2001). Bayesian Calibration of Computer Models. Journal of the Royal Statistical Society, Series B 63:425-464

Lehunger, S., Gabrielle, B., Van Oijen, M., Makowski, D., Germon, J.-C., Morvan, T. and Hénault, C. (2009). Bayesian calibration of the nittrous oxide emission module of an agroecosystem modell. Agriculture, Ecosystems and Environment 133:208-222

Liu, J. S. (2001). Monte Carlo stratigies in Scientific Compution. Springer-Verlag, New york

Mila, A.L., Yang, X.B., Carriquiry, A. L. (2003). Bayesian Logistic Resgression of Soybean Sclerotinita Stem Root Prevalence in the U.S, North-Central Region: Accountion for Uncertainty in Parameter Estimation. The American Phytopathology Society 758-764

Ntzoufras, I. (2009). Bayesian Modelng using WinBUGS. Wiley

Radtke, P. J., Robinson, A. P. (2006). A Bayesian strategy for combining predictions from empirical and process-based models. Ecological Modelling 190:287-298

Reinds, G.J., Van Oijen, M., Heuvelink, G.B.M., Kros, H. (2008). Bayesian calibration of VSD soil acidification modell using European forest monitoring data. Geoderma 146:475-488.

Roberts, B. O. (1996). Markov Chain concepts related to sampling algorithms. In: Gilks, W. R., Richardson, S., Spiegelhalter, D. J., (Eds.), Markov Chain Monte Carlo in Practice. Chapman & Hall.

Rougier, J. (2007). Probabilistic Inference for Future Climate Using an Ensemble of Climate Model Evaluations. Climatic Change 81:247-264

Sivia, D. S. (2006). Data Analysis, A Bayesian Turtorial. Oxford University press.

Smith, T.J. & Marshall, L.A. (2008). Bayesian methods in hydrologic modeling: A study of recent advancements in Markov chain Monte Carlo techniques. Water Resources Research 44: W00B05

Spiegelhalter, D. J., Best, N. G., Carlin, P. B., Van der Linde A. (2002). Bayesian measures of modell complexity an fit (discussion). Journal of the Royal Slatisrical Society 54:583-616

Thyer, M., Renard, B., Kavetski, D., Kuczera, G., Franks, S.W. and Srikanthan, S. (2009). Critical evaluation of parameter consistency and predictive unccertainty in hydrological modeling: A case study in using Bayesian total error analysis. Water Resources Research 45: W00B14

Vieira S, Hoffman R. (1977). Comparison of the Logistic and the Gompertz Growth Functions Considering Additive and Multiplicative Error terms. Applied Statistics. 23:143-148

Yang, Z. (2006). Computational Molecular Evolution. Oxford University press

I

Fine-tuning Bayesian calibration for complex systems with application to a snow depth model

Anne-Grete Roer*[ab], Stig Morten Thorsen[cd], Trond Rafoss[a], Marcel van Oijen[e], Trygve Almøy[b]

[a] Norwegian Institute for Agricultural and Environmental Research, Plant Health and Plant Protection Division. Høgskoleveien 7, N-1432 Ås, Norway

[b] Norwegian University of Life Science, Department of Chemistry, Biotechnology and Food Science, N-1432 Ås, Norway

[c] Norwegian Institute for Agricultural and Environmental Research, Grassland and Landscape Division. Postvegen 213, N-4353 Klepp st., Norway

[d] Norwegian University of Life Science, Department of Mathematical Science and Technology, N-1432 Ås, Norway

[e] CEH-Edinburgh, Bush Estate. Penicuick, EH26 0QB, UK

*Corresponding author. Tel.: +47 922 83 427; fax: +47 649 46 110.

Email address: anne-grete.roer@bioforsk.no, (Anne-Grete Roer)

Abstract

This paper explores the challenges emerging when applying Bayesian calibration to a complex deterministic dynamic model. The Bayesian approach regards parameters as random and allows integration of prior knowledge. It is here demonstrated how prior information and new data affect the calibration process, parameters and model outputs, with focus on uncertainty. Point estimates and uncertainties are calculated and visualized for both parameters and model outputs. Generally, uncertainty decreased when new data were incorporated. Uniformly distributed priors gave the best fit for this model according to root mean square error, while the more informative beta distributed priors gave more physically meaningful parameter estimates. Markov chains of samples from the posterior distribution of

the parameters were obtained by the random walk Metropolis algorithm. Crucial points when using these methods are reaching and determining convergence of the chains. In order to reach convergence faster, informative priors, Sivia's likelihood, reflection and updating the proposal distribution with parts of the data gave successful results. To determine convergence objectively and correctly, the use of multiple chains and the Gelman Rubin method was found useful. Several decisions must be made when implementing Bayesian calibration, and we highlight and visualize the choices that were found to be most effective.

Key words: convergence diagnostics, model uncertainty, parameter estimates, parameter uncertainty, random walk Metropolis

INTRODUCTION

The potential effects of climate change on Norwegian agriculture are studied in the Norwegian Research program WINSUR (winter survival). A primary goal is to predict the impact of climate change on winter wheat and grass by making climate scenario driven plant growth models. Van Oijen et al. (2005b) developed a plant model for timothy and perennial ryegrass to forecast winter climate impacts on forage crops. Motivated by the need for daily information about snow depth, which is an important climate factor for winter survival of perennial plants, a model is built for predicting the not-commonly-measured variable snow depth, based on two commonly-measured variables, air temperature and precipitation (Thorsen and Haugen 2007). The model is based on a model computing the snow water equivalent developed by Vehvilainene (1992) and the parametrization is based on previous modeling work of Riley and Bonesmo (2005) for a site located at Bioforsk Arable Crops Division, Kise, Norway.

Our snow model is an example of a large category of environmental models, which are deterministic and dynamic and aim to represent the processes underlying the behaviour of the system. The processes are defined by differential equations which the model solves by – computationally demanding - numerical simulation. No such environmental model simulates the real world system perfectly, but still predictions are often made conditional on the model being correct. Predictions related to climate change will not be tested until several decades or maybe a century ahead. It is therefore important to provide decision makers with predictions that are transparent with respect to uncertainty (Thyer et al. 2009). There are three major

sources of uncertainty related to any model (Goldstein and Rougier 2006): (1) the model contains parameters whose values are not certain, (2) the model is an imperfect analog of the system and (3) the collected data contain measurement error. In this paper, we shall focus on uncertainty derived from (1) and (3). Our approach is that of Bayesian calibration (Van Oijen et al. 2005a) which unifies the two goals of model parameterization and uncertainty quantification. Uncertainty with respect to model structure (2) can be addressed in this framework as well, provided multiple models of the same system are available, but we do not carry out Bayesian model comparison in this paper.

In practice, parameter values of environmental models are either inferred from the literature or found by trial and error when little information is available. Calibration is the process of finding the best parameter estimate for the model using data from the system. Maximum likelihood (Miller and Miller 1999) is a well used traditional calibration routine that maximizes the probability of the data given the parameters, $f(\boldsymbol{x}|\boldsymbol{\theta})$. Limitations of the maximum likelihood approach are that uncertainties can not readily be quantified and conclusions made by the modeler are conditional on the model being correct. An alternative Bayesian approach is more rarely used for complex models (Van Oijen et al. 2005a, Hue et al. 2008, Luo et al. 2009), partly because of practical problems addressed in the present paper. A key issue is the computationally demanding numerical solution of differential equations, which limits the number of model evaluations for calibration that is feasible. Despite these computational problems, the application of the Bayesian method to environmental models has been increasing in recent years (e.g. Reinds et al. 2008, Lehuger et al. 2009) because it improves on the traditional approach by automatically including uncertainty quantification (Campbell 2006). It also allows for prior information about the parameters and conclusions are made conditional on the data.

Much pioneering work on the Bayesian calibration of environmental models has been carried out in hydrology (e.g. Kavetski et al. 2006, Smith and Marshall 2008, Thyer et al. 2009), often for stochastic models of water flow in response to precipitation. However, experience with the approach for the slow environmental models is sill limited and, as Campbell (2006) states in a recent review of calibration of computer simulators, “much work is still to be done to place calibration on a sound and practical statistical footing”.

The main objective for this paper has been to apply Bayesian calibration to a complex model in order to explore practical problems with the calibration as well as work out solutions. Point

estimates are calculated and uncertainties visualized for both the parameters and the model outputs. In order to obtain convergence of the chains (and thus reasonable results) by simulation in a limited amount of time, the usefulness of informative priors, Sivia's constrained likelihood, the reflection method and different proposal distributions including optimizing the proposal distribution with parts of the data are tested. To detect the state of convergence, we have checked the usefulness of multiple chains and Gelman-Rubin.

SNOW DEPTH MODEL

The SnowFrost model, described in detail in Thorsen and Haugen (2007) is a one dimensional model which simulates the dynamics of depth of snow cover S_{depth} (m) and soil frost F_{depth} (m). SnowFrost is integrated in a grassland model which simulates the regrowth dynamics of timothy (*Phleum pratense* L.). This grassland model by Van Oijen et al. (2005a) is under further development. There are two main modules in SnowFrost; one module relates to the dynamics of the snow cover, and one module relates to the formation of soil frost. In SnowFrost, the formation of soil frost is affected by the presence of a snow cover, but the snow cover is not affected by the presence of soil frost. In this paper the focus is on the snow module (Figure 1) and the calibration of its parameters, and thus we leave out issues related to soil frost. Preliminary calibration of SnowFrost suggested some modifications, and this new snow depth model is presented below.

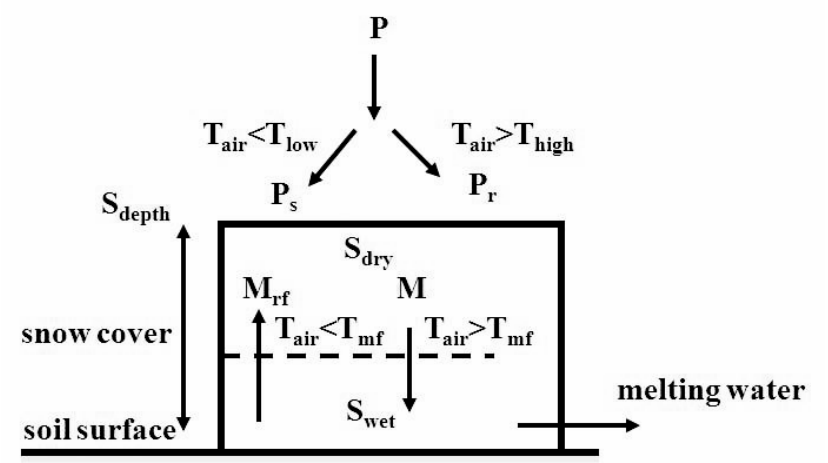


Figure 1: Description of the system simulated in the snow depth model.

Table 1: Symbol and description of the 9 parameters in the snow depth model.

$\{\boldsymbol{\theta_i}\}$	**Interpretation**	**Symbol**
1	Precipitation falls as rain if $T_{air} > T_{high}$ (°C)	T_{high}
2	Precipitation falls as snow if $T_{air} < T_{low}$ (°C)	T_{low}
3	Threshold temperature for snow melt M (°C) and refreezing S_{wet}	T_{mf}
4	Densification of snow cover (mm mm-1 day-1)	ξ
5	The difference between the maximum and minimum value for the melting rate of snow pack K (mm °C-1 day-1)	ΔK_{max}
6	Minimum value for the melting rate of snow pack K (mm °C-1 day-1)	K_{min}
7	Degree-day temperature refreezing index (mm °C-1 day-1)	SW_{rf}
8	The density of fresh snow (kg m-3)	ρ_{ns}
9	The retention capacity of snow cover (mm mm-1)	SW_{ret}

Based on precipitation rate P (mm day-1), mean air temperature T_{air} (ºC) and two threshold temperatures T_{high} (ºC) and T_{low} (ºC) (see Table1), the snow model determines the precipitation form (rain P_r (mm day-1) or snow P_s (mm day-1), where falling new snow has the density ρ_{ns} (kg m-3)) by calculating a fraction of liquid water f_w (mm mm-1) of the precipitation according to

$$f_w = \begin{cases} 1.0 & \text{if } T_{high} < T_{air} \\ \dfrac{T_{air} - T_{low}}{T_{high} - T_{low}} & \text{if } T_{low} \leq T_{air} \leq T_{high} \\ 0 & \text{if } T_{air} < T_{low} \end{cases}$$

where the corresponding amounts of P_r and P_s are

$$P_r = f_w P$$
$$P_s = (1 - f_w) P$$

The snow cover consists of water in solid state S_{dry} (mm) (snow and ice), and liquid water within the snow cover S_{wet} (mm). In SnowFrost snow melt occurs when T_{air} exceeds the base temperature T_{bm} (ºC), and refreezing of S_{wet} occurs when T_{air} drops below T_{bf} (ºC). Preliminary calibration of the SnowFrost model showed that the marginal posterior distributions for the two threshold temperatures T_{bm} and T_{bf} was practically equal. We therefore replaced T_{bm} and T_{bf} with one threshold temperature T_{mf} (ºC), that determines whether snow is melting M (mm day-1), when $T_{air} > T_{mf}$, or liquid water within the snow is refreezing M_{rf} (mm day-1), when $T_{air} < T_{mf}$. The snow cover, being a porous medium, can retain a limited amount of liquid water S_{wet} resulting from rain or melted snow. Similar to

Engseth et al. (2000), we estimate the potential retention capacity of the snow cover as $SW_{ret} \cdot S_{dry}$ where SW_{ret} (mm mm-1) is the retention capacity of the snow cover. Liquid water within the snow cover may refreeze at the rate SW_{rf} (mm ºC-1 day-1). Also, following the idea of Engseth et al. (2000), we calculate the rate of snow melt using a temperature dependent rate K (mm ºC-1 day-1) described by a sinusoidal curve; the period is one year with maximum snow melt rate K_{max} (mm ºC-1 day-1) occurring on 23. June, and minimum snow melt K_{min} (mm ºC-1 day-1) on 23. December. To avoid situations like $K_{max} < K_{min}$ during the calibration, we replaced K_{max} by $\Delta K_{max} = K_{max} - K_{min}$ and calibrate ΔK_{max} (mm ºC-1 day-1). If the entire snowpack melted instantaneously, the resulting depth of water is known as the snow water equivalent SWE (mm). SWE is defined as the sum of S_{dry} and S_{wet}, and the density of the snow cover ρ_s (kg m-3) is defined as SWE/S_{depth} (note: 1 mm of precipitation equals 1 kg m-2). Densification of the snow cover due to change in physical properties (e.g. change in shape of snowflakes and the increase in weight of overlying snow following accumulation) is incorporated through the empirical compaction parameter ξ (mm mm-1 day-1). We use the following equations for the snow cover dynamics:

$$
\begin{aligned}
\Delta \mathrm{S_{dry}}/\Delta \mathrm{t} &= P_s + M_{rf} - M \\
\Delta \mathrm{S_{wet}}/\Delta \mathrm{t} &= P_r + M - M_{rf} \\
\Delta \mathrm{S_{depth}}/\Delta \mathrm{t} &= P_s/\rho_{ns} - M/\rho_s - \xi S_{depth} \\
K &= \Delta K_{max}/2 \sin(2\pi t/365 + 3\pi/8) + (K_{min} + \Delta K_{max}/2) \\
M &= K(T_{air} - T_{mf}) \\
M_{rf} &= SW_{rf}(T_{mf} - T_{air})
\end{aligned}
$$

Snow depth model parameters to be calibrated are listed in Table 1.

STATISTICAL METHOD

The model, $Y(\boldsymbol{\theta}, \boldsymbol{x})$ simulates output variables $\hat{\boldsymbol{y}}$ using input variables $\boldsymbol{x}$ and parameters $\boldsymbol{\theta}$. In the Bayesian calibration approach, parameters are regarded as random variables and thus follow some probability distribution. Instead of searching for the best parameter estimates $\boldsymbol{\theta}^*$, we actually search for the probability distribution of these parameters. The calibration routine collects samples from these distributions and parameter uncertainties may be visualized together with point estimates.

Bayesian learning

Bayes theorem is the building block in Bayesian calibration. It was formulated by Thomas Bayes in 1763 (Berger 1985), and may be reformulated as

$$\pi(\boldsymbol{\theta}|\boldsymbol{d}) = \frac{f(\boldsymbol{d}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\boldsymbol{d})}$$

where the parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ ($\boldsymbol{\Theta}$ is the whole parameter space) and $\boldsymbol{d} = (d_1, d_2, \dots, d_M)$ is the collected data. The formula reverses conditional probabilities by looking at the unknown parameter set $\boldsymbol{\theta}$ as random variables. The posterior probability distribution $\pi(\boldsymbol{\theta}|\boldsymbol{d})$, is the probability distribution of the parameters given the collected data. According to Bayes formula it is found by combining the original parameter uncertainty, expressed by a prior probability distribution $\pi(\boldsymbol{\theta})$ and the conditional probability density function of the collected data given the parameters, $f(\boldsymbol{d}|\boldsymbol{\theta})$ (often called the likelihood function and denoted $L_d(\boldsymbol{\theta})$). The so called 'evidence' or 'integrated likelihood' term $f(\boldsymbol{d})$ is constant and found by the integral $f(\boldsymbol{d}) = \int_{\boldsymbol{\Theta}} f(\boldsymbol{d}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\theta$. This gives us the proportionality

$$\pi(\boldsymbol{\theta}|\boldsymbol{d}) \propto f(\boldsymbol{d}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \qquad (1)$$

which shows that the posterior information is a combination of prior knowledge and new information incorporated through the likelihood function of the collected data.

The prior distribution

The prior distribution quantifies the original uncertainty we have about the parameters. According to Ellison (1996) there are three different interpretations of the prior distribution: (1) a frequency distribution based on existing data. As long as the same data is not used twice, a part of the collected data can be used, or existing data from an earlier investigation, (2) an "objective" statement of what is rational to believe about the parameters and (3) a subjective measure of what the investigator actually believes about the parameter values.

Although limited, the prior information that reflects the initial population basis will assist in the probability distribution of the posterior prediction (Gelman et al. 1996b and Marshall et al. 2004). If the prior dominates the likelihood, the prior will have much greater effect on the posterior probability function than the subsequent experiment can supply. Most of the

criticism of Bayesian inference is that Bayesian analysis can produce results consistent with any point of view when specifying a subjective prior based on personal belief (Dennis 2004). It is therefore of great importance, not to use unrealistically informative prior. If non-informative prior distributions were used for all the individual parameters, then the model would fit the data very closely, but often also with scientifically unreasonable parameters. This may motivate the researcher to specify a prior distribution using external information (Gelman 2002). If no prior information of the parameters is available, non-informative priors (one approach introduced by Jeffrey (1961)) may be used, so that the inferences are unaffected by information external to the experiment (Gelman et al. 1996 b). As usual, we will assume prior independence between the parameters. When having more than one parameter, the joint prior density can be written

$$\pi(\boldsymbol{\theta}) = \pi(\theta_1, \theta_2, \ldots, \theta_L) = \prod_{l=1}^{L} \pi_l$$

where L is the total number of parameters in the model.

The Likelihood function

The likelihood function is the data distribution, conditional on the model used and, expressed as a function of the model parameter values. Measurements $\boldsymbol{d}$ made in the true observable quantity $\boldsymbol{y}$ are not perfect. At the same time, the model $Y(\boldsymbol{\theta},\boldsymbol{x})$ is a simplification of the real world system.

$$\boldsymbol{d} \equiv Y(\boldsymbol{\theta}, \boldsymbol{x}) + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\varepsilon}$ is both measurement and representational error. After some simplifications (Rougier 2007), the likelihood function can be written as:

$$\begin{aligned} L_d(\boldsymbol{\theta}^*) &= f(\boldsymbol{d}|\boldsymbol{\theta}^*, \boldsymbol{x}^*) \\ &= \prod_{m=1}^{M} f(d_m|\boldsymbol{\theta}^*, \boldsymbol{x}^*) \\ &= \prod_{m=1}^{M} \varphi(d_m - Y_m(\boldsymbol{\theta}^*, \boldsymbol{x}^*); 0, \sigma_m) \end{aligned}$$

where φ is the univariate normal probability density function and σ_m is the m'th diagonal element of the diagonal variance matrix Σ of the errors. The likelihood function can then be written:

$$L_d(\boldsymbol{\theta}^*) = \prod_{m=1}^{M} \frac{1}{\sigma_m (2\pi)^{1/2}} exp\left(-\frac{1}{2}(d_m - Y_m(\boldsymbol{\theta}^*, \boldsymbol{x}^*))^2/\sigma_m^2\right) \quad (2)$$

Outliers in the collected data may produce bad results. Sivia (2006) solves this problem by formulating a constraint on the Gaussian likelihood function. He used a variant of Jeffreys' prior to specify a lower boundary (σ_0) for the standard deviation

$$\pi(\sigma|\sigma_0) = \sigma_0/\sigma^2$$

for $\sigma \geq \sigma_0$, and zero otherwise.

The formula of the constraint likelihood function with the unknown σ integrated out is written

$$L_{d_m}^{(1)}(\boldsymbol{\theta}^*) = \frac{1}{\sigma_{0_m}(2\pi)^{1/2}}(1 - \exp(-R_m^2/2))/R_m^2$$

Where σ_{0_m} is the lower bound of the standard deviation and $R_m = (d_m - Y_m(\boldsymbol{\theta}^*, \boldsymbol{x}^*))/\sigma_{0_m}$. The equation is not defined for $R_m = 0$, but the limit likelihood when $R_m$goes through zero is found as

$$L_{d_m}^{(0)}(\boldsymbol{\theta}^*) = \lim_{R_m \to 0} L_1(\boldsymbol{\theta}^*) = \frac{1}{2\sigma_{0_m}(2\pi)^{1/2}}$$

By series expansion (not shown here). The total constraint Gaussian likelihood function is finally defined as

$$L_d(\boldsymbol{\theta}^*) = \prod_{R_m \neq 0} L_{d_m}^{(1)}(\boldsymbol{\theta}^*) \cdot \prod_{R_m = 0} L_{d_m}^{(0)}(\boldsymbol{\theta}^*) \quad (3)$$

The variance-covariance matrix, Σ_0 of model and measurement error in the likelihood function is unknown. The measurement error may be found by investigating how reliable the measurement instrument used is. The representational error including both simplifications in the model and the fact that the model and the data are not talking about the exact same parameter (the model predicts mean snow depth while the measurements are point estimates),

would be much harder to find. The problem of estimating the covariance matrix of model errors has been simplified by using a fixed diagonal covariance matrix. Conform Van Oijen et al (2005a), we set the standard deviation of each measurement to 30 % of the mean observed value.

$$\hat{\sigma}_{0m} = 0.3 \cdot d_m$$

To avoid a standard deviation of zero, when no snow is observed, the standard deviation was redefined to be

$$\hat{\sigma}_{0m} = \max\left(0.1, 0.3 \cdot d_m\right)$$

This gives a standard deviation of 0.1 if the mean collected snow depth is less than 0.33m.

The difference between observed and simulated output ($d_m - Y_m(\boldsymbol{\theta}^*, \boldsymbol{x}^*)$), will be dominated by zeros, because no snow depth will be both observed and simulated most of the year. A student-t distribution, having a fatter tail (Miller and Miller 1999) is an alternative recommended when outliers occurs, but is not tested here. Probably a zero-inflated distribution (Agarwal et al. 2002) would be an even better choice

The likelihood function is what modifies the prior knowledge into a posterior distribution. According to Bayes theorem, the more experimental data added, the more will the likelihood dominate the prior, and have much greater effect on the posterior probability distribution.

Random Walk Metropolis

There are four different techniques that can be used to find the posterior distribution in Bayesian calibration; exact calculation, analytical approximation, numerical integration and Monte Carlo simulation. Integration problems makes the exact calculation impossible, especially when $\boldsymbol{\Theta}$ is high dimensional. We will use a MCMC algorithm, random walk Metropolis.

The Metropolis Algorithm is the cornerstone of all Markov chain-based Monte Carlo methods. It was proposed as early as in 1953 in a short paper by Metropolis et al. (1953). The idea is of great simplicity and power, and its variations are in use by many researchers in several different scientific fields.

We implemented the random walk Metropolis algorithm in Matlab. We start with some initial parameter values, $\boldsymbol{\theta}^0 = (\theta_1^0, \theta_2^0, \ldots, \theta_L^0)$, where L is the number of parameters in the model. For each iteration step $i \in 1{:}\, I$, we have these steps:

1. Draw $\boldsymbol{\theta}' \sim g_\sigma$, where g_σ is a spherically symmetric distribution, independent distributed for different i, centered at the current state.
2. Compute the ratio $r = \frac{\pi(\boldsymbol{\theta}'|\boldsymbol{d})}{\pi(\boldsymbol{\theta}^{i-1}|\boldsymbol{d})} = \frac{\pi(\boldsymbol{\theta}\prime)\cdot f(\boldsymbol{d}|\boldsymbol{\theta}\prime)}{\pi(\boldsymbol{\theta}^{i-1})\cdot f(\boldsymbol{d}|\boldsymbol{\theta}^{i-1})}$
3. Draw $u \sim \mathrm{U}[0,1]$, where U is the uniform probability density function, and set

$$\boldsymbol{\theta}^i = \begin{cases} \boldsymbol{\theta}' & \text{if } u \le r \\ \boldsymbol{\theta}^{i-1} & \text{otherwise} \end{cases}$$

The draws $\theta^1, \theta^2, \ldots, \theta^I$ will in the long run converge to the posterior distribution of the parameter set (Liu 2001).

To avoid the joint likelihood to be too large to be represented by a digital computer, that round off to infinity, the natural logarithm was used in all steps in the random walk Metropolis algorithm.

The step length δ is the distance between the current and the proposed parameter vector. Small δ ensures that the proposed parameter vector is close to the current position, so the probability of accepting it is high. With small average δ, the Markov chain will converge slowly since all its moves will be small. On the other hand, a large step length δ places the new proposed parameter further away from the current parameter vector, which leads to a low probability of accepting it. The Metropolis algorithm will then reject a too high proportion of its proposed moves. Most of the computation time goes to costly evaluation of the posterior density. The step length δ therefore controls the effectiveness of the Metropolis algorithm. According to Roberts et al. (1997), an acceptance rate of roughly 0.23 is desired. We considered an acceptance rate between 0.15 and 0.5 to be acceptable (Roberts 1996).

The choice of a proposal distribution may be a crucial factor for convergence of the algorithm. Adaptive MCMC algorithms (Andrieu and Thoms 2008, Smith and Marshall 2008) solve this problem by using the chain history in order to continually tune the proposal distribution.

Convergence Diagnostics

The random walk Metropolis algorithm produces a Markov chain whose stationary distribution is the target posterior distribution. If the iterations before stationarity are used to summarize the target distribution, they can give false answers. To detect the state of stationarity (``burn-in'' state) or lack of stationarity, different methods exist. Gelman and Rubin (1992) pointed out that in many problems, lack of convergence can be easily determined from multiple independent sequences but can not be diagnosed using simulation outputs from any single sequence. The sequence may remain in a region heavily influenced by the starting point, although it has not converged to the true posterior distribution (Gelman et al. 1996b). In contrast, Geyer (1992) states that one should concentrate all computational resources in a single chain, since it is more likely that for example the latter 90.000 iterations from a single run of 100.000 iterations come from the target distribution than the final samples from 10 parallel runs of 10.000.

When running parallel sequences, the most obvious approach to assess convergence is to plot the chains as time series and assess by visual inspection whether the sequences have converged to each other. A more quantitative approach based on the ratio of between- and within-chain variance was formulated by Gelman and Rubin (1992). Convergence is identified when the empirical distribution of simulations obtained separately from each sequence is approximately the same as the distribution obtained by mixing all the sequences together. Before the parallel sequences have converged, the simulations from each sequence will be much less variable than the sequence combined. We assume J parallel simulations, each of length I and with starting points drawn randomly from the prior distribution that is over-dispersed in the sense of being more variable than the target posterior distribution. The first $I/2$ iterations are discarded to diminish the effect of the starting distribution. The estimated potential scale reduction factor $\sqrt{\hat{R}}$ is calculated at each iteration step

$$\sqrt{\hat{R}} = \sqrt{\left(\frac{I/2-1}{I/2} + \frac{J+1}{J\,I/2}\frac{\hat{B}}{\widehat{W}}\frac{df}{df-2}\right)}$$

where B is the variance between the sequence means and W is the average within-sequence variance. df refers to the degree of freedom in a t-distribution approximation to the empirical distribution of θ. For large number of samples, $df/(df-2)$ can be ignored.

$$\hat{B} = \frac{I/2}{J-1}\sum_{j=1}^{J}\left(\overline{\theta}_{J\cdot} - \overline{\theta}_{\cdot\cdot}\right)^2$$

$$\widehat{W} = \frac{1}{J(I/2-1)}\sum_{i=1}^{I/2}\sum_{j=1}^{J}\left(\theta_{ji} - \overline{\theta}_{J\cdot}\right)^2$$

When $\sqrt{\hat{R}}$ is close to 1 (less than 1.2 in practice (Gelman 1996)), the parallel Markov chains are essentially overlapping. We should also make sure that the mixture of sequence variance $\hat{V}$ and the within sequence variance $\widehat{W}$ stabilizes as a function of I (Brooks and Gelman 1998).

$$\hat{V} = \frac{I/2-1}{I/2}\widehat{W} + \frac{J+1}{J\,I/2}\hat{B}$$

The iterations before the ``burn-in'' state are discarded. Typically one will discard only a small fraction of the run. So, if after ``burn-in'' state you are left with less than half the run, you haven't run the iterations for long enough (Kass et al. 1998).

DATA

The snow depth model is calibrated using snow depth data from Kise, Norway, which is situated 60.77N, 10.8E, 127 meters above sea level. Kise has a continental climate, and the landscape is dominated by arable land and the largest lake in Norway, Mjøsa. The model is calibrated using data from 10 years, 1988 to 1998, and it is tested using data from the 5 following years. Temperature and precipitation observations are obtained from Bioforsk Agrometeorological service, the snow depth observations are obtained from both Bioforsk Agrometeorological service (1988-1997) and from The Norwegian Water Resources and Energy Directorates service (1997-2003). On average there was snow cover 120 days of the year, with an average depth of 0.16 m. Variation between years was from 160 days with snow cover and an average of 0.36 m the winter 1993/1994 to only 77 days with snow cover and an average of 0.06 m the winter 1989/1990,

RESULTS AND DISCUSSION

Tuning the MCMC

To run the Bayesian calibration algorithm, several decisions must be made by the researcher.

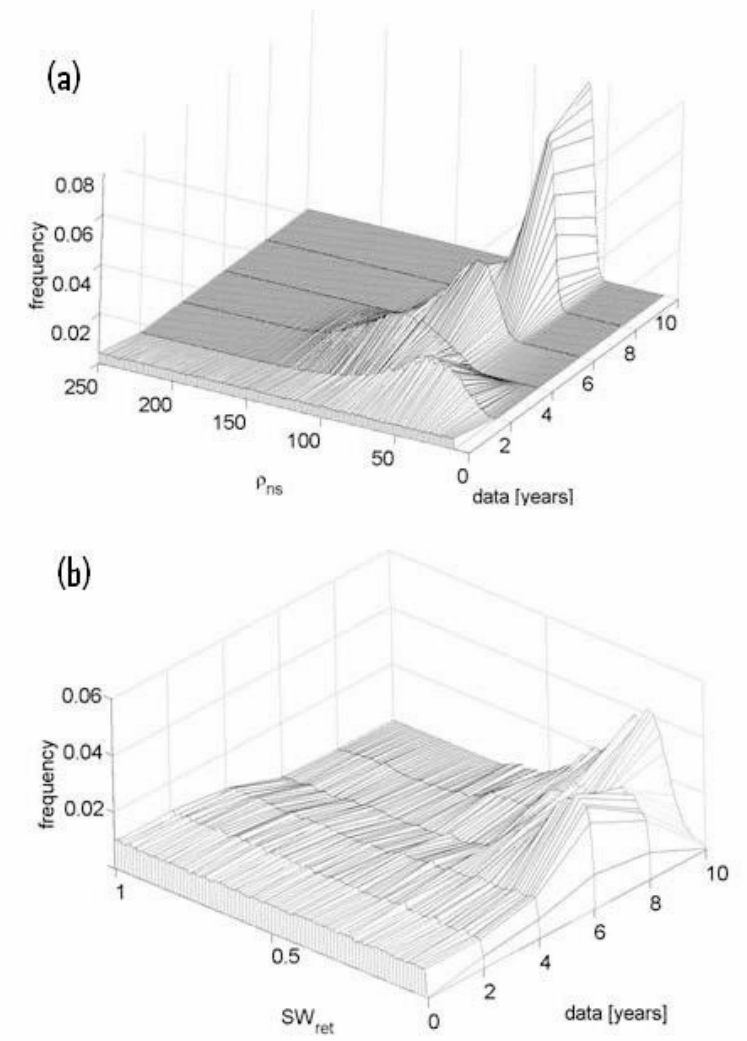
The number of observed data

Usually, we will use all available data. Here, 10 years of snow depth data are used to fit the model and 5 years to test the model. To see the effect of the number of data used to fit the model, the calibration algorithm was also run using 2, 4, 6 and 8 years of collected data. The results from ρ_{ns} (θ_8) and SW_{ret} (θ_9) are plotted as three dimensional figures to visualize the change in uncertainty about the respective parameter when adding data (Figure 2). In agreement with Bayesian learning (Equation 1), we can see that the uncertainty, i.e. the width of the histograms, decreases when adding more data. Also the position of the histograms changes, most dramatically up to 8 years, but also from 8 to 10 years. The weather situation varies from year to year, and the parameter estimates do depend on what kind of years used. A period of 8 years does not contain all variability in weather, and the estimates do therefore change further when adding two more years. Most probably, the estimates will still change when adding more years of data, until the whole specter of weather situations are included.

Reaching convergence for the posterior chains was easy when 2 or 4 years of collected data were used. With 6 or more years of data, convergence became much harder to reach.

The prior distribution

We have chosen to use relatively wide uniform prior distributions. Usually, we will not consider all values between the upper and lower limit in the prior distribution as equally believable. We therefore constructed a beta distribution between the boundaries and used results from Engseth et al. 2000 (Table 2) as modal values. Comparing the results of calibration starting from beta distributions rather than uniform ones, showed that the more informative beta priors gave much easier convergence and different point and interval estimates for the parameters. These new estimates permitted more meaningful physical interpretation, but showed worse fit according to RMSE for both the training and the test data.



Figur 2: Changes in parameter uncertainty when using respectively 0, 2, 4, 8 and 10 years with collected data in the calibration routine. Figure (a) shows parameter uncertainty for $\boldsymbol{\rho_{ns}}$ $(\boldsymbol{\theta_8})$and $\boldsymbol{SW_{ret}}$ $(\boldsymbol{\theta_9})$.

RMSE is now commonly reported in environmental modeling in comparsions of different calibration approaches (e.g. Reinds et al. 2008, Lehuger et al. 2009). The lower RMSE with the uniform prior was not necessarily expected because our likelihood function was not Gaussian nor did our data all have the same standard deviation.

The Likelihood function

Both the Gaussian likelihood function (Equation 2) and the Gaussian likelihood function with constraints (Equation 3) were tested and the constrained function gave much faster parameter convergence than the ordinary Gaussian. Less iteration were needed to reach convergence and since each iteration step is time requiring; only the constrained function was used.

The proposal distribution

The choice of an effective proposal distribution for the MCMC is essential in order to obtain convergence in a limited amount of time. Both a uniform and a Gaussian proposal distribution centered at the current state were tested. The Gaussian distribution, which predominantly samples close to zero turned out to be the most efficient one and was therefore used. The covariance matrix was defined as diagonal with the *l*'th diagonal element proportional to the width of the prior interval for the respective parameter $\left(a_l = c \cdot \left(\theta_l^{max} - \theta_l^{min}\right)\right)$. In order to achieve efficient convergence, the constant c was set by trial and error to produce an acceptance rate of roughly 0.23 (Roberts et al. 1997). To prevent the most sensitive parameters T_{high} (θ_1), T_{low} (θ_2), ξ (θ_4) and ρ_{ns} (θ_8) (according to sensitivity analysis calculated for the entire *SnowFrostIce* model (Thorsen et al. 2009)) to control the whole accept/reject procedure, the constant c was individually corrected up for all other parameters. The sensitive parameter ξ (θ_4) proved most challenging. If its proposal stepsize was not weighted down, it controlled most of the accept/reject procedure and made it impossible for all other parameters to converge. At the same time, if weighted down enough, convergence is not reached within a proper time when having a starting value for the parameter far away from the target posterior distribution. Our solution was to weigh the parameter down after a number of iterations. To keep the rule that the step length has to be identically distributed for different iterations (Liu 2001), this is done during the ``burn-in'' phase only.

Adaptive MCMC algorithms (Andrieu and Thoms, 2008, Smith and Marshall 2008) were not used, but we did use adaptation informally, as follows. The information obtained by the calibration using two years of observations was used to form a suitable proposal distribution for the calibration using all ten years of observations. The new covariance matrix was calculated from the variances of the proposal distribution and the correlation matrix calculated from the resulting parameter chains after convergence when using two years of observations.

Then only a scaling factor for the entire covariance matrix had to be found by trial and error to produce an efficient acceptance rate. Preliminary tests of this method showed it to be highly efficient for the calibration of our model, but it was not used to produce the results reported here.

Convergence

The calibration algorithm was run for 300.000 iterations, requiring about 40 hours computing time. The usefulness of running parallel sequences to detect convergence was found during model development from the first version of the snow model (Thorsen and Haugen 2007) to this version. The Markov chain for T_{bm} (Lower limit temperature for snow melt) is plotted in Figure 3a for 150.000 iterations, which objectively seems like a large enough number. Running only one chain, we would determine ``burn-in" after 10.000 iterations by eyes, and treat the remaining 140.000 iterations as draws from the true posterior distribution. When running two chains in parallel instead (Figure 3b), we found that the two chains had not converged to each other during this run. We therefore ran the algorithm for 150.000 more

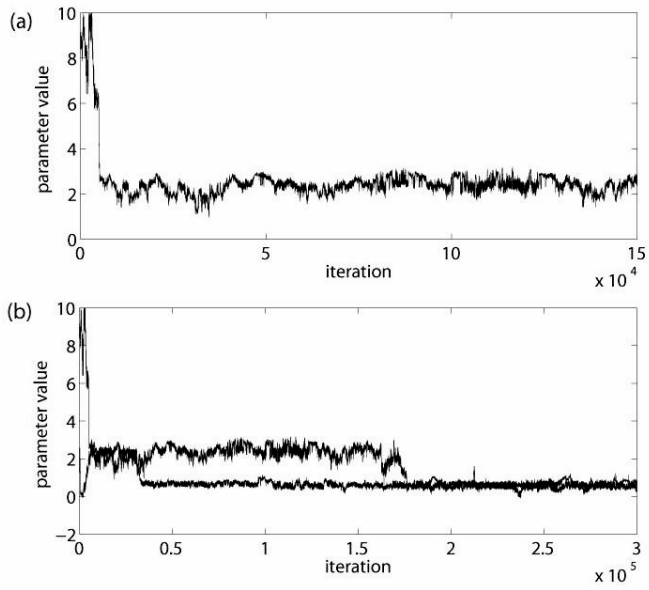


Figure 3: Markov chain of $\boldsymbol{T_{bm}}$ (parameter not included in the finial version of the snow depth model) (a) one chain for 150.000 iterations and (b) two chains in parallel for 300.000 iterations

iterations and clearly they converge to each other after 175.000 iterations. With more confidence, we can now treat the last 125.000 iterations as draws from the posterior distribution. In most cases, four sequences were run in parallel, but during the development of the model only two. The method of Gelman and Rubin was used to detect ``burn-in''.

Reflection

When using upper and lower limits in the prior distribution, new proposal parameters may be generated outside these boundaries and consequently be rejected in the random walk Metropolis algorithm. Here, T_{low} (θ_2), which stabilized near the lower boundary of the prior interval, caused a high rejection rate. To avoid proposal parameters generated outside the prior boundaries, reflection at the boundaries (Yang 2006) is used. If the proposal parameter θ' is outside the prior interval ($[\theta_{min}, \theta_{max}]$) the excess is reflected back into the interval; that is, if $\theta' > \theta_{max}$, θ' is reset to $\theta_{max} - (\theta' - \theta_{max})$, and if $\theta' < \theta_{min}$, θ' is reset to $\theta_{min} + (\theta_{min} - \theta')$. The proposal parameter distribution will still be symmetric($p(\theta' | \theta^{i-1}) = p(\theta^{i-1} | \theta')$) and thereby the acceptance of the Metropolis algorithm correct, since $p(\delta) = p(-\delta)$ and a step length δ from the current state θ^{i-1} with reflection will give the state θ' ($\theta' = \theta^{i-1} + \delta = 2\theta_{min} - (\theta^{i-1} + \delta)$), while the same step length δ from θ' will reflect the proposal parameter back to current state θ^{i-1} ($\theta' + \delta = 2\theta_{min} - (\theta' + \delta) = \theta^{i-1}$).

Uncertain vs. fixed values of the parameters

To reach convergence for all parameters turned out to be difficult. We therefore reduced the number of parameters by setting some of them to fixed values in the early investigation. This showed that convergence problems emerged when four or more parameters were included in the calibration. It was therefore tempting to treat some of the less sensitive parameters as fixed values. But this has the problem of producing inaccurate estimates and it underestimates uncertainty when the parameters to be fixed are not known accurately (Gelman et al. 1996a). Especially the posterior distribution of T_{low} (θ_2) changed dramatically when the last three parameters were fixed. In the results, all 9 parameters are treated as uncertain. By performing a sensitivity analysis, e.g. using the Morris method as described in Campolongo et al. 2007, it

is possible to identify the least important parameters which are candidates for fixed values. This was done for a similar snow depth model by Thorsen et al. 2009.

Statistical Inference

The model is calibrated using wide uniform prior distributions (Table 2) and four sequences are run in parallel, each for 300.000 iterations. All parameters converged after less than 100.000 iterations (according to Gelman and Rubins criteria) with an acceptance rate of approximately 0.25. The potential scale reduction factor, $\sqrt{\hat{R}}$ at the end of the calibration is listed in Table 3. The highest value is 1.02 for both ΔK_{max} (θ_5) and SW_{ret} (θ_9) suggesting that additional simulation might reduce the posterior interval for these parameters by only up to a factor of 1.02. The 95 % highest posterior density (HPD) interval is calculated for each parameter and listed in Table 3. Each interval is a measure of how certain we are about the respective parameter, and we can clearly see a decreased uncertainty compared to the prior information (Table 2) for all parameters except SW_{rf} (θ_7).

Three point estimates are calculated, the mean $(\overline{\boldsymbol{\theta}})$, the mode $\left(\widehat{\boldsymbol{\theta}}_{mode}\right)$ and the maximum a posteriori estimate ($\widehat{\boldsymbol{\theta}}_{MAP} = \arg\max\ \pi(\boldsymbol{\theta}|\boldsymbol{x})$) (Gilks et al. 1996). Both the mean and the mode estimates are calculated for each parameter one by one, while the MAP estimate is the largest mode for the joint posterior distribution. All three estimates are different (Table 3)

Table 2: Minimum and maximum values used to define limits in the uniform prior interval and parameter estimates from Engseth et al. (2000), $\widehat{\boldsymbol{\theta}}_E$. Lack of number indicates that the specific parameter does not occur in Engseth's model.

$\{\boldsymbol{\theta}_i\}$	Symbol	$\boldsymbol{\theta}_{min}$	$\boldsymbol{\theta}_{max}$	$\widehat{\boldsymbol{\theta}}_E$
1	T_{high}	-5	10	0.5
2	T_{low}	-10	5	0.5
3	T_{mf}	-10	10	0.5
4	ξ	0	1	
5	ΔK_{max}	0	10	1.25
6	K_{min}	0	10	2
7	SW_{rf}	0	10	0.01
8	ρ_{ns}	10	250	
9	SW_{ret}	0	1	0.1

which may be explained by skewness and several peaks in the parameter densities and by correlations (Table 3) between the different parameters. The MAP estimate, which is the only estimate considering the entire parameter set concurrently is the estimate giving the smallest root mean square error (RMSE) for both the training data and for the test data. It is important to note that the MAP estimate is the parameter set having the largest posterior density among our 800.000 iterations. Each of our four parallel chains gave a different MAP estimate, suggesting that the parameter space is not totally searched and additional simulation might give even better MAP estimates. The Bayesian calibration method does not search for the best parameter estimates, but for the posterior distribution of them.

Bayesian calibration simply combines prior parameter information with the likelihood of the data given the model. Since the model is not a perfect representation of the system, parameter estimates may deviate from physically meaningful values. Here, especially T_{low} (θ_2) seems unrealistic (Table 3). To reach a more realistic estimate, the model could have been improved, a more informative prior distribution could have been used or a separating parameter could replace T_{high} (θ_1) and T_{low} (θ_2). Since the purpose of our model is not to learn about the system, but prediction and we do not want a too complex model, both the model and the parameters are retained while a more informative prior was tested.

Table 3: Results from Bayesian Calibration using MCMC with chains of 300000 iterations. The potential scale reduction factor $\left(\sqrt{\widehat{R}}\right)$, the mean parameter value $(\overline{\boldsymbol{\theta}})$, the mode parameter value $(\widehat{\boldsymbol{\theta}}_{\boldsymbol{mode}})$, the maximum a' posteriori estimate $(\widehat{\boldsymbol{\theta}}_{\boldsymbol{MAP}})$, the coefficient of variation (CV), the 95 % HPD interval and the parameter with which parameter is correlated at greater absolute values than 0.3 (underlined if negative).

$\{\boldsymbol{\theta}_i\}$	$\sqrt{\widehat{\boldsymbol{R}}}$	$\overline{\boldsymbol{\theta}}$	$\widehat{\boldsymbol{\theta}}_{\boldsymbol{mode}}$	$\widehat{\boldsymbol{\theta}}_{\boldsymbol{MAP}}$	CV	95 % HPD	Correlated $\{\boldsymbol{\theta}_i\}$
1	1.00	1.95	1.87	1.99	0.12	[1.60, 2.42]	
2	1.00	-9.32	-10.00	-9.92	-0.06	[-10.00, -8.11]	$\{\theta_8\}$
3	1.01	0.58	0.54	0.61	0.14	[0.43, 0.74]	$\{\theta_6\}$ $\{\underline{\theta_9}\}$
4	1.01	0.03	0.03	0.03	0.10	[0.02, 0.03]	$\{\underline{\theta_8}\}$
5	1.02	2.27	1.05	1.63	0.73	[0, 5.40]	
6	1.01	6.69	7.01	6.69	0.10	[5.36, 7.86]	$\{\theta_3\}$
7	1.00	5.07	1.05	1.56	0.56	[0.51, 10]	
8	1.00	65.12	63.25	62.10	0.08	[54.96, 75.75]	$\{\theta_2\}$ $\{\underline{\theta_4}\}$
9	1.02	0.25	0.18	0.31	0.46	[0.06, 0.50]	$\{\underline{\theta_3}\}$

The coefficient of variation (CV) is a normalized measure of dispersion of a probability distribution defined as the ratio of the standard deviation to the mean. Three parameters $\left(\Delta K_{max}\ (\theta_5), SW_{rf}\ (\theta_7) \text{and } SW_{ret}\ (\theta_9)\right)$ stand clearly out with greatest CV values (Table 3). These parameters are also the parameters having the largest relative distance between the three point estimates, having the longest ``burn-in'' phase and having the smallest relative decrease in uncertainty when comparing the prior interval with the 95 % HPD interval. This indicates that the information from new data had little effect on these three parameters, not only on their parameter values, but on general knowledge about the parameter characterized by their posterior distribution. A sensitivity analysis done for the whole SnowFrostIce model (Thorsen et al. 2009) gave the result that changes in these three parameters also gives the smallest rate of changes in the output of the model. In summary, the results from the Bayesian calibration can tell us how new information from data are allocated within the model and accumulated as increased knowledge for some parameters while leaving others unaffected. While sensitivity analysis tells us about the sensitivity of model outputs to changes in parameters, Bayesian calibration tells us about how new information affects our knowledge about the parameters and model outputs.

Predictive uncertainty in model outputs is visualized in Figure 4 together with snow depth observations for the two test years 1998/1999 and 1999/2000. Prior and posterior uncertainties are calculated by sampling randomly 100.000 samples from the prior distribution and from the posterior chains respectively. Model outputs are then calculated for each parameter set, and the uncertainty plotted as one standard deviation above and below the mean model output for each day. Wide prior intervals were used for the parameters and we can see that the predictive posterior uncertainty is much reduced compared to the prior uncertainty for the outputs. The calibrated model approximates the data fairly closely, except for some underestimation during periods of prolonged large snow depth. Standard goodness of fit assessment is also done by constructing a predictive qq plot (Dawid 1984 and Thyer et al. 2009), checking if the predictive distribution is consistent with the observed data (Figure 5). If the cumulative distribution function (cdf) of the predictive distribution of snow cover (assumed Gaussian distributed with mean and standard deviation calculated from the 100.000 random samples drawn from the posterior chains) is independent uniform U[0,1] variables, the observations are realizations of the predictive distribution. The shape of the qq plot (Figure 5) indicates underestimated predictive uncertainty (according to Thyer et al. (2009)). We expect that this underestimation in the current case is caused by representational error,

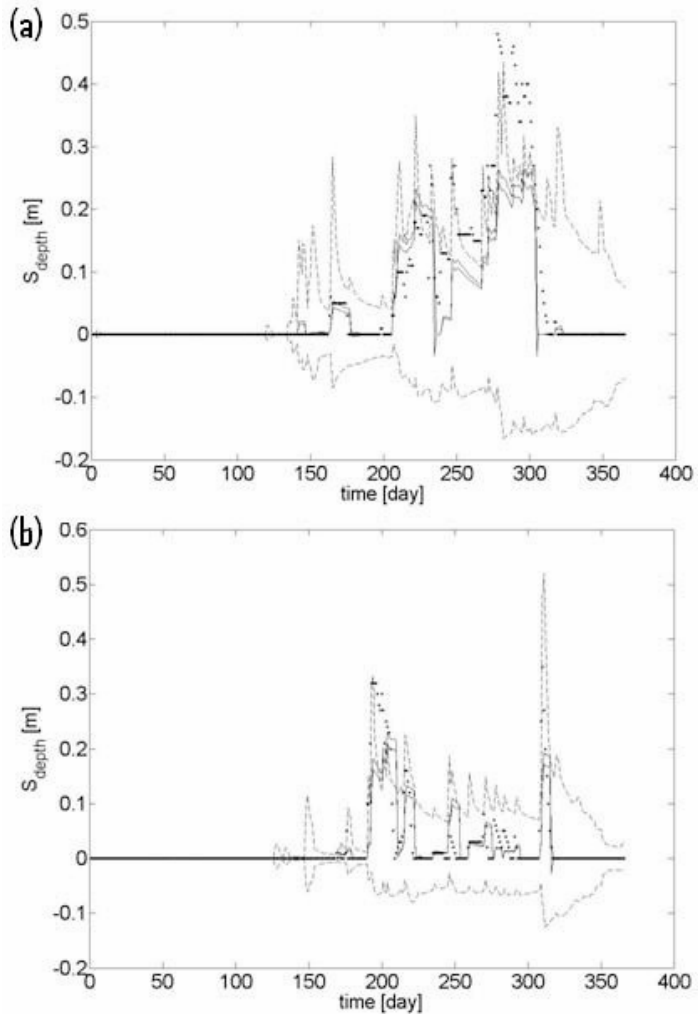


Figure 4: Prior and posterior uncertainties in the model output snow depth plotted as one standard deviation above and below the mean output from 100.000 model runs for (a) 1998/1999 (b) 1999/2000. Dotted line denotes prior uncertainty and solid line denotes posterior uncertainty. The stars denote observed values.

which we have not quantified (see introduction), but also measurement error not included in the likelihood of the data may add to the predictive uncertainty. Further more, the information in the calibration data may not span the variation range sufficiently, causing an additional parameter uncertainty (e.g. Figure 2). The posterior uncertainty is also visualized by frequency histograms of estimated snow depth values for each day during the winter 1998/1999 (Figure 6). We can see snowing and snow melt periods as changes in the snow

depth position of the frequency histograms over time. We can also see a larger uncertainty for larger snow depth values as wider frequency histograms for the larger snow depth values.

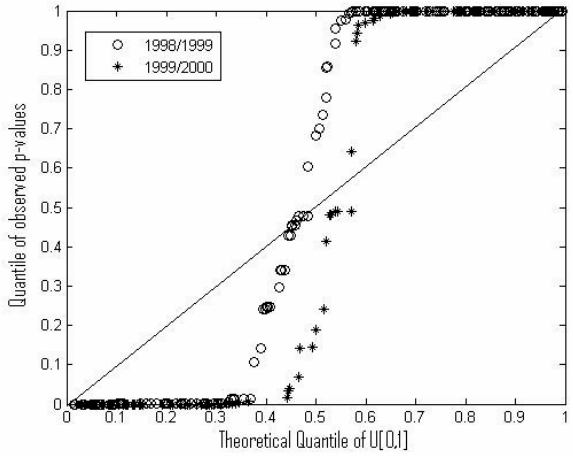


Figure 5: Predictive QQ plot.

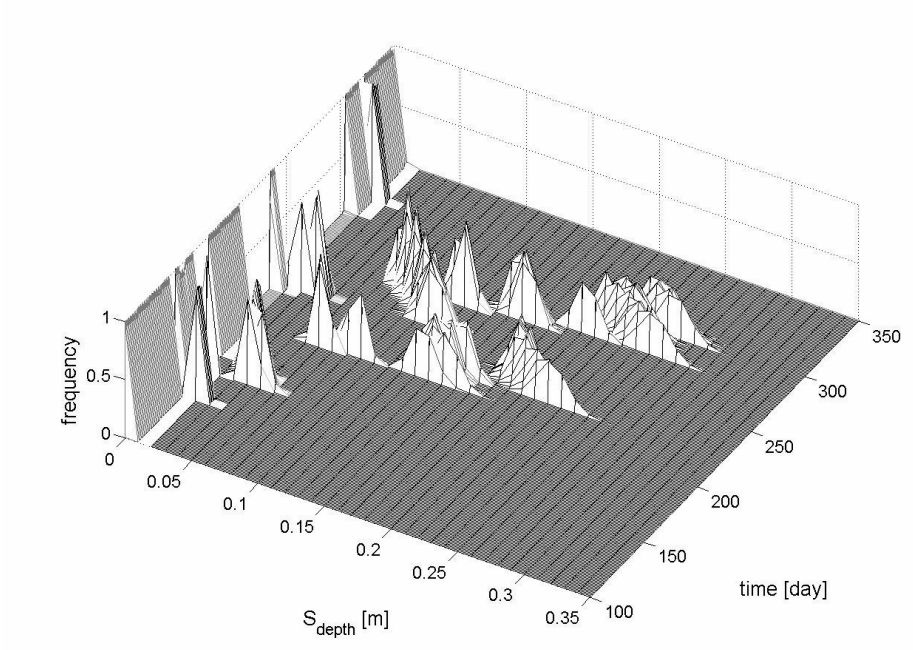


Figure 6: Posterior uncertainties on the output for the period 1. October 1998 to 1. May 1999, plotted as frequency histograms of the output each day.

CONCLUSIONS

We have used Bayesian calibration to calibrate a complex model of snow depth. The Bayesian approach regards parameters as random and prior information of the parameters is combined with observed data to form a joint posterior parameter distribution.

Here, point estimates were calculated and uncertainties visualized for both parameters and model outputs. Clearly, the uncertainty of both the parameters and the model outputs decreased when adding more data. Also, the amount of data affected the parameter estimates since the input data varied from year to year and the data used did not include the whole specter of varieties in the input space. The best fit of the model was found when using less informative priors, while more informative priors gave more meaningful physical values for the parameters. To detect ``burn-in'' both objectively and correctly, both multiple chains and Gelman Rubin were found to be useful. The choice of treating some uncertain parameters as fixed values simplified the calibration procedure, but changed the parameter estimates and led to underestimated parameter uncertainty.

We used the Markov chain Monte Carlo algorithm, random walk Metropolis. Both the idea and the implementation of the algorithm are relatively simple, but the use of the method for calibration of the complex model was far from straightforward in practice. The major problem was to obtain convergence of the chains in a limited amount of time. With regard to the prior parameter distributions, we found that informative beta distributions led to faster convergence of the posterior parameter chains than less informative uniform priors. Faster convergence was also achieved by the use of Sivia's constraint likelihood rather than the more common Gaussian likelihood function. The choice of an effective proposal distribution was difficult, but optimizing the proposal distribution with parts of the data was found to be useful. To avoid spending time on proposal parameters generated outside the prior interval, the reflection method was successfully used.

ACKNOWLEDGEMENTS

We want to thank the two anonymous referees on the first version of this paper who made extensive detailed suggestions on how to improve it. We thank them for their effort, which has been of great help in rewriting the paper.

REFRENCES

Agarwal, D. K., Gelfand, A. E., Citron-Pousty, S., 2002. Zero-inflated models with application to spatial count data. Environmental and Ecological Statistics **9 341-355**.

Andrieu, C., Thoms J., 2008. A tutorial on adaptive MCMC. Stat. Comput., **18 343-373**.

Berger, O. J., 1985. Statistical Decision Theory and Bayesian Analysis. Springer-Verlag, New York.

Brooks, S. P., 1998. General Methods for Monitoring Convergence of Iterative Simulations. Journal of Computational and Graphical Statistics **7 434-455**.

Campbell, K. 2006. Statistical calibration of computer simulations. Reliability Engineering and System Safety **91 1358-1363**.

Campolongo, F., Cariboni J., Saltelli A., 2007. An effective screening design for sensitivity analysis of large models. Environmental Modelling and Software 22 **1509-1518**.

Dawid, A. P., 1984. Statistical theory: The Prequential Approach. J. R. Stat. Soc., Ser. A, **147 278-292**

Dennis, B., 2004. Statistics and scientific method in ecology. The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations **327-378**.

Ellison, A. M., 1996. An Introduction to Bayesian Inference for Eqological Research and Environmental Decision-Making. Ecological Applocation **6 1036-1046**.

Engeseth, R. V., Sorteberg, H. K., Udnæs, H., 2000. NOSiT, utvikling av NVEs operasjonelle snøinformasjonstjeneste. Norges vassdrags- og energidirektorat.

Gelman, A., 1996. Inference and Monitoring Convergence. In: Gilks, W. R., Richardson, S., Spiegelhalter, D. J., (Eds.), Markov Chain Monte Carlo in Practice. Chapman & Hall.

Gelman, A., 2002. Prior distribution. Encyclopedia of Environmetrics **3 1634-1637**

Gelman, A., Rubin, D. B., 1992. Inference form Iterative Simulation Using Multiple Sequences. Statistical Science **7 457-511**.

Gelman, A., Bois, F., Jiang, J., 1996a. Physiological pharmacokinetic analysis using population modeling and informative prior distributions. Journal of the American Statistical Association **91 1400-1412**.

Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B., 1996b. Bayesian Data Analysis. Chapman & Hall.

Geyer C., 1992. Practical Monte Carlo Markov chain (with discussion). Statistical Science **7 473-511**.

Goldstein, M., Rougier, J., 2006. Bayes Linear Calibrated Prediction for Complex Systems. Journal of the Americal Statistical Association **101 1131-1143**.

Hue, C., Tremblay, M., Wallach, D. J., 2001. A Bayesian Approach to Crop Model Calibration Under Unknown Error Covariance. Journal of the Americal Statistical Association **101 355-365.**

Jeffreys, H., 1961. Theory of probability. Oxford University Press

Kass, R. E., Carlin, B. P., Gelman, A., Neal, R. M., 1998. Markov Chain Monte Carlo in Practice: A Roundtable Discussion. The American Statistican **52 93-100**.

Kavetski, D., Kuczera, G. & Franks, S.W. 2006. Bayesian analysis of input uncertainty in hydrological models: 1. Theory. Water Resources Research 42: W03407.

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

Liu, J. S., 2001. Monte Carlo Strategies in Scientific Compution. Springer-Verlag, New York.

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

Marchall, L., Nott, D., Sharma, A., 2004. A comparative study of Markov chain Monte Carlo methods for conceptual rainfall-runoff modeling. Water Resource Research. **40**

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

Miller, I., Miller, M., 1999. Freund's Mathematical Statistics. Prentice-Hall, Upper Saddle River.

Reinds, G.J., Van Oijen, M., Heuvelink, G.B.M. & Kros, H. 2008. Bayesian calibration of the VSD soil acidification model using European forest monitoring data. Geoderma **146: 475-488.**

Riley, H., Bonesmo, H., 2005. Modelling of snow and freezing-thaw cycles in the EU-rotate_N secision support system. Grønn kunnskap **9,1-8**.

Roberts, B. O., 1996. Markov Chain concepts related to sampling algorithms. In: Gilks, W. R., Richardson, S., Spiegelhalter, D. J., (Eds.), Markov Chain Monte Carlo in Practice. Chapman & Hall.

Roberts, G. O., Gelman, A., Gilks, W. R., 1997. Weak convergence and optimal scaling of random walk Metropolis algorithms. The Annals of Applied Probability **7 110-120**.

Rougier, J., 2007. Probabilistic Inference for Future Climate Using an Ensemble of Climate Model Evaluations. Climatic Change **81 247-264**.

Sivia, D. S., 2006. Data Analysis, A Bayesian Turtorial. Oxford University press.

Smith, T.J. & Marshall, L.A. 2008. Bayesian methods in hydrologic modeling: A study of recent advancements in Markov chain Monte Carlo techniques. Water Resources Research 44: W00B05.

Thorsen, S. M., Haugen, L. E., 2007. Development of the SnowFrost model for the simulation of Snow Fall and Soil Frost. Bioforsk FOKUS **2 1-23**.

Thorsen, S. M., Roer, A-G., Van Oijen, M., 2009. Modelling the dynamics of snow cover, soil frost and surface ice in Norwegian grasslands. Submitted for publication.

Thyer, M., Renard, B., Kavetski, D., Kuczera, G., Franks, S.W. & Srikanthan, S. 2009. Critical evaluation of parameter consistency and predictive uncertainty in hydrological

modeling: A case study using Bayesian total error analysis. Water Resources Research 45: W00B14.

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

Van Oijen, M., Höglind, M., Hanslin, H. M., Caldwell, N., 2005b. Process-Based Modelling of Thimothy Regrowth. Agronomy Journal **97 1295-1303**.

Vehviläinen., 1992. Snow cover models in operational watershed forecasting. PhD Thesis, Helsinki University of Technology.

Yang, Z., 2006. Computational Molecular Evolution. Oxford University press.

II

Polar Research

Modelling the dynamics of snow cover, soil frost and surface ice in Norwegian grasslands

Stig Morten Thorsen,[1,2] Anne-Grete Roer[3] & Marcel van Oijen[4]

1 Grassland and Landscape Division, Norwegian Institute for Agricultural and Environmental Research, Postvegen 213, NO-4353 Klepp Stasjon, Norway
2 Department of Mathematical Sciences and Technology, Norwegian University of Life Sciences, PO Box 5003, NO-1432 Ås, Norway
3 Plant Health and Plant Protection Division, Norwegian Institute for Agricultural and Environmental Research, Høgskoleveien 7, NO-1432, Norway
4 Centre for Ecology and Hydrology, Edinburgh, Natural Environment Research Council, Bush Estate, Penicuik, EH26 0QB, UK

Keywords
Calibration; frost depth; ice cover; modelling; sensitivity analysis; snow cover.

Correspondence
Stig Morten Thorsen, Grassland and Landscape Division, Norwegian Institute for Agricultural and Environmental Research, Postvegen 213, NO-4353 Klepp Stasjon, Norway. E-mail: stig.thorsen@bioforsk.no

doi:10.1111/j.1751-8369.2010.00157.x

Abstract

Studying the winter survival of forage grasses under a changing climate requires models that can simulate the dynamics of soil conditions at low temperatures. We developed a simple model that simulates depth of snow cover, the lower frost boundary of the soil and the freezing of surface puddles. We calibrated the model against independent data from four locations in Norway, capturing climatic variation from south to north (Arctic) and from coastal to inland areas. We parameterized the model by means of Bayesian calibration, and identified the least important model parameters using the sensitivity analysis method of Morris. Verification of the model suggests that the results are reasonable. Because of the simple model structure, some overestimation occurs in snow and frost depth. Both the calibration and the sensitivity analysis suggested that the snow cover module could be simplified with respect to snowmelt and liquid water content. The soil frost module should be kept unchanged, whereas the surface ice module should be changed when more detailed topographical data become available, such as better estimates of the fraction of the land area where puddles may form.

Grasslands are important components of Norwegian terrestrial ecosystems. In order to investigate the impacts of climate change, parts of the Norwegian research programme Climate Change Effects on Winter Survival of Perennial Forage Crops and Winter Wheat, and Plant Diseases and Weed Growth and Control at High Latitudes (WINSUR) are dedicated to developing a grassland model to study the winter survival of different crops. The grassland model, currently simulating the regrowth dynamics of timothy (*Phleum pratense* L.), has been developed by van Oijen, Höglind et al. (2005). The same model will be adapted to simulate the regrowth dynamics of perennial ryegrass (*Lolium perenne* L.). During the winter, a significant number of plants may die as a result of frost, ice encapsulation, and other physical and biological stresses (Larsen 1994). Snow cover provides insulation from lethal freezing temperatures, while also reducing the amount of photosynthetically active radiation at plant level. However, a more variable winter climate in Norway (Beldring et al. 2008) may lead to less snow cover, and may thereby increase plant exposure to killing frosts (Bélanger et al. 2002).

If the ground is frozen, water (rain or snowmelt) can accumulate in small depressions, freeze and cause plants to be encapsulated in ice. Ice encasement can severely reduce gas exchange between the plant and the surrounding atmosphere, leading to a transition from aerobic to anaerobic respiration, and to the accumulation of respiration products (especially CO_2) to toxic levels (Gudleifsson & Larsen 1993).

In order to make predictions about the effects of climate change on plant performance over more than one growing season, the grassland model needs an additional set of functions to describe the winter survival of the sward. The grassland model must be able to simulate effects of winter climate on soil and soil surface processes. The main objective of this work is to develop a simple winter module that can easily be incorporated into the

existing grassland model. Therefore, the structure of the winter module needs to be kept as simple as possible.

Regarding the simulation of winter climate effects on soil and soil surface conditions (e.g., snow cover and soil frost), the literature provides examples of different approaches (Benoit & Mostaghimi 1985; Flerchinger & Saxton 1989; Jordan 1991; Vehvilainen 1992; Melloh 1999; Engeset et al. 2000; DeGaetano et al. 2001; Jansson & Karlberg 2001; Kokkonen et al. 2006). We implemented and tested different algorithms for snow cover and soil frost that have already been applied to Nordic conditions. Based on preliminary modelling work, including site-specific model calibration, we developed a new snow module using ideas from a snow model currently being used by the Norwegian Water Resources and Energy Directorate (NVE) (Engeset et al. 2000). The NVE model has 10 parameters, and is used throughout Norway for operational snow forecasts. This model simulates snow accumulation based on daily precipitation rates and daily mean air temperature. Snowmelt is a function of a degree–day temperature index, described by a sinusoidal curve and daily mean air temperature. The NVE model is mainly designed for hydrological purposes (hydroelectricity production and spring flood warnings), and thus simulates the liquid water equivalent of snow water equivalent (SWE) (mm) and snowmelt run-off, but not snow depth.

Different models for simulating snow accumulation and snowmelt are described in the literature, ranging from hydrological (Jordan 1991; Engeset et al. 2000; Kokkonen et al. 2006) to combined agricultural and hydrological applications (Flerchinger & Saxton 1989) and soil–plant–atmosphere systems (Jansson & Karlberg 2001). These models simulate point estimates of a single-layered homogeneous one-dimensional (z-direction) snow cover, whereas Jordan (1991) presents a multi-layered one-dimensional snow model. Melloh (1999) provides a review of several snowmelt models. Comprehensive state-of-the-art snow models such as the COUP model (Jansson & Karlberg 2001) with graphical user interface and the SNTHERM model (FORTRAN-77 code; Jordan 1991) are very complex and rich in parameters (>100). The COUP model was considered as a potential candidate early in the project, but the model version available at that time required a special graphical user interface, and therefore could not be incorporated into the grassland model, which was developed using another programming environment (MATLAB and Simulink). The combination of a special user interface and extensive data requirements (as regards number of parameters and driving climate variables) makes it very difficult to incorporate state-of-the-art snow cover and soil frost models as sub-modules into other models. The ability to incorporate a snow and soil frost model into a larger grassland model was our main motivation for developing a new model. Our proposed model is simple: it only requires nine calibrated parameters and two input variables to simulate daily values of the depths of snow cover, soil frost and surface ice, and the temperature between the soil surface and the snow cover.

A study comparing four models simulating soil frost (Kennedy & Sharratt 1998)—the two finite difference models SHAW and SOIL, and two energy balance models—concluded that the simpler energy balance models generally overestimate the frost depth. However, one weakness of all four models compared by Kennedy & Sharratt (1998) is the estimation of snow depth (one of the energy balance models uses snow depth as an input). Snow cover has a strong influence on the estimation of soil frost depth, e.g., through snow depth and snow density, with both affecting the thermal conductivity of the snow cover. Therefore, accurate simulation of snow cover is important for the simulation of soil frost depth.

As regards modelling the formation of ice on the soil surface, we did not find examples in the literature of models simulating this process or ice encapsulation of the ground vegetation.

Following the conclusions by Kennedy & Sharratt (1998), the present work describes a new model that simultaneously simulates the depths of snow, soil frost and surface ice, and explains how it was calibrated for sites across Norway using Bayesian methods. We also conducted a sensitivity analysis of the model using the Morris method, which identifies the parameters to which the model is most sensitive.

Materials and methods

The snow model

Our snow module is based on ideas presented by Melloh (1999, and references therein) and Engeset et al. (2000). Whereas snow models used for hydrological purposes usually simulate SWE, SnowFrostIce also simulates the actual depth of the snow cover S_{depth} (m). To run SnowFrostIce, the only required meteorological inputs are daily values of mean air temperature T_{air} (°C) and precipitation rate P (mm d^{-1}). The parameters, which need to be locally calibrated, are listed in Table 1.

In SnowFrostIce, the precipitation form is determined by a threshold temperature T_{rs} (°C). If $T_{\mathrm{air}} > T_{\mathrm{rs}}$, precipitation falls as rain, P_{r} (mm d^{-1}). Otherwise it falls as snow P_{s} (mm day^{-1}), with density ρ_{ns} (kg m^{-3}). There is no intermediate form for sleet. The snow cover consists of water in solid state S_{dry} (mm) (snow and ice), and liquid state S_{wet} (mm). The threshold temperature T_{mf} (°C) determines

Table 1 Parameter description for the SnowFrostIce model. $\boldsymbol{\theta}^{min}$ and $\boldsymbol{\theta}^{max}$ represent parameter lower and upper boundaries; $\boldsymbol{\theta}^{mode}$ and $\boldsymbol{\theta}^{def}$ represent parameter mode and default values, respectively. When $\boldsymbol{\theta}_i^{mode}$ values are presented, a beta prior distribution is used for parameter $\boldsymbol{\theta}_i$, otherwise a uniform prior distribution is assumed between $\boldsymbol{\theta}^{min}$ and $\boldsymbol{\theta}^{max}$.

Symbol	Unit	θ_i^{min}	θ_i^{max}	θ_i^{mode}	θ_i^{def}	References
T_{rs}	°C	–5	5	0.5	0.5	Engeset et al. (2000)
T_{mf}	°C	–5	5	0.5	0.5	Engeset et al. (2000)
ξ	mm mm^{-1} day^{-1}	0	1	—	0.02	Thorsen & Haugen (2007)
ΔK_{max}	mm °C^{-1} day^{-1}	0	5	1.25	1.25	Engeset et al. (2000)
K_{min}	mm °C^{-1} day^{-1}	0	5	2	2	Engeset et al. (2000)
SW_{rf}	mm °C^{-1} day^{-1}	0	5	0.01	0.01	Engeset et al. (2000)
ρ_{ns}	kg m^{-3}	10	250	—	100	Judson & Doesken (2000)
SW_{ret}	mm mm^{-1}	0	1	0.1	0.1	Engeset et al. (2000)
λ_{fs}	J m^{-1} °C^{-1} day^{-1}	8.6×10^4	21.6×10^4	—	17.3×10^4	Jansson & Karlberg (2001)

whether snow is in the process of melting M (mm day^{-1}), when $T_{air} > T_{mf}$, or when liquid water within the snow cover is in the process of refreezing M_{rf} (mm day^{-1}), when $T_{air} < T_{mf}$. The numerical values of T_{rs} and T_{mf} are sampled from the posterior distribution obtained in the Bayesian calibration. As the model is calibrated locally, the estimates of T_{rs} and T_{mf} are different for each location. Instead of using a constant melt rate (mm snowmelt per degree celsius and day, also known as the degree–day temperature index method), we use a degree–day temperature index K (mm °C^{-1} day^{-1}), which is described by a sinusoidal curve (see Eqn. 4). The reason for describing K by a sinusoidal curve is to incorporate the seasonal variation. Incoming radiation increases and the albedo of the snow cover decreases in the spring. Thus K increases in spring. In Norway, located between latitudes 58° and 71°N in the Northern Hemisphere, the dates of the solstice are 21 December and 21 June. The sinusoidal curve is therefore defined as having a period of 1 year, with a trough, termed K_{min} (mm °C^{-1} day^{-1}), on 21 December, and a crest, termed K_{max} (mm °C^{-1} day^{-1}), on 21 June. The simulated snowmelt intensity M is proportional to the number of degrees above T_{mf} (see Eqn. 5). To avoid situations such as $K_{max} < K_{min}$ during the calibration, we replaced K_{max} by $\Delta K_{max} = K_{max} - K_{min}$, and calibrated ΔK_{max} (mm °C^{-1} day^{-1}) (see Table 1).

Liquid water within the snow cover may refreeze. The simulated refreezing intensity M_{rf} is proportional to the number of degrees below T_{mf} (see Eqn. 6), where SW_{rf} (mm °C^{-1} day^{-1}) is the degree–day temperature index for refreezing. We calculated the potential retention capacity of the snow cover as SW_{ret}*S_{dry}, where SW_{ret} (mm mm^{-1}) is the retention capacity of the snow cover. The snow water equivalent, SWE, is defined as the sum of S_{dry} and S_{wet}, and the density of the snow cover ρ_s (kg m^{-3}) is defined as SWE/S_{depth}. As snow is accumulated on the surface of the snow cover, there is a rapid metamorphosis as snow crystals break down, and at lower snow depths densification occurs at a slower rate, which is largely determined by the overburden pressure (Gray & Morland 1995). In SnowFrostIce we make the assumption that the combined effects of the metamorphosis of snow crystals and the densification of the lower snow layers is captured by the empirical compaction parameter ξ (mm mm^{-1} day^{-1}). We use the following equations (1–6) to describe the snow cover dynamics.

$$\frac{\Delta S_{dry}}{\Delta t} = P_s + M_{rf} - M \tag{1}$$

$$\frac{\Delta S_{wet}}{\Delta t} = P_r + M - M_{rf} \tag{2}$$

$$\frac{\Delta S_{depth}}{\Delta t} = \frac{P_s}{\rho_{ns}} - \frac{M}{\rho_s} - \xi S_{depth} \tag{3}$$

$$K = \frac{\Delta K_{max}}{2} \sin\left(\frac{2\pi t}{365} + \frac{3}{8}\pi\right) + \left(K_{min} + \frac{\Delta K_{max}}{2}\right) \tag{4}$$

$$M = K(T_{air} - T_{mf}) \tag{5}$$

$$M_{rf} = SW_{rf}(T_{mf} - T_{air}) \tag{6}$$

The snow model parameters to be calibrated are listed in Table 1.

The soil frost model

When modelling soil frost we use an energy balance approach. Our simple approach does not include an annual energy budget for the soil system. SnowFrostIce simulates only the lower frost boundary F_{depth} (m), resulting in one frozen soil layer ranging from the soil surface to F_{depth}. For the soil–water balance, we use the routines implemented in the grassland model by Höglind et al. (2001) to obtain daily values of available soil water content x_w (m^3 m^{-3}) (i.e., what is left from surplus liquid water after transpiration and evaporation is subtracted),

which is used in the calculation of F_{depth}. The soil layer is parameterized as in the grassland model. SnowFrostIce requires site-specific soil type parameters for soil water retention, but the only soil parameter to be calibrated is the thermal conductivity of the frozen soil λ_{fs} (J m^{-1} °C^{-1} day^{-1}).

Our way of estimating the lower frost boundary F_{depth} is based on certain assumptions. Regarding surface temperature, we follow along the lines of the assumption made by Benoit & Mostaghimi (1985), that in any given 24-h period, the mean surface temperature of the soil or snow cover can be approximated by the daily mean air temperature for that same period. However, instead of using the daily mean air temperature at the snow cover surface when calculating F_{depth}, like Benoit & Mostaghimi (1985), whenever a snow cover is simulated we use a simulated soil surface temperature T_{surf} (°C) from Eqn. 15 as an approximation to the soil surface temperature to incorporate the insulating effect of the snow cover. (Note to Eqn. 7: during snow-free periods we assume T_{surf} can be approximated by T_{air}.) We assume a unidirectional stationary flow of heat between F_{depth} and the soil surface, ignoring additional heat from, e.g., lower unfrozen soil layers, percolating water, radiation and no freeze-point depression. We further assume a linear variation in soil temperature $T(z)$ (°C) with respect to soil depth z (m) in the frozen soil layer, and that all available soil water x_{w} within this layer freezes. It is the temperature difference between the soil surface and F_{depth} that drives the process of soil frost formation in the model:

$$T(z) = T_{surf} + z\frac{T^* - T_{surf}}{F_{depth}} \tag{7}$$

where T_{surf} is the simulated temperature just above the soil surface, T^* (°C) is the temperature where soil water freezes (we assume $T^* = 0$ °C). Following the assumption regarding $T(z)$, Eqn. 7 is only valid when $F_{\text{depth}} > 0$. We denote the heat flux density released when the soil water freezes Q_{E} (J m^{-2} day^{-1}). Following an existing idea (Thorsen & Haugen 2007), we express Q_{E} using the above assumptions as:

$$Q_E = -x_w \rho_w L_f \frac{\partial F_{depth}}{\partial t} \tag{8}$$

where x_{w} is available soil water content, ρ_{w} (1000 kg m^{-3}) is the density of water and L_{f} (335 kJ kg^{-1}) is the latent heat of fusion. When the soil cools down during autumn and winter, the heat released (Q_{E}) when the soil frost penetrates deeper into the soil is transported through the previously frozen soil. Using Fourier's equation for heat transport in one-dimensional form, we express the heat transport through the frozen soil, termed Q_{fs} (J m^{-2} day^{-1}), as:

$$Q_{fs} = -\lambda_{fs}\frac{\partial T(z)}{\partial z} \tag{9}$$

From the assumption of linear variation in soil temperature $T(z)$ with depth z in frozen soil, we obtain $\frac{\partial T(z)}{\partial z}$ from Eqn. 7, and insert this into Eqn. 9:

$$Q_{fs} = -\lambda_{fs}\frac{T^* - T_{surf}}{F_{depth}} \tag{10}$$

Equating Eqns. 8 and 10 and using the assumption $T^* = 0$ °C, we obtain an algebraic expression for the rate of change in F_{depth}:

$$\frac{\partial F_{depth}}{\partial t} = -\frac{\alpha}{F_{depth}} \tag{11}$$

where $\alpha = \frac{\lambda_{fs} T_{surf}}{x_w \rho_w L_f}$. If we neglect the diurnal variation in T_{surf} and x_{w}, and consider Eqn. 11 as $\frac{\mathrm{d}F_{depth}}{\mathrm{d}t} = -\frac{\alpha}{F_{depth}}$, by solving this equation we can express the daily increase in frost depth as $F_{depth}^{(t+1)} = \sqrt{\left(F_{depth}^{t}\right)^2 - 2\alpha}$. Provided $\left(F_{depth}^{(t)}\right)^2 - 2\alpha > 0$, we can express the rate of change in F_{depth} as follows:

$$\frac{\Delta F_{depth}}{\Delta t} = \frac{1}{\Delta t}\left(\sqrt{\left(F_{depth}^{(t)}\right)^2 - 2\alpha} - F_{depth}^{(t)}\right) \tag{12}$$

The presence of snow cover has an insulating effect on the soil. Following Jansson & Karlberg (2001), we assume a steady state heat flow through the frozen soil layer and the snow cover. The heat flux density through the frozen soil Q_{fs} from Eqn. 10 thereby equals the heat flux density through the snow cover Q_{snow} (J m^{-2} day^{-1}):

$$Q_{fs} = Q_{snow} \tag{13}$$

$$-\lambda_{fs}\frac{T^* - T_{surf}}{F_{depth}} = -\lambda_s \frac{T_{surf} - T_{air}}{S_{depth}} \tag{14}$$

where λ_{s} (J m^{-1} °C^{-1} day^{-1}) is the thermal conductivity of the snow cover. The parameter λ_{s} is treated as a constant, and is not calibrated. According to Jansson & Karlberg (2001), a reasonable estimate for the ratio $\lambda_{\text{fs}}/\lambda_{\text{s}}$ in our situation is $\lambda_{\text{fs}}/\lambda_{\text{s}} \approx 10$. We rearrange the above equation to derive the following approximation of T_{surf}:

$$T_{surf} \approx T_{air} / (1 + 10(S_{depth} / F_{depth})) \tag{15}$$

(Note: for the calculations, $F_{\text{depth}} > 0$ when soil frost is present.) In the case of an existing snow cover but no soil frost ($F_{\text{depth}} = 0$), we assume T_{surf} to lie around 0 °C. This assumption is in accordance with observations made by Iwata et al. (2008), and it is incorporated by an additional

Table 2 Locations in Norway used for calibrating and validating the SnowFrostIce model. The fifth location, Karasjok, was only included in the validation, and was not used in the calibration.

Location	Grid	Elevation (m a.s.l.)	Climate	Measurement calibration	Period validation
Kise	60°77′N, 10°8′E	127	Interior, lake	1993–96	1996–99
Kvithamar	63°49′N, 10°88′E	40	Coastal	2001–03	2003–05
Vågønes	67°28′N, 14°45′E	30	Coastal	1998–2001	2001–03
Holt	69°65′N, 18°91′E	20	Coastal	1996–99	2005–07
Karasjok	69°28′N, 25°31′E	149	Interior	—	1998–99

empirical expression preserving the insulating effect of the snow cover:

$$T_{surf} \approx T_{air} e^{(-\gamma S_{depth})} \quad (16)$$

where the empirical parameter γ (m^{-1}) is set to 65. This γ parameter is not calibrated.

Puddle formation and infiltration of meltwater

As we were unable to obtain topographical information for any location during this study, we assume the hypothetical field of interest to be an even, rectangular surface sloping at a low angle towards a water-blocking barrier at the lower end. The height of this barrier determines the maximum depth of the surface puddle. This maximum storage level is set to 50 mm. Baker & Spaans (1997) report that infiltration from puddles can occur despite the presence of a frozen soil layer of 20–40 cm. Based on this observation, surface water (snowmelt and rain) in SnowFrostIce is allowed to infiltrate into the soil if $F_{\text{depth}} < 20$ cm. This assumption is also confirmed by Iwata et al. (2008). In reality, the surface water transfers heat to the soil, and because the frozen soil initially remains cold this may create a thin ice layer at the soil surface, which impedes water infiltration and increases surface run-off (Stähli et al. 2004). Therefore, when F_{depth} penetrates below 20 cm, we assume that the soil becomes impermeable to any further infiltration, and that the surface water is re-directed to the puddle area. If the maximum depth of the barrier at the end of the field is exceeded, the additional surface water runs off. When the soil starts thawing we let the infiltration rate of the puddle water follow the thawing rate (in accordance with observations by Hayashi et al. [2003]), until $F_{\text{depth}} > 20$ cm, when the remaining puddle water is drained as if the soil were unfrozen.

Formation of ice layer

When a surface puddle is formed, the water may freeze and form a basal ice layer. By regarding the puddle as an extremely dilute soil, and setting the water content to unity, we use the same approach to calculate I_{depth} (mm) as we do for the soil frost. Provided $(I_{depth}^{(t)})^2 - 2\beta > 0$, we get the following expression for the daily change in I_{depth}:

$$\frac{\Delta I_{depth}}{\Delta t} = \frac{1}{\Delta t}\left(\sqrt{(I_{depth}^{(t)})^2 - 2\beta} - I_{depth}^{(t)}\right) \quad (17)$$

where $\beta = \frac{\lambda_i T_{surf}}{\rho_w L_f}$, the thermal conductivity of ice is λ_i (19.4×10^4 J m^{-1} °C^{-1} day^{-1}), the density of water is ρ_w and the latent heat of fusion is L_f.

Description of the locations and data used in calibration

The SnowfrostIce model was calibrated using observed depths of snow cover and the lower frost boundary. The snow cover depth was measured in cm in accordance with the Norwegian Meteorological Institute. The depth of the lower frost boundary was measured in cm using a frost tube, as described by DeGaetano et al. (2001) and Iwata et al. (2008). We were unable to obtain information on the accuracy of the observations. We were also unable to obtain information on normal depths of snow cover and soil frost. We therefore present values of mean air temperature and precipitation sums from autumn to spring, and frost sums. Table 2 presents a geographical description of the locations, and Tables 3–6 provide a summary of the climate for each location for the current normal period in Norway (1961–1990), and for the calibration and validation periods. For each location we calculated the following from autumn to spring (i.e., from 1 September to 30 April): the mean 2 m air temperature, denoted as mean(T_{air}); the temperature sum for days when $T_{\text{air}} < 0$, denoted as ΣT_{air}; and the sum of daily precipitation rates, denoted as ΣPrec.

During the calibration period at Kise (Table 3), the first and third winters were both colder and had more frost compared with the normal period. The second winter was milder and had less frost. The first winter received more precipitation compared with the normal period, whereas the latter two winters were dryer. In the validation period, all winters were slightly milder and had less frost than normal: the first winter was dryer than normal, whereas the latter two were wetter.

Table 3 Climate summary for Kise. Values are calculated for the months September–April for the current normal period in Norway (1961–1990), and for the respective calibration and validation periods. Mean(T_{air})(°C) is the average 2-m air temperature, ΣT_{air} (°C day) is the temperature sum on frost days and ΣPrec (mm) is the recorded precipitation.

Sept–Apr	1961–90	1993/94	1994/95	1995/96	1996/97	1997/98	1998/99
Mean(T_{air})	–1	–2.5	1.1	2.2	0.8	1.2	0.2
ΣT_{air}	–761	–1068	–400	–1214	–629	–439	–611
ΣP_{rec}	340	368	294	188	273	421	436

Table 4 Climate summary for Kvithamar.

Sept–Apr	1961–90	2001/02	2002/03	2003/04	2004/05
Mean(T_{air})	1.5	3.2	1.7	3.2	3.3
ΣT_{air}	–269	–272	–385	–245	–225
ΣP_{rec}	597	682	508	604	891

See Table 3 for abbreviations.

Table 5 Climate summary for Vågønes.

Sept–Apr	1961–90	1998/99	1999/2000	2000/01	2001/02	2002/03
Mean(T_{air})	1.3	1.9	2.6	2.3	2.6	1.2
ΣT_{air}	–284	–323	–264	–368	–330	–372
ΣP_{rec}	811	902	1156	561	983	735

See Table 3 for abbreviations.

Table 6 Climate summary for Holt and Karasjok.

Sept–Apr	1961–90	1996/97	1997/98	1998/99	2005/06	2006/07	1998/99*
Mean(T_{air})	–0.8 (–8.3)	–0.1	0.2	0.3	1.6	1.1	–8.5
ΣT_{air}	–375 (–2199)	–468	–483	–432	–317	–322	–2295
ΣP_{rec}	765 (172)	804	627	578	831	817	207

Values within brackets represent the normal period for the Karsjok location.

* 1998/99 represents the Karasjok location.

See Table 3 for abbreviations.

At Kvithamar (Table 4), both winters in the calibration period were milder than normal, but they had more frost. The first winter was wetter, and the second winter was dryer than normal. In the validation period, both winters were milder and wetter compared with the normal period.

At Vågønes (Table 5), all winters in the calibration period were milder compared with the normal period, but the first and third winters had more frost days, whereas there were fewer frost days in the second winter. The first two winters were wetter, and the third winter was dryer than normal. In the validation period, both winters had more frost than normal, but only the first winter was milder than normal. The first winter was wetter than normal, and the second was dryer.

At Holt (Table 6), all winters in the calibration period were milder and had more frost than normal. The first winter was wetter, whereas the latter two winters were dryer than normal. Both winters in the validation period were milder, had less frost and were wetter when compared with the normal period.

The winter in the validation period at Karasjok (Table 6) was approximately the same as the normal period, but slightly wetter.

In addition to simulating S_{depth} and F_{depth}, SnowFrostIce simulates the thickness of ice (I_{depth}) resulting from the freezing of soil surface puddles. However, data on surface ice were scarce, and there was no description of field topography available, forcing us to make assumptions on field topography. We therefore present full simulation results for only two locations: Holt in Troms county and Karasjok in Finnmark county. Based on data availability, we chose four locations for site-specific calibration of the model spanning the south–north variation in regional climate. Table 2 gives a brief description of these locations. Karasjok was not included in the calibration.

Observations of surface ice cover were scarce, and data were only available for two sites: Holt (1997/98 and

1998/99) and Karasjok (1998/99). Ice observations from Holt came at a later stage in the project, so we had to use observations on snow cover and frost depth from the calibration period.

Bayesian calibration of the SnowFrostIce model

The SnowFrostIce model represents a simplification of different physical processes. Parameters used in process-based models have a physical meaning, but these are seldom precisely known, or are at best difficult to measure. We represented this uncertainty as a probability distribution over the parameters. Thus, if we define a parameter vector $\boldsymbol{\theta}$ for the model, then $\pi(\boldsymbol{\theta})$ is said to be a joint probability density function (pdf) expressing our initial prior belief in the parameters. Given a data set $\mathbf{D}$ of model outputs, we update the joint pdf of the parameters by applying the Bayes theorem: $\pi(\boldsymbol{\theta}|\mathbf{D}) = \pi(\boldsymbol{\theta}) f(\mathbf{D}|\boldsymbol{\theta}) / f(\mathbf{D})$, where $\pi(\boldsymbol{\theta}|\mathbf{D})$ is the posterior distribution of $\boldsymbol{\theta}$ given the data $\mathbf{D}$, $f(\mathbf{D}|\boldsymbol{\theta})$ is the likelihood of the data given the model outputs using parameters $\boldsymbol{\theta}$, and $f(\mathbf{D})$ is a normalization constant. In the Bayesian calibration of dynamic models, a large number of model runs are carried out, often using a Markov chain Monte Carlo (MCMC) approach. We used the MCMC algorithm known as the Metropolis Random Walk. For further details on using Bayesian methods to calibrate complex models see van Oijen, Rougier et al. (2005). The target posterior distribution was the stationary distribution of the Markov chain produced by the Metropolis Random Walk.

Metropolis Random Walk

The general idea of the Metropolis Random Walk is to walk randomly through the parameter space, running the model at each visited point, eventually forming a Markov chain. The starting point of this chain, $\boldsymbol{\theta}_0$, is randomly chosen from the prior distributions for the parameters. A new proposal parameter vector $\boldsymbol{\theta}'$ is then chosen based on the current parameter vector $\boldsymbol{\theta}^t$:

$$\boldsymbol{\theta}' = \boldsymbol{\theta}^t + \boldsymbol{\delta} \tag{18}$$

where $\boldsymbol{\delta}$ is the step length vector. It is also important that $p(\boldsymbol{\delta}) = p(-\boldsymbol{\delta})$, i.e., that there is an equal probability of stepping in either direction from the current point. We then compute the so-called Metropolis ratio:

$$r = \frac{\pi(\boldsymbol{\theta}'|\mathbf{D})}{\pi(\boldsymbol{\theta}^t|\mathbf{D})} = \frac{\pi(\boldsymbol{\theta}') f(\mathbf{D}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}^t) f(\mathbf{D}|\boldsymbol{\theta}^t)} \tag{19}$$

The next step is to generate a uniform random number $u \sim U(0,1)$, and to accept the proposal parameter vector $\boldsymbol{\theta}'$ as the new $\boldsymbol{\theta}^{t+1}$ if $u \leq r$. Otherwise, let $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t$. The chain consisting of all $\boldsymbol{\theta}^t$ forms our Markov chain, which is our sample from the posterior distribution.

The posterior distribution is therefore a combination of prior knowledge and new information obtained from the data using the likelihood function. Measurement errors are used in the determination of how likely a model–data mismatch might be, i.e., if the data are informative and have a sharply peaked distribution (i.e., a small variance), the resulting posterior distribution will be narrower and more peaked than the prior distribution. This indicates that the parameter uncertainty is reduced.

Defining prior probability distributions of the parameters

Based on a literature review, we defined the likely ranges $[\theta_i^{\min}, \theta_i^{\max}]$ and mode values for the nine parameters. For parameters where range and mode value were suggested, we used a beta distribution as prior. A suitable range was only found for parameters ξ, ρ_{ns} and λ_{fs}. For these three parameters we selected a flat uniform distribution within their range $[\theta_i^{\min}, \theta_i^{\max}]$. In the calibration process we assumed the parameters to be independent a priori, implying that their joint prior distribution is equal to the product of their individual marginal pdfs. The parameters, along with their prior distributions, are presented in Table 1.

Defining the data-likelihood function

We used measurements on snow depth and lower frost boundary for the calibration of SnowFrostIce. Specific information about the precision of the measurements was not available, so we used the same approach as van Oijen, Rougier et al. (2005), and chose the standard deviation of each measurement to be 30% of the mean value. To avoid a standard deviation of zero (if the observed variable was zero), the standard deviation was redefined as $\sigma_{ij}^o = \max(0.1; 0.3 \cdot \mathrm{D_{ij}})$ where $\mathrm{D_{ij}}$ are the measurements on output j at time i. Assuming the measurement errors to be independent and Gaussian, we used Sivia's (2006) formulation, which was slightly modified to account for model discrepancy:

$$f(\mathbf{D}|\boldsymbol{\theta}) = \prod_{j=1}^{G} \left[\prod_{R_{ij} \neq 0} \prod_{i=1}^{M} \frac{1}{\sigma_{ij}^o \sqrt{2\pi}} \left[\frac{1 - \mathrm{e}^{-R_{ij}^2/2}}{R_{ij}^2} \right] \prod_{R_{ij}=0} \frac{1}{2\sigma_{ij}^o \sqrt{2\pi}} \right] \tag{20}$$

where σ_{ij}^o represent the lower bounds on the data noise, and the residual is represented by $R_{ij} = (\mathrm{D}_{ij} - M_{ij}(\boldsymbol{\theta}, \mathbf{X})) / \sigma_{ij}^o$, where $M_{ij}(\boldsymbol{\theta}^*, \mathbf{X})$ are model outputs using input variables $\mathbf{X}$ and parameterization $\boldsymbol{\theta}^*$.

Determining jumps in the Metropolis Random Walk algorithm

The step length vector $\boldsymbol{\delta}$ in the Metropolis Random Walk algorithm is very important in order to obtain convergence of the Markov chain produced, i.e., the targeted posterior distribution of the parameters. In our implementation, the new candidate value θ_i' for parameter i was $\theta_i' = \theta_i^t + \delta_i$, where $\delta_i \sim N(0, a_i)$. If the elements in the step length vector $\boldsymbol{\delta}$ are too small, the random walk algorithm will not move far enough from the current point in parameter space $\boldsymbol{\theta}^t$ when proposing a new candidate parameter vector $\boldsymbol{\theta}'$, and consequently the acceptance rate will be too large, and vice versa. In our case, choosing a_i so that the acceptance rate was between 0.15 and 0.5 (in accordance with Roberts 1996) was attained by trial and error. Each element a_i of the vector $\boldsymbol{\delta}$ was chosen according to $a_i = c_i(\theta_i^{\max} - \theta_i^{\min})$, where c_i is a constant found by trial and error, and $(\theta_i^{\max} - \theta_i^{\min})$ is the width of prior pdf of parameter θ_i.

Determining convergence of the Markov chains

A central issue when using an iterative simulation method such as the Metropolis Random Walk algorithm is to determine when the chain has converged to the desired posterior distribution. One option, suggested by Gelman & Rubin (1992), is to generate multiple chains followed by calculating the scale reduction factor $\sqrt{\hat{R}}$, which is used to determine the length of the "burn-in" phase. The "burn-in" of the chain is the first part where the chain is influenced by the starting point until it reaches stationarity. We determined the "burn-in" phase to last until $\sqrt{\hat{R}} < 1.2$, following Gelman (1996): when $\sqrt{\hat{R}}$ nears 1 it means that the Markov chains are essentially overlapping. We randomly sampled two starting points from the prior distribution, and used the $\sqrt{\hat{R}}$ to determine when the two chains had converged to the desired posterior distribution.

Sensitivity analysis of SnowFrostIce

When working with models, sensitivity analysis (hereafter referred to as SA) is recommended as part of the process (Kokkonen et al. 2006). For the SA to be meaningful, the practitioner should decide beforehand on how to define the importance of the parameters, i.e., the type of question the SA is expected to answer (Saltelli et al. 2008). In our case, we would like to know which of the parameters can be fixed anywhere within their prior bounds without affecting model outputs, i.e., which parameters are not important. This is helpful in relation to model simplification.

In order to identify non-important parameters in the model, we carried out a screening exercise using the improved sensitivity indices from the Morris method, as described by Campolongo et al. (2007). This method is relatively simple to implement.

The Morris method proposes two sensitivity measures, the main purpose of which are to determine the model k parameters that can be considered to be (i) not important, (ii) linear and additive, or (iii) non-linear or involved in interactions with other parameters. For each of the parameters, two sensitivity measures are computed: $\boldsymbol{\mu}$, which evaluates the overall influence of the parameter on the model output (main effect or elementary effect [EE]), and $\boldsymbol{\sigma}$, which collectively evaluates all the higher order effects resulting from non-linearity and/or interactions with other parameters. The Morris method was originally used for parameters following uniform distributions in [0, 1]. If the k parameters follow other distributions, Campolongo et al. (1999) suggest that rather than sampling directly from these distributions, the sampling should be performed in the space of the quantiles of these k distributions (i.e., each parameter is discretized into p levels, and each quantile q_p varies in [0,1], producing a k-dimensional unit hypercube as the sampling space). The actual parameter values would subsequently be derived from their known distributions. In this SA of SnowFrostIce, we investigated the $k = 9$ parameters from the calibration (Table 1). The input space we used was the sub-space Ω comprised of the k-dimensional unit hypercube of the $p = 6$ equidistant quantiles in [0,1] from the prior distribution of the parameters $\pi(\boldsymbol{\theta})$. Outputs from SnowFrostIce are time series, and for this SA we needed a scalar value. Thus, for the simulation runs required in the SA, we used as output the log-transformed likelihood from Eqn. 20, i.e., $\log[f(\mathbf{D}|\boldsymbol{\theta})]$, with the likelihood being the probability of the observed data $\mathbf{D}$ given a certain model parameterization $\boldsymbol{\theta}$.

By randomly sampling parameter vectors $\boldsymbol{\theta}$ from Ω, and calculating EE (for details, see Campolongo et al. 2007) for each of the nine parameters, we obtained a sample from the distribution for each EE, termed $\mathrm{EE}_i \sim F_i(\mu_i, \sigma_i)$. The sensitivity measures μ_i and σ_i proposed by Morris are the mean and standard deviation of F_i, respectively. To estimate μ_i and σ_i, the sampling strategy proposed by Morris is to create r trajectories in parameter space Ω. Each of these r trajectories contains $(k + 1)$ points, and results in k elementary effects (i.e., estimates of one EE per parameter), leading to a total of $r(k + 1)$ sample points corresponding to the number of model runs required for the complete SA. A very nice stepwise presentation of this method is presented in Saltelli et al. (2008).

A high σ_i value for parameter θ_i implies that the corresponding EE_i value for θ_i at one point in Ω is considerably

different from another EE_j value ($I \neq j$) for the same parameter θ_i located somewhere else in Ω, i.e. that this particular EE value is influenced by the values of the other parameters or nonlinearities. A low value for σ_i suggests that the EE_i value associated with θ_i is independent from the values of the other parameters, and thus it is not involved in interactions or nonlinearities.

To avoid type II errors of failing to identify important parameters, Campolongo et al. (2007) suggest replacing $\boldsymbol{\mu}$ by $\boldsymbol{\mu}^*$, an estimate of the mean of the distribution of the absolute values of the elementary effects G_i, i.e., $|EE_i| \sim G_i(\mu^*_i, \sigma_i)$. To properly characterise non-influential parameters, one must therefore simultaneously consider the vectors $\boldsymbol{\mu}^*$ and $\boldsymbol{\sigma}$ (see Fig. 1 for SA results for the Kise site).

When conducting the SA, we tried the same approach for all locations. First, we sampled trajectories from the prior distribution and calculated $\boldsymbol{\mu}^*$ and $\boldsymbol{\sigma}$. Then, we sampled from the posterior distribution and calculated $\boldsymbol{\mu}^*$ and $\boldsymbol{\sigma}$. Sampling trajectories from the prior distribution gave very similar results for all sites (as did those for Kise; Fig. 1). When we sampled from the correlated posterior distribution, the results in $\boldsymbol{\mu}^*$ and $\boldsymbol{\sigma}$ were different when comparing sites. For all but one site the same non-important parameters were identified, but highly correlated parameters influenced the results. For example, at Kise, the parameter T_{rs} was wrongly recognized as being non-important. This illustrates that the Morris method can produce different results depending on whether the parameters are correlated or not. We did not find examples in the literature of how to handle correlated parameters when using the Morris method. T_{rs} is an important parameter, as was clearly shown when sampling trajectories from the prior distribution. Based on this observation, we decided to use the assumed uncorrelated prior distribution when sampling trajectories for the screening exercise.

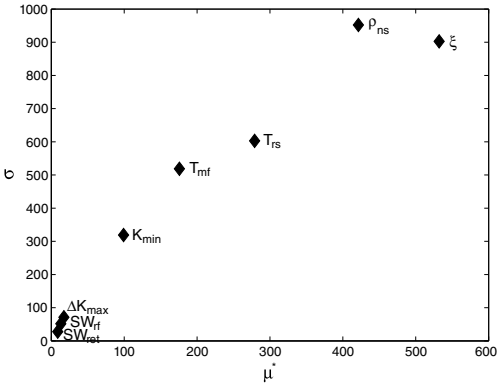


Fig. 1 Sensitivity analysis results for the location Kise showing $\boldsymbol{\mu}^*$ versus $\boldsymbol{\sigma}$, based on $r = 100$ trajectories. Low values for both $\boldsymbol{\mu}^*$ and $\boldsymbol{\sigma}$ identify SW_{ret}, SW_{rf} and ΔK_{max} as the least important parameters.

Model validation and predictive uncertainty

The data sets for each location were divided in two: one part was used for calibration and the other was used for validation (Table 2). To evaluate the predictive uncertainty of the model after calibration, we sampled 20 parameter sets from the posterior distribution, and calculated the subsequent mean and standard deviations of the model outputs.

Results

Results from the Bayesian calibration

The main result of the Bayesian calibration procedure is the estimated joint posterior distribution of the model parameters. This correlated multidimensional joint distribution is difficult to visualize, so we present the marginal posterior distribution for single parameters.

We determined the success of the calibration by evaluating the estimated marginal posterior distributions. If they are narrower than their corresponding prior distribution, this indicates that the parameter uncertainty has been reduced. The calibration at each location used two chains of length 300 000, and a unique step length vector for that location.

The part of the Markov chains succeeding the burn-in point, which we determined as the point from where $\sqrt{\hat{R}}$ remains below 1.2, comprises the marginal posterior distribution of the parameter (Gelman 1996). The right column of Fig. 2 shows plots of $\sqrt{\hat{R}}$ for the parameters ξ, ρ_{ns} and T_{rs}, and the centre column shows the estimated marginal posterior distribution for the same parameters. Panels in the left column in Fig. 2 show trace plots of the Markov chains for parameters ξ, ρ_{ns} and T_{rs} calibrated at the Kise site. These trace plots are used to verify that the two chains for each parameter stabilize around the same value, and that the posterior distribution is properly explored.

In order to visualize the marginal posterior distributions for all locations simultaneously, we fitted continuous distributions to the samples from the posterior generated by the MCMC. They are shown, together with the prior distributions, in Fig. 3. The marginal posterior distributions are either multimodal, skewed or both. It was therefore informative to present both the maximum posterior estimate and the median value of $\boldsymbol{\theta}$ ($\boldsymbol{\theta}_{MAP}$ and $\tilde{\boldsymbol{\theta}}$, respectively) from the marginal posterior

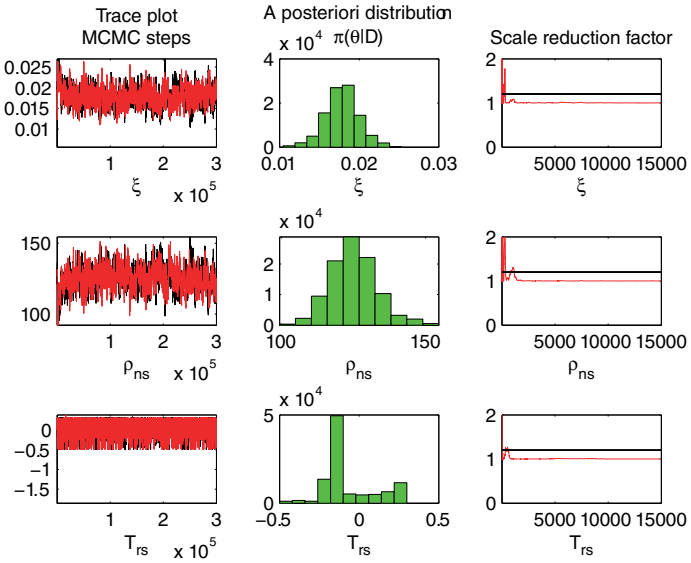


Fig. 2 Panels in top row show results for parameter ξ; centre row shows results for parameter ρ_{ns}; bottom row shows results for parameter T_{rs}. Panels in the left column show trace plots of the two parallel chains (red, chain 1; black, chain 2). Panels in the centre column show the marginal posterior distribution of the parameter $\pi(\theta|\mathbf{D})$. Panels in the right column show the scale reduction factor $\sqrt{\hat{R}}$, calculated at every 20th iteration.

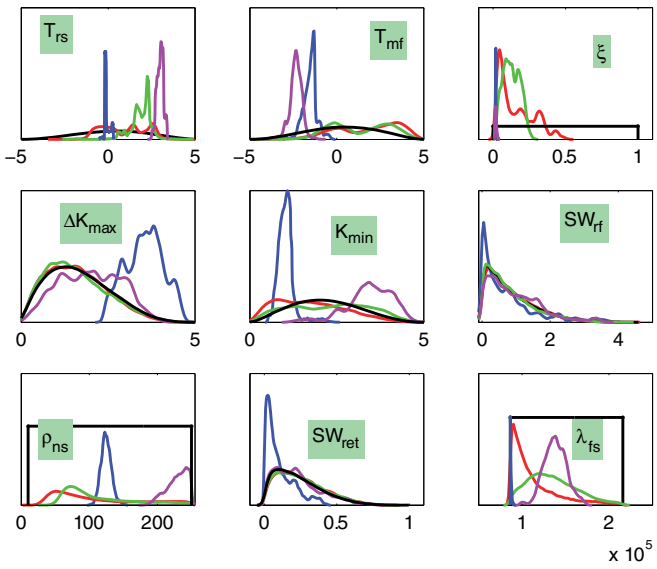


Fig. 3 Continuous density function estimations of the prior distributions (black) and the marginal posterior distributions for all locations: Kise (blue), Kvithamar (red), Vågønes (green) and Holt (magenta).

distributions as summary statistics (see Table 7), complemented by plots of the marginal posterior distributions in Fig. 3 showing posterior parameter uncertainty. The parameter vector $\boldsymbol{\theta}_{MAP}$ represents the single best parameter vector at the different locations. For most of the parameters, when comparing the marginal posterior distributions in Fig. 3 with their respective prior distribution (black lines), it is clear that the calibration process reduced the prior parameter uncertainty. However, for the parameters related to liquid water in the snow cover, SW_{rf} and SW_{ret}, we can see that measurements on snow depth alone did not provide enough information to

Table 7 Parameter values for SnowFrostIce that gave the highest posterior density $\boldsymbol{\theta}_{MAP}$, and the median values $\tilde{\boldsymbol{\theta}}$ for the sites Kise, Kvithamar, Vågønes and Holt.

Parameter	Kise		Kvithamar		Vågønes		Holt	
θ_i	$\theta_{MAP(i)}$	$\tilde{\theta}_{(i)}$	$\theta_{MAP(i)}$	$\tilde{\theta}_{(i)}$	$\theta_{MAP(i)}$	$\tilde{\theta}_{(i)}$	$\theta_{MAP(i)}$	$\tilde{\theta}_{(i)}$
T_{rs}	–0.1	–0.1	–0.6	1	2.3	2	3.1	3
T_{mf}	–1.4	–1.5	3.1	2.1	0.7	1.3	–3	–2.3
ξ	0.02	0.02	0.025	0.12	0.15	0.13	0.01	0.02
ΔK_{max}	4.5	3.6	0.79	1.5	1.8	1.5	0.5	2
K_{min}	1.1	1	0.43	1.6	0.2	2.3	2.6	3.5
SW_{rf}	0.002	0.48	0.87	0.68	2.61	0.63	3.65	0.78
ρ_{ns}	128	124	216	89	84	95	250	231
SW_{ret}	0.32	0.07	0.21	0.22	0.18	0.22	0.35	0.2
λ_{fs} ($\times 10^4$)	8.6	8.8	12.6	10.3	17.6	13.1	13.7	13.8

depart from our prior estimates (for the Kise site, they are more peaked). For the precipitation threshold temperature, T_{rs}, the parameter uncertainty was least reduced at Kvithamar compared with the other locations. For T_{mf}, the parameter uncertainty was reduced more at Kise and Holt than at Kvithamar and Vågønes. For Kvithamar, the median value of T_{mf} (see Table 7) was larger than the median value of T_{rs} (this is shown in Fig. 4b, where the green line is located slightly above the red line). The uncertainty in ΔK_{max} and K_{min} was reduced for Kise and Holt, but for Kvithamar and Vågønes there was not much improvement. The parameter uncertainty was reduced for the remaining ξ, ρ_{ns} and λ_{fs}.

Results from the sensitivity analysis

At each of the locations used in the calibration, we randomly generated $r = 100$ different trajectories for the computation of EE, i.e., $r(k + 1) = 1000$ parameter vectors were sampled from Ω, and thus 1000 model runs were used for the SA. The results were very similar for each location. Figure 1 shows the sensitivity indices μ^*_i and σ_i for each parameter for the Kise site. We find the parameters SW_{ret}, SW_{rf} and ΔK_{max} in the lower left-hand corner, and the remaining parameters are almost linearly spread. Inspection of histograms of the sampled parameter values suggests that the ranges of the prior intervals were adequately explored.

The parameter λ_{fs} was excluded from the SA because S_{depth} affects F_{depth}, and not vice versa.

Validation of the model

The SnowFrostIce model was validated at all locations used in the calibration. For each of the locations we sampled 20 parameter vectors from the posterior distribution, and calculated the mean and standard deviation of the model output. Variation in model output is shown as the mean ± one standard deviation (Fig. 5). If the median value of T_{rs} is close to that of T_{mf} they appear as one line in the sub-figures. See Table 7 for these parameter values. The validation at the Kise site shows little variation in model output. At this site, S_{depth} is overestimated for the winter of 1997/98. This is as expected when considering that $T_{air} < T_{rs}$ for most of the precipitation events (see Fig. 4a). Frost depth at Kise during 1997/98 is initiated earlier than observed, in addition to being slightly underestimated. F_{depth} during the 1998/99 winter is overestimated: frost rates that were too high initially caused F_{depth} to be shifted downwards compared with observations. The validation for Kvithamar (Fig. 5b) shows more variability in model output compared with Kise, especially towards the end of springtime for S_{depth}. The data points here are captured within this variation. At Vågønes (Fig. 5c), model performance for S_{depth} is quite good, but F_{depth} is overestimated (more severely during 2001/02 than 2002/03). At Holt (Fig. 5d), the S_{depth} is overestimated during 2005/06 (as with Kise in 1997/98) because $T_{air} < T_{rs}$ for most of the precipitation events of that winter. Note the events between January and May 2006 with $P > 20$ mm (Fig. 4d), where precipitation is simulated as snow. F_{depth} looks reasonably accurate, but a complete thaw is predicted too early for both validation years. Variation in model output is in general higher for Kvithamar and Holt than for Vågønes and Kise. Figure 6 shows all output (snow cover, soil frost and surface ice) for Holt (1997/98) and Karasjok (1998/99). We had no data to calibrate SnowFrostIce for Karasjok. For Karasjok, we sampled the parameter values from $\pi(\boldsymbol{\theta}|\mathbf{D})$ obtained for Kise, as both locations have an interior climate. Holt and Karasjok were the only locations where ice observations were available.

Discussion and conclusions

In this paper we present a new model for the simulation of snow depth, soil frost depth and depth of surface

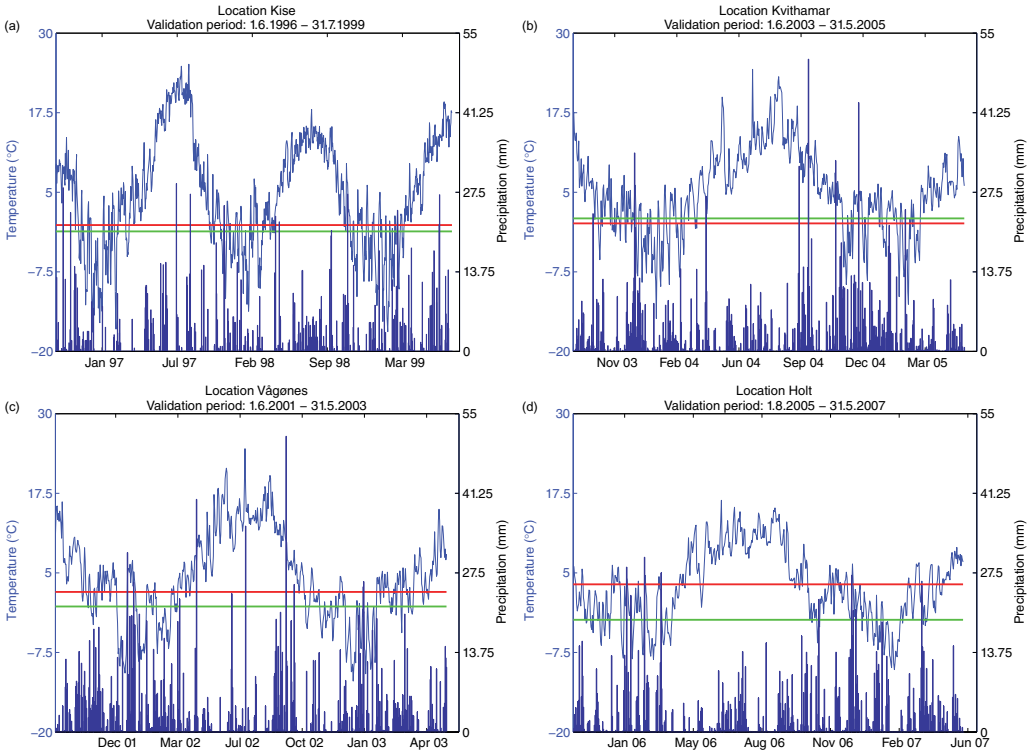


Fig. 4 Climate during validation period for locations (a) Kise, (b) Kvithamar, (c) Vågønes and (d) Holt. Solid lines show the daily mean air temperature T_{air} (blue) and bars show daily precipitation; median values from the posterior distribution of the threshold temperatures for precipitation T_{rs} (red) and snowmelt/refreezing T_{mf} (green). See Table 7 for parameter values.

ice cover. We calibrated the model by means of well-documented Bayesian methods, and conducted a qualitative sensitivity analysis. As far as we know this practice is still relatively new for this kind of model. The results presented here, both regarding assumptions about prior pdfs and the resulting posterior pdfs, and the simple yet very effective method of sensitivity analysis, are useful for the modelling community.

A study comparing four models simulating soil frost (Kennedy & Sharratt 1998)—the two finite difference models SHAW and SOIL, and two energy balance models—concluded that the simpler energy balance models generally overestimate frost depth. However, one weakness of the models (investigated by Kennedy & Sharratt) that also simulate snow cover is the estimation of snow depth. Snow cover has a strong influence on the estimation of soil frost depth, e.g., through snow depth and snow density, both affecting the thermal conductivity of the snow cover. Therefore, accurate simulation of snow cover is important for the simulation of soil frost depth.

Our new model SnowFrostIce for simulating the effects of winter climate on the soil surface is designed to be included in a grassland model. This restricts SnowFrostIce with regards to the number of parameters included. We calibrated SnowFrostIce against independent data from four locations in Norway, capturing climatic variation from south to north and from coastal to inland areas. We also identified the key parameters by conducting a sensitivity analysis.

It is important to bear in mind that SnowFrostIce represents simplifications of real-world processes, which are described at various levels of complexity. Some of the parameters used have a physical interpretation, but they are seldom measured, and quantitative data are scarce in the literature. This means that the parameters, and thereby the model outputs, are subject to uncertainty.

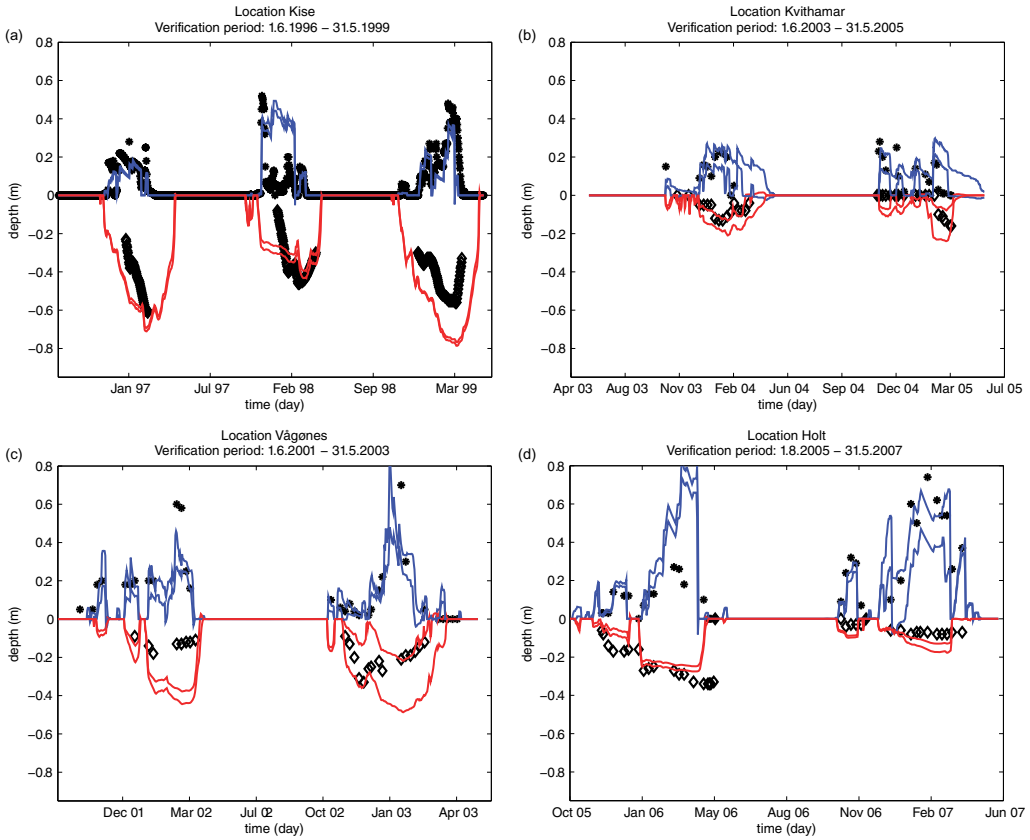


Fig. 5 Validation of the SnowFrostIce model describing the variation between model output and observed values on depths of snow and soil frost at (a) Kise, (b) Kvithamar, (c) Vågønes and (d) Holt. Solid lines (mean ± SD) show S_{depth} (blue) and F_{depth} (red); observed snow cover depth (*); observed lower frost boundary (◇).

The Bayesian method we used aims to quantify and reduce these uncertainties, rather than maximizing the model fit. When selecting an optimal parameter set for a simulation run for a specific location, we chose the parameter values that maximized the posterior distribution $\boldsymbol{\theta}_{MAP}$ (Table 7). A consequence of this procedure was that these specific parameter values must be interpreted accordingly (i.e. reducing model uncertainty), rather than given a clear-cut physical interpretation.

When carrying out the Bayesian calibration, it was difficult to obtain convergence of the Markov chains for the parameters relating to liquid water in the snow cover (SW_{ret} and SW_{rf}). This may imply that the calibration data were not sufficient for improving the prior knowledge related to these parameters.

The estimated posterior distribution is different for each location. We expected some regional differences for the melting parameters K_{min} and ΔK_{max} as a result of regional differences in radiation, altitude (m a.s.l.) and ocean vicinity, for example, but not for the threshold temperatures for precipitation T_{rs} and snowmelt/refreezing T_{mf}, or for the density of new snow ρ_{ns}. This might indicate that the model needs geographical adjustments and a functional description of ρ_{ns}. The differences in the results for the thermal conductivity of frozen soil λ_{fs} were expected, as the soil types are different for each of the locations.

A reason for the erroneous estimation of S_{depth} could be that the calibrated value of T_{rs} is wrong, leading to observed rain being simulated as snow, or vice versa. In addition, by using daily mean air temperatures, the model

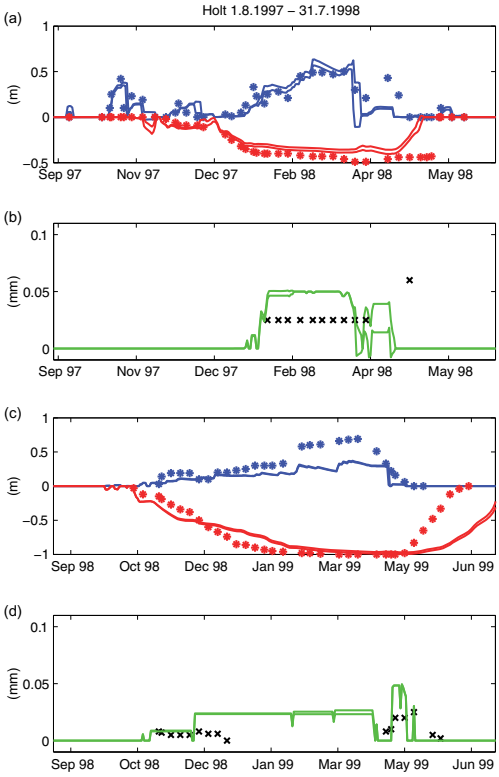


Fig. 6 Simulation results for (a, b) Holt and (c, d) Karasjok. In (a) and (c), solid lines show simulated values (mean ± SD) and (*) indicate observed values; blue represents snow cover, and red represents the lower frost boundary. In (b) and (d), solid green lines represent simulated (mean ± SD) ice cover depth, and (×) indicate observed values for ice cover depth.

might associate incorrect air temperatures with precipitation events. For instance, the observed air temperature could be below 0 °C for most of the day, followed by above 0 °C at the end of the day, resulting in a mean daily temperature below T_{rs}. If precipitation had been observed as rain by the end of the day, it would still have been simulated as snow. The overestimation of S_{depth} might result from important processes being omitted, e.g., the heat content of rain is not incorporated in the model, so this kind of additional snowmelt is not included. A third reason for the erroneous estimation of S_{depth} might be the redistribution of snow by wind, a factor not taken into account in the model.

The number of available observations for the calibration is important. Using data from two and three years is not sufficient to capture the interannual variation in snow cover and soil frost. The limited number of observations on both snow cover and soil frost at the same location has an effect on the results of the calibration. In a preliminary study, the snow module of SnowFrostIce was calibrated for the Kise location using two, four, six, eight and finally 10 years of snow depth observations (Roer et al. unpubl. ms.). Including more data resulted in a narrower posterior distribution, but convergence was also increasingly harder to obtain. Including more observations also resulted in a shift in the location of the posterior parameter distribution. This showed that the interannual variation in winter weather will affect the results of the calibration. As long as more data are included, the results are likely to keep varying until the whole spectrum of weather conditions is included. Ideally, we should have had observations comprising a full climate period (30 years) to capture the variation within a normal period. In the study comparing 33 snowpack models by Rutter et al. (2009) only two years of observations were available. In the present study, the data set was split in two in order to conduct the validation, which would otherwise have had to be postponed until more observations became available.

The parameters related to snowmelt (T_{mf}, K_{min} and ΔK_{max}) are less uncertain for Kise than for the other locations (see Fig. 3). This contributes to less uncertainty in the snow depth simulation at Kise compared with the other locations. At Kvithamar, in addition to the uncertainty of T_{mf}, this parameter also has a high numerical value compared with the other locations (Table 7). This leads to more uncertainty in the melting period at Kvithamar, and also to a delayed onset of snowmelt in the simulations compared with, for example, Kise. The results from the sensitivity analysis showed that T_{mf} and K_{min} were the most important parameters related to snowmelt. It is therefore reasonable to attribute the uncertainty and delay in snowmelt mainly to the uncertainty of the parameters T_{mf} and K_{min}.

In this study we used the likelihood of a sampled parameter set, given the data (see Eqn. 20), as a scalar output when calculating the sensitivity indices $\boldsymbol{\mu}^*$ and $\boldsymbol{\sigma}$. If, on the other hand, we were to use daily simulated snow depth values as the scalar output in the SA, for example, we would have to calculate one pair of $\boldsymbol{\mu}^*$ and $\boldsymbol{\sigma}$ for each of these S_{depth} values. This would provide an answer to the question of which parameters were most important on which day during the whole simulation period. However, performing two SA—where the first SA uses depth of snow cover on a specific day during midwinter, and the second SA uses depth of snow cover on a specific day towards the end of winter—might give further indications

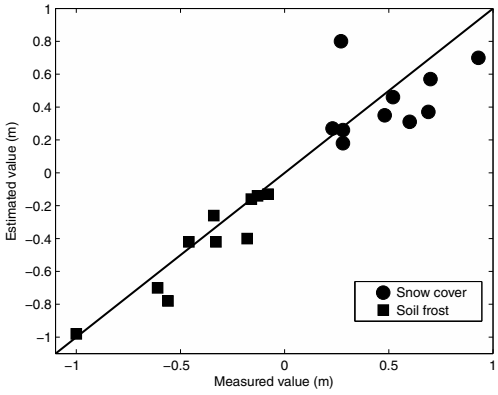


Fig. 7 Collective scatterplot of measured and estimated maximum values of the depths of snow cover S_{depth} and soil frost F_{depth}. Each point represents one independent set of data from the validation periods: three points correspond to Kise, two points each correspond to Kvithamar, Vågønes and Holt, and one point corresponds to Karasjok.

of which parameters are most important regarding snowmelt in cold and mild periods, respectively.

The purpose of our SA was to identify key parameters in the model. Here, we used the Morris screening method to identify the non-important parameters. In Fig. 1, the parameters SW_{rf}, SW_{ret} and ΔK_{max} are recognized as being less important (low values for $\boldsymbol{\mu}^*$ and $\boldsymbol{\sigma}$). Following the way in which we defined parameter importance in our SA, the SA results suggest that varying these parameters within their prior bounds would not markedly affect the model output. We can also find support for this conclusion in the calibration results. Figure 3 shows that the posterior distribution for the three non-important parameters has not changed much compared with the prior distribution, implying that no new information is added through the data.

The ability of SnowFrostIce to simulate the maximum depths of snow cover and soil frost is presented in Fig. 7. The points lie close to the 1:1 line, indicating satisfactory model performance. The maximum depths of snow cover and soil frost are good indices to show the trends of the snow cover and soil freezing in each winter, and they are both appropriately estimated by the model.

The approach to calculating soil frost, by balancing energy, is similar to that proposed by Benoit & Mostaghimi (1985). Although we made some critical assumptions (e.g., estimation of the soil surface temperature, a constant thermal conductivity of frozen soil and a constant thermal conductivity of snow), we have shown that when tested on independent data sets (see Fig. 7), the ability of SnowFrostIce to estimate the maximum lower frost boundary is also quite good.

The simulation of ice cover at Holt and Karasjok were based on the assumptions outlined in the sections "Puddle formation and infiltration of meltwater" and "Formation of ice layer". The results shown in Fig. 6 indicate that our simple approach is a sound starting point for further development of the ice layer module.

We conclude that our simple yet effective method for modelling depths of snow cover, lower frost boundary and soil surface ice provides reasonable results, making it suitable for incorporation into more complex models.

Continued work

In order to simulate damage to plants as a result of ice encasement, for example, there is a need for a better description of local field topography, such as quantifying the part of the study area that can potentially be covered by surface puddles. This will be of help when simulating the number of plants dying because of ice-related stresses. These refinements should be followed by further model validation.

The results from the calibration and SA indicate scope for model improvement. A modification motivated by the calibration results is a functional description of ρ_{ns}. In addition, the results from the SA suggest lumping together (or disregarding) the processes related to liquid water within the snow cover, and replacing the sinusoidal snowmelt function by a constant melt rate.

Acknowledgements

Comments from Dr Mats Höglind were much appreciated. This study was funded by the Norwegian Research Council as part of the programme WINSUR. The authors are grateful to the two anonymous reviewers for their insightful comments on an earlier version of this manuscript.

References

Baker J. & Spaans E.J.A. 1997. Mechanics of meltwater movement above and within frozen soil. In I. Iskandar et al. (eds.): *Proceedings of the International Symposium on Physics, Chemistry, and Ecology of Seasonally Frozen Soils, Fairbanks, Alaska, June 10–12, 1997. CRRL Special Report 97-10*. Pp. 31–36. Hanover, NH: US Army Cold Regions Research and Engineering Laboratory.

Bélanger G., Rochette P., Gastonguay Y., Bootsma A., Mongrain D. & Ryan D. 2002. Climate change and winter survival of perennial forage crops in eastern Canada. *Agronomy Journal 94*, 1120–1130.

Beldring S., Engen-Skaugen T., Forland E. & Roald L. 2008. Climate change impacts on hydrological processes in

Norway based on two methods for transferring regional climate model results to meteorological station sites. *Tellus Series A 60*, 439–450.

Benoit G. & Mostaghimi S. 1985. Modeling soil frost depth under three tillage systems. *Transactions of the ASAE 28*, 1499–1505.

Campolongo F., Cariboni J. & Saltelli A. 2007. An effective screening design for sensitivity analysis of large models. *Environmental Modelling and Software 22*, 1509–1518.

Campolongo F., Tarantola S. & Saltelli A. 1999. Tackling quantitatively large dimensionality problems. *Computer Physics Communications 117*, 75–85.

DeGaetano A., Cameron M. & Wilks D. 2001. Physical simulation of maximum seasonal soil freezing depth in the United States using routine weather observations. *Journal of Applied Meteorology 40*, 546–555.

Engeset R., Sorteberg H. & Udnæs H. 2000. *NOSIT—utvikling av NVEs operasjonelle snøinformasjonstjeneste. (NOSIT—development of the operational snow information services of NVE.) Dokument 1*. Olso: Norwegian Water Resources and Energy Directorate.

Flerchinger G. & Saxton K. 1989. Simultaneous heat and water model of a freezing snow–residue–soil system. I. Theory and development. *Transactions of the ASAE 32*, 565–571.

Gelman A. 1996. Interference and monitoring convergence. In W.R. Gilks et al. (eds.): *Markov chain Monte Carlo in practice*. Pp. 131–144. Suffolk: Chapman & Hall.

Gelman A. & Rubin D. 1992. Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science 7*, 457–511.

Gray J. & Morland L. 1995. The compaction of polar snow packs. *Cold Regions Science and Technology 23*, 109–119.

Gudleifsson B. & Larsen A. 1993. *Advances in plant cold hardiness*. Boca Raton, FL: CRC Press.

Hayashi M., van Der Kamp G. & Schmidt R. 2003. Focused infiltration of snowmelt water in partially frozen soil under small depressions. *Journal of Hydrology 270*, 214–229.

Höglind M., Schapendonk A. & van Oijen M. 2001. Timothy growth in Scandinavia: combining quantitative information and simulation modelling. *New Phytologist 151*, 355–367.

Iwata Y., Hayashi M. & Hirota T. 2008. Comparison of snowmelt infiltration under different soil-freezing conditions influenced by snow cover. *Vadose Zone Journal 7*, 79–86.

Jansson P. & Karlberg L. 2001. *Coupled heat and mass transfer model for soil–plant–atmosphere systems*. Stockholm: Department of Civil and Environmental Engineering, Royal Institute of Technology.

Jordan R. 1991. *A one-dimensional temperature model for a snow cover. Technical documentation for SNTHERM.89. CRRL Special Report 91-16*. Hanover, NH: US Army Cold Regions Research and Engineering Laboratory.

Judson A. & Doesken N. 2000. Density of freshly fallen snow in the central Rocky Mountains. *Bulletin of the American Meteorological Society 81*, 1577–1587.

Kennedy I. & Sharratt B. 1998. Model comparisons to simulate soil frost depth. *Soil Science 163*, 636–645.

Kokkonen T., Koivusalo H., Jakeman T. & Norton J. 2006. Construction of a degree–day snow model in the light of the ten iterative steps in model development. In A. Voinov et al. (eds.): *Proceedings of the iEMSs Third Biennial Meeting: Summit on Environmental Modelling and Software. Environmental Modelling and Software Society, Burlington, USA, July 2006*. CD-ROM. Manno, Switzerland: International Environmental Modelling and Software Society.

Larsen A. 1994. Breeding winter hardy grasses. *Euphytica 77*, 231–237.

Melloh R. 1999. *A synopsis and comparison of selected snowmelt algorithms. CRRL Report 99-8*. Hanover, NH: US Army Cold Regions Research and Engineering Laboratory.

Roberts G.O. 1996. Markov chain concepts related to sampling algorithms. In W.R. Gilks et al. (eds.): *Markov chain Monte Carlo in practice*. Pp. 45–57. Suffolk: Chapman & Hall.

Rutter N., Essery R., Pomeroy J., Altimir N., Andreadis K., Baker I., Bartlett P., Boone A., Deng H., Douville H., Dutra E., Elder K., Ellis C., Feng X., Gelfan A., Goodbody A., Gusev Y., Gustafsson D., Hellström R., Hirabayashi Y., Hirota T., Jonas T., Koren V., Kuragina A., Lettenmaier D., Li W.-P., Luce C., Martin E., Nasonova O., Pumpanen J., Pyles R.D., Samuelsson P., Sandells M., Schädler G., Shmakin A., Smirnova T.G., Stähli M., Stöckli R., Strasser U., Su H., Suzuki K., Takata K., Tanaka K., Thompson E., Vesala T., Viterbo P., Wiltshire A., Xia K., Xue Y. & Yamazaki T. 2009. Evaluation of forest snow process models (SnowMIP2). *Journal of Geophysical Research—Atmospheres 114*, D06111, doi: 10.1029/2008JD011063.

Saltelli A., Ratto M., Andres T., Campolongo F., Cariboni J., Gatelli D., Saisana M. & Tarantola S. 2008. *Global Sensitivity Analysis. The primer*. West Sussex: Wiley & Sons.

Sivia D. 2006. *Data analysis: A Bayesian tutorial*. 2nd edn. Oxford: Oxford University Press.

Stähli M., Bayard D., Wydler H. & Fluhler H. 2004. Snowmelt infiltration into alpine soils visualized by dye tracer technique. *Arctic, Antarctic, and Alpine Research 36*, 128–135.

Thorsen S.M. & Haugen L.E. 2007. *Development of the SnowFrost model for the simulation of snow fall and soil frost. Bioforsk FOKUS 2(9)*. Ås, Norway: Bioforsk.

van Oijen M., Höglind M., Hanslin H. & Caldwell N. 2005. Process-based modelling of timothy regrowth. *Agronomy Journal 97*, 1295–1303.

van Oijen M., Rougier J. & Smith R. 2005. Bayesian calibration of process-based forest models: bridging the gap between models and data. *Tree Physiology 25*, 915–927.

Vehvilainen B. 1992. *Snow cover models in operational watershed forecasting*. PhD thesis, Helsinki University of Technology.

Nomenclature

F_{depth}	simulated depth of lower frost boundary (m)
I_{depth}	simulated thickness of surface ice cover (m)
K	degree–day temperature index for snowmelt (mm °C^{-1} day^{-1})
K_{min}	minimum value of K (mm °C^{-1})
K_{max}	maximum value of K (mm °C^{-1})
L_f	latent heat of fusion (J kg^{-1})
M	snow melt rate (mm day^{-1})
M_{rf}	refreezing rate (mm day^{-1})
P	precipitation rate (mm day^{-1})
P_r	simulated daily precipitation rate as rain (mm day^{-1})
P_s	simulated daily precipitation rate as snow (mm day^{-1})
Q_E	heat flux density from freezing of soil water (J m^{-2} day^{-1})
Q_{fs}	heat flux density through frozen soil (J m^{-2} day^{-1})
Q_{snow}	heat flux density through snow cover (J m^{-2} d^{-1})
S_{dry}	water constituent of snow cover in solid state (snow and ice) (mm)
S_{wet}	liquid water constituent of snow cover (mm)
SWE	snow water equivalent (mm)
SW_{ret}	retention capacity of snow cover (mm mm^{-1})
SW_{rf}	degree–day temperature index for refreezing of liquid water within snow cover (mm °C^{-1} day^{-1})
S_{depth}	depth of simulated snow cover (m)
T_{air}	daily mean air temperature at 2 m height (°C)
T_{surf}	simulated temperature in void between soil surface and snow cover (°C)
T_{rs}	daily mean air temperature below which precipitation is simulated as snow (°C)
T_{mf}	daily mean air temperature below which water within snow cover refreezes (°C)
T^*	daily mean air temperature below which soil water freezes (°C)
x_w	volumetric content of available soil water (m^3 m^{-3})
z	soil depth (m)
γ	empirical parameter (m^{-1})
λ_{fs}	thermal conductivity of frozen soil (J m^{-1} °C^{-1} day^{-1})
λ_i	thermal conductivity of surface ice cover (J m^{-1} °C^{-1} day^{-1})
λ_s	thermal conductivity of snow cover (J m^{-1} °C^{-1} day^{-1})
ρ_{ns}	density of falling new snow (kg m^{-3})
ρ_w	density of water at 0°C (kg m^{-3})
ρ_s	density of snow cover (kg m^{-3})
ξ	snow cover compaction parameter (mm mm^{-1} d^{-1})

III

Modelling spore maturation in a Bayesian framework with application to spore release of *Venturia inaequalis*

Anne-Grete Roer*[ab], Håvard Eikemo[a], Arne Stensvand[a], Trygve Almøy[b], Piet Creemers[c], Trond Rafoss[a]

[a] Norwegian Institute for Agricultural and Environmental Research, Plant Health and Plant Protection Division. Høgskoleveien 7, N-1432 Ås, Norway

[b] Norwegian University of Life Science, Department of Chemistry, Biotechnology and Food Science, N-1432 Ås, Norway

[c] Proefcentrum Fruitteelt – Applied Scientific Research, Department of Mycology, B-3800 Sint-Truiden, Belgium

*Corresponding author. Tel.: +47 900 26 519; fax: +47 649 46 110.

Email address: anne-grete.roer@bioforsk.no, (Anne-Grete Roer)

Abstract

A pest forecasting model of ascospore maturity of *Venturia inaequalis* developed in New Hampshire, US, were further developed in order to predict spore release more accurately. Five alternative models and three tuning alternatives for model optimization were proposed and the different model versions compared. The data were split in one set for model development and one set for model assessment ten times, and all calculations were done for each split individually, to ensure not fitting the model only to the specific data but more generally to novel data from the same population. The split of the data turned out to be a crucial factor for the conclusions made. As mean value over the ten splits of the data, the best model were 9.2 % improved compared to the New Hampshire model according to root mean square error of prediction (RMSEP), and 19.0 % improved when only looking at the most important spore release periods between 5 and 95 % matured ascospores only. Because of a high variance between the different splits of the data, the improvements were only significant ($p<0.05$) for the 5 to 95 % interval. Additionally the deviance information criterion (DIC) was used for

model comparison, giving a more complex model as the overall best. The criteria used for model assessment were thereby crucial in order to which model selected as the best, but the two models were not significantly different. The Bayesian statistical approach was used to calibrate the models. By regarding parameters as random variables following some probability distribution, not as fixed values, calculations of parameter uncertainties were automatically included. Model robustness was visualized by showing the effect of adding data on the resulting parameter distribution and thereby its uncertainty.

Key words: apple scab, model robustness, model selection, impact of the data, random walk Metropolis

INTRODUCTION

Computer models constitute a key component of modern forecasting systems for plant pests and diseases. Generally, they are systemized biological knowledge, which usually takes biological observations and weather prognoses as inputs. The main purpose of these models is to optimize the precision of disease/pest management applications for optimal crop protection, and to avoid unnecessary effort when there is no risk. The models are included in warning systems that alert farmers about when and where the disease/pest may attack the crop.

Model predictions will never be completely perfect. Uncertainty is related to how close the true real world quantity will be to the model simulation, but still uncertainties are usually not addressed in disease/pest forecasting. According to Goldstein and Rougier (2006), there are three sources of uncertainty related to models; (i) the model itself is an imperfect representation of the underlying system, (ii) it generally contains unobserved or uncertain quantities (parameters), and (iii) collected data from the system used to calibrate (where calibration is learning about the parameters using data from the system) the model are imperfectly measured. Included in the third source, there will be an uncertainty related to how well the specific collected data represents the population. Bøvelstad et al. (2007), pointed out that splitting the data into a training and a test set in several different splits is important for the model to fit well to novel data from the same system, and not only to the specific data at hand.

Model parameters mainly have a physical meaning also outside the model, and prior information about them is often known. Although it is a natural desire to incorporate all available information in decision making (Wolfson et al. 1996), classical statistical methods consider the collected data to be the only available source of information. Also, in classical statistical methods, uncertainties may only be established through large samples (asymptotic) arguments. Therefore, a combination of prior knowledge and substantial uncertainties makes it natural to adapt a Bayesian statistical approach (Kennedy and O'Hagan 2001). In Bayesian statistics, probability theory is used to describe uncertainty, and parameters are regarded as random variables (not as fixed values as in classical frequentistic statistics) that follow some probability distribution.

The aim of Bayesian calibration is to reduce the prior uncertainty by making use of the collected data. Initially, our parameter knowledge before collecting data is described by a prior probability distribution. The prior knowledge is then updated by new incorporated information, through the likelihood distribution of the collected data. The resulting combination, describing our new parameter uncertainty, is called the posterior parameter distribution.

As a case study we used ascospore maturation in the ascomycete *Venturua inaequalis*, the cause of apple scab, a serious disease on apple worldwide. The primary overwintering site for the fungus is leaf litter on the ground, and in spring and early summer ascospores are released from pseudothecia during periods of rain. The fungus targets its maturation and first release of ascospores to match the time when first susceptible tissue is present on the apple trees. Numerous studies have been undertaken to examine the seasonal distribution of ascospore release in *V. inaequalis* in various apple producing countries in the world (Childs 1917, Stover and Johnson 1924, Frey and Keitt 1925, Schneiderhan 1925, Keitt and Jones 1926, Weber 1934-35, Wiesman 1935, Fjelddalen 1948, Weber and Jørgensen 1953, Gjærum 1954, Hårdh 1955, Szkolnik 1969, Moller et al. 1971, Szkolnik 1974, Brook 1976, Gadoury and MacHardy 1982, Ylämäki 1989, Norin 1989, Rossi et al. 1999). Based on temperature sums or temperature sums in combination with moisture, several models have been developed for estimation of ascospore maturation in *V. inaequalis* (Gadoury and MacHardy 1982, James and Sutton 1982, Lagarde 1988, Massie and Szkolnik 1974, Stensvand et al. 2005). Both the apple tree and the fungus have a development rate dependent on temperature. However, ascosporic development is also dependent on available moisture, and protracted periods of dry weather slow down or stop the maturation process (Keitt and Jones 1926, Wilson 1928, James and

Sutton 1982a, James and Sutton 1982b, O'Leary and Sutton 1986, Schwabe et al. 1989, Stensvand 2005).

In this paper, we have further developed an existing forecasting model for the specific plant disease, and the Bayesian statistical framework was used to calibrate the model alternatives. The main objective of this work was to select the best model among the model alternatives developed. Also, model robustness and comparison of the model selection criteria used was investigated.

MATERIALS AND METHODS

Data used in deriving and testing the model

Volumetric spore traps (Burkard Manufacturing Co. Ltd, Hertfordshire, UK) sampled ascospores during 1992-1995, 1997-2001 and 2003 at Ås (south eastern Norway), 1993 and 1994 at Hjelmeland (south western Norway), and 2002-2005 at Gorsem (north eastern Belgium). A leaf bed of heavily infected, overwintered apple leaves surrounded the spore traps. The soil of the study area was drenched with a benzimidazole at 0.5 g a.i. litre^{-1}, to preserve the leaf litter during winter and spring. The spore traps sampled air at approximately 10 litre min^{-1}. The microscope tape attached to the clock cylinder was coated with a thin layer of a preheated mixture of Vaseline, toluene and kerosene. The tape was mounted on microscopic glass slides, and the number of ascospores recorded was adjusted for the proportion of tape examined and volume of air sampled, and recorded as spores m^{-3} air. Electronic data loggers provided records of precipitation, temperature, RH, and leaf wetness. Temperature and RH were recorded 1.5 to 2 m aboveground in weather shelters or radiation shields. Additionally, ascospores were trapped using the microscope slide technique in 2002-2004 at Lake Konstanz, southern Germany (Triloff 1997). Microscope slides were placed 4 mm above heavily infected leaves lying on a metal plate and placed in the orchard next to a weather station. Ascospores were released during periods of rain, captured by the slides and counted.

Daily accumulations of number of ascospores were used in the model development below. Environmental data were provided with daily mean air temperature T (°C), daily accumulated precipitation P (mm) and number of hours of leaf wetness LW (day^{-1}). Missing meteorological observations of air temperature and precipitation at Ås were filled with data from Bioforsk

Agrometeorological Service or with data from the Norwegian University of Life Sciences, both measured in the same area as the spore trap, but not inside the orchard. In a few cases, there were missing meteorological data that were estimated by the least square method.

The data were randomly divided ten times into one training set (2/3) used for model development and one test set (1/3) used to evaluate the model. All following calculations were done for the ten splits of the data individually, to ensure not fitting the model to only the specific data at hand, but to novel data from the same population **(**Bøvelstad et al. 2007).

Model development

A model to estimate ascospore maturation in *V. inaequalis* was developed by Gadoury and MacHardy (1982) in New Hampshire, USA, based on a linear relationship between the probit transform of matured ascospores and the degree day accumulation. The model was written

$$\Phi^{-1}(\boldsymbol{p}) + 5 = 2.51 \cdot 0.01\boldsymbol{DD} \quad (1)$$

where Φ is the standard normal cumulative distribution function, $\boldsymbol{p}$ is the proportion of mature ascospores, and $\boldsymbol{DD}$ is degree day accumulation calculated using a base temperature of 0 °C from time of green tip. This model is widely used, but it is not accurate in dry seasons (St-Arnaud and Neuman 1990, Stensvand et al. 2005). Stensvand et al. (2005) showed that a moisture frequency threshold of 7 consecutive dry days (a dry day was defined to have less than 0.2 mm precipitation and dew or fog occurring for less than 12 h) to adjust the degree day accumulation greatly improved the model for dry years, without substantially affecting the accuracy in wet seasons.

Five different mathematical models for the ascospore maturation in *V. inaequalis* were further developed from the New Hampshire (NH) model. While the original model was a linear model of the probit transformed cumulative proportions of mature ascospores, the new models were based on generalized linear models (GLM) (Agresti 2007) to directly model the cumulative proportion of mature ascospores. GLM consist of both logit and probit link functions that are variants of the same modeling, but based on different data assumptions. While logit link makes use of the natural logarithm of the odds of the proportion matured ascospores, probit makes use of the standard normal cumulative distribution function. The main difference is that logit link has slightly flatter tails, i.e; the probit curve approaches the

axes more quickly than the logit link curve, and in an early stage of this study, logit link was chosen over the probit link function because of a slightly better fit. The logit link function is defined as follow

$$p_i = \frac{e^{\alpha+\beta\cdot DD_i}}{1+e^{\alpha+\beta\cdot DD_i}} \quad (2)$$

in which p_i is the cumulative proportions of matured ascospores at day i, DD_i is the adjusted degree day accumulation at day i, and both α and β are unknown model parameters.

Model 1:

Based on the NH model, the degree day accumulation was calculated with a base temperature of 0 ℃. The degree day accumulation at day i was then defined by

$$DD_i = \sum_{j=1}^{i} \max(0, T_j) \quad (3)$$

with starting point j=1 at bud break of the apple tree and daily mean air temperature T_j at day j.

Model 2:

To adjust for the slow down or stop in the maturation process during dry periods, a halting of degree day accumulation was used if periods of 7 consecutive dry days or more occurred, according to Stensvand et al. (2005). The adjusted degree day accumulation could then be written

$$DD_i = \sum_{j=1}^{i} \max(0, T_j)\, I(n_j \leq 7) \quad (4)$$

were I is the identity function giving the value 1 if the statement is true and 0 otherwise and n_j is the number of consecutive dry days at day j.

Model 3:

A linear function between 1 and N_d consecutive dry days was formed to slow down or stop the maturation process for dry periods. The temperature at day j is weighted by 1 if only one consecutive dry day occurred, with 0 if N_d consecutive dry days occurred and with a linear function between. The accumulation could then be written

$$DD_i = \sum_{j=1}^{i} \max(0, T_j)\,(1 - n_j/N_d) I(n_j \le N_d) \quad (5)$$

where n_j is the number of consecutive dry days at day j.

Model 4:

In addition to stop the maturity process when periods of 7 or more consecutive dry days (Equation 4), the temperatures were weighted differently based on a weighting function for daily rate of plant growth based on three cardinal temperatures (Yan and Hunt 1999).

$$t_j = \max\left(0, \left(\frac{T_{max} - T_j}{T_{max} - T_{opt}}\right)\left(\frac{T_j - T_{min}}{T_{opt} - T_{min}}\right)^{\left(\frac{T_{opt} - T_{min}}{T_{max} - T_{opt}}\right)}\right) \quad (6)$$

with T_{min} fixed at 0 ℃. Following, the degree day accumulation could be written

$$DD_i = \sum_{j=1}^{i} t_j \max(0, T_j)\, I(N_d \le 7) \quad (7)$$

Model 5:

In addition to slowing down or stop the maturity process during periods of dry days (Equation 5), air temperatures were weighted differently (Equation 6). The degree day accumulation could then be written

$$DD_i = \sum_{j=1}^{i} \mathrm{t_j} \max(0, T_j)(1 - N_d/N) I(N_d \le N) \quad (8)$$

Model optimization

All five models were tuned according to technical and statistical approaches to optimize the prediction abilities of each model.

The total number of ascospores released is relatively low both early and late in the season (Gadoury and MacHardy 1982), and thus the most important periods of ascospore release and infection are between 5 and 95 % maturation. To see whether the prediction model could be improved by removing the tails of observations at both ends, calibration was also run when using only the observations between 5 and 95 % ascospores trapped in the training data.

Both the logit link function of GLM and the probit transformation generate models with a symmetrical slope around its maximum at 50 % matured ascospores. To account for a real world system not behaving symmetrically, the GLM was redefined according to Hosmer and Hjort (2002) to allow for asymmetry. The redefined model for the proportion at day *i* (Equation 2) was written

$$p_i = \left(\frac{e^{\alpha+\beta \cdot DD_i}}{1 + e^{\alpha+\beta \cdot DD_i}} \right)^{\lambda} \quad (9)$$

where λ is an unknown parameter.

The weighting function for air temperatures (Equation 6), originally used a fixed minimum cardinal air temperature (T_{min}) of 0 ℃, but was additionally estimated by using Kelvin degrees to allow for negative air temperatures.

Statistical analysis

Bayesian calibration

The Bayesian framework was used to calibrate the models. Based on Bayes theorem (Berger 1985), the prior parameter knowledge was updated with the new incorporated information through sampled data to form posterior functions of distributions of the parameters. The prior distributions of the parameters are our knowledge about the parameters before new data is collected. The model parameters mostly have physical meaning and available prior information based on existing data, expert opinions or literature review. Uninformative priors (one approach introduced by Jeffrey, 1961) may be used if no prior information is available.

Here, the prior knowledge was described by uniform distributions with minimum, maximum and reference given in Table 1. Prior independence was assumed, and the joint distribution found as the product of the marginal parameter distributions. If the sample size of collected data is small, or if the available data only provided indirect information about the parameters (Gelman 2002), the prior parameter knowledge will have a larger effect on the posterior. The numbers of mature ascospores were binomially distributed, and the likelihood function (Gelman et al. 1996) was defined by

$$L_{\boldsymbol{D}}(\boldsymbol{\theta}) = \prod_{i=1}^{m} \binom{m_i}{r_i} p_i^{r_i} (1-p_i)^{m_i - r_i} \quad (10)$$

where m_i is the total number of ascospores, r_i is the number of mature ascospores and p_i is the proportion of matured ascospores at day i determined as a function of the model parameters $\boldsymbol{\theta}$ (Equation 2). Calculations were done using the Markov chain Monte Carlo (MCMC) algorithm random walk Metropolis (Liu 2001), that was implemented in the computer program Matlab. The algorithm is iterative and starts with an initial guessed parameter set $\boldsymbol{\theta}^0=(\theta^0_1, \theta^0_2, \ldots \theta^0_n)$. Then, a candidate parameter set $\boldsymbol{\theta}'$ is generated, here using a multivariate normal distribution centered at the current point and with standard deviations found by trial and error to give an efficient acceptance rate between 0.15 and 0.5 (Gilks et al. 1996). A random variable u is then generated from the standard uniform distribution, and the proposed parameter set accepted ($\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}'$) if the ratio

$$\alpha = \frac{\pi(\boldsymbol{\theta}'|\boldsymbol{D})}{\pi(\boldsymbol{\theta}^t|\boldsymbol{D})} = \frac{\pi(\boldsymbol{\theta}')\cdot L_{\boldsymbol{D}}(\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}^t)\cdot L_{\boldsymbol{D}}(\boldsymbol{\theta}^t)} \quad (11)$$

Table 1: Symbols, units and limits for the prior interval and reference for all parameters used in this study together with widened priors used for some parameters.

Symbol	Unit	Prior interval	References	Widened prior interval
$\boldsymbol{\alpha}$		[-10 0]	This study	[-10 10]
$\boldsymbol{\beta}$		[0 0.7]	This study	
$\boldsymbol{N_d}$		[1 10]	Stensvand et. al (2005)	
$\boldsymbol{\lambda}$		[1 10]	This study	[1 100]
$\boldsymbol{T_{opt}}$	℃	[16 22]	MacHardy (1996)	
$\boldsymbol{T_{min}}$	℃	0	MacHardy (1996)	[-2 2]
$\boldsymbol{T_{max}}$	℃	[23 33]	MacHardy (1996)	

is greater or equal to u. Otherwise, the current parameter set is repeated ($\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^{t}$). The main idea of the MCMC algorithm is that the resulting chain of parameter sets will in the long run converge to the posterior distribution of the parameters. Monitoring convergence ("burn-in") of the chain is an important step in Bayesian analysis, since only the iterations after convergence can be regarded as samples from the posterior distribution. We ran two chains in parallel and detected burn-in by using the Gelman-Rubin diagnostic (Gelman and Rubin 1992) that compares the variability between and within the sequences by estimating a potential scale reduction factor (GR) for each parameter. Convergence was determined if GR was close to 1, less than 1.2 in practice (Gilks et al. 1996).

Model assessment and selection

The Deviance Information Criterion (DIC) (Spiegelhalter et al. 2002) is a model comparison method, combining model fit and penalty on model complexity, using the training data. It is defined by

$$DIC = \overline{D} + q_D \quad (12)$$

where $\overline{D}$ is the posterior mean of the deviance (quality of fit, calculated as -2 times the log-likelihood ratio of the reduced model compared to the full model, Agresti 2007) and q_D is the estimated model complexity. Here, it was individually calculated for all sets of training data. The model with the lowest DIC was preferred. The models generated using only the data from 5 to 95 % trapping and the initial NH model is not comparable with the other models according to DIC, since the models were based on different training data.

The model prediction abilities were compared by Root Mean Square Error of Prediction (RMSEP) (Hastie et al. 2001) for the ten sets of test data individually

$$RMSEP = \sqrt{\frac{1}{K}\sum_{k}^{K}(p_k - \hat{p}_k)^2} \quad (13)$$

where K is the number of test data, p_k is the proportion matured ascospores observed at time k, m_k is the total number of ascospores and $\hat{p}_k$ is the estimated proportion matured ascospores (Equation 2).

Additionally, RMSEP was calculated using only the conservative (most important infection period) test data between 5 and 95 % matured ascospores ($RMSEP_c$).

Analysis of variance (Montgomery 2005) was used to determine significance of model improvements. Both model and the optimization alternatives asymmetry and updating T_{min} were used as factors and the different splits of the data were treated as a block. The updating T_{min} option was only valid for Model 4 and Model 5 and for these models a three factors mixed model was ran with common restrictions on the fixed effects.

$$RMSEP_{ijkl} = \mu + M_i + A_j + T_k + (MA)_{ij} + (MT)_{ik} + (AT)_{jk} + (MAT)_{ijk} + B_l + \varepsilon_{ijkl} \quad \begin{cases} i = 4,5 \\ j = 0,1 \\ k = 0,1 \\ l = 1,\dots,10 \end{cases} \qquad (14)$$

where μ is the general effect, M_i the *i*th model effect, A_j is the effect of asymmetry ($j=0$ denotes symmetry and $j=1$ asymmetry) and T_k is the effect of the optimization alternative updating T_{min} ($k=0$ denotes a fixed T_{min} at zero, and $k=1$ denotes updating T_{min}). The interaction effects are determined by $(MA)_{ij}$, $(MT)_{ik}$, $(AT)_{jk}$ and $(MAT)_{ijk}$, B_l is the random block effect of the splits ($B_l \sim N(0,\sigma_1^2)$) and ε_{ijkl} is the error ($\varepsilon_{ijkl} \sim N(0,\sigma^2)$).

Also, the effect of T_{min} was removed from Equation 14, and the effect of model and asymmetry tested for all models ($i=1,2,\dots,5$). Finally, to test all 19 model versions that were calibrated in the study plus the original NH model, a mixed model with only one factor, namely the model version was ran, the splits of the data were still treated as a block.

The analysis of variance models were additionally constructed for $RMSEP_c$, but not for DIC, which is not comparable for all models.

RESULTS

The five models including different alternatives for model optimization were calibrated for all ten splits of the data into a training and a test set. Convergence were determined within the first half of the run for all model versions according to the shrink factor of Gelman-Rubins (1992) and the 50.000 iterations after burn-in were regarded as samples from the posterior parameter distributions.

Both DIC and RMSEP (including RMSEPc) were calculated for the splits individually and listed in Table 2 as means and standard deviations over the splits. RMSEP were calculated using the maximum posterior parameter estimate ($\boldsymbol{\theta}_{MAP}$) which was the parameter set after burn-in that had the highest joint posterior probability (arg max $\pi(\boldsymbol{\theta}|\boldsymbol{D})$) (Berger 1985). Analysis of variance models were constructed for RMSEP and $RMSEP_c$ to show significant effects.

The two factors mixed analysis of variance model (Equation 14 without T_{min}) with RMSEP as response, gave (according to an F-test) significance of both main effects; model ($p<0.0001$)

Table2: DIC, RMSE and RMSEc for the model alternatives calculated as mean and standard deviation over the ten splits of the data. Values written in italics are only comparable with other values in italics.

Model	**DIC $\cdot 10^7$**	**RMSEP**	**$RMSEP_c$**
New Hampshire		0.1811 (0.0343)	0.2228 (0.0395)
Original			
Model 1	9.3581 (1.9446)	0.1967 (0.0225)	0.2143 (0.0345)
Model 2	9.3479 (2.0154)	0.1912 (0.0240)	0.2139 (0.0344)
Model 3	9.1286 (1.9309)	0.1699 (0.0120)	0.1863 (0.0209)
Model 4	9.4485 (2.0590)	0.1982 (0.0260)	0.2192 (0.0385)
Model5	9.2519 (1.9674)	0.1864 (0.0181)	0.2044 (0.0286)
Conservative			
Model 1	*8.2059 (1.7948)*	0.2297 (0.0226)	0.2204 (0.0166)
Model 2	*8.2320 (1.8299)*	0.2213 (0.0169)	0.2172 (0.0139)
Model 3	*8.0835 (1.7803)*	0.1954 (0.0177)	0.1934 (0.0119)
Model 4	*8.2719 (1.8534)*	0.2301 (0.0192)	0.2221 (0.0157)
Model 5	*8.1420 (1.8048)*	0.2080 (0.0223)	0.2034 (0.0124)
Asymmetric			
Model 1	9.3016 (1.9206)	0.1889 (0.0239)	0.2085 (0.0345)
Model 2	9.2950 (1.9944)	0.1868 (0.0247)	0.2106 (0.0333)
Model 3	9.0600 (1.9122)	0.1645 (0.0133)	0.1804 (0.0208)
Model 4	9.3954 (2.0404)	0.1925 (0.0272)	0.2150 (0.0385)
Model 5	9.1813 (1.9433)	0.1775 (0.0151)	0.1955 (0.0232)
Updating T_{min}			
Model 4	9.1572 (2.0222)	0.1880 (0.0218)	0.2152 (0.0299)
Model 5	8.9938 (1.9028)	0.1679 (0.0107)	0.1864 (0.0169)
Asymmetric and updating T_{min}			
Model 4	9.1814 (1.9560)	0.1857 (0.0211)	0.2122 (0.0275)
Model5	8.9463 (1.8900)	0.1655 (0.0106)	0.1822 (0.0145)

and asymmetry (*p=0.0011*), but not for the interaction effect (*p=0.9447*). The highly insignificant interaction terms determined that the factors produced the same effect in the response at different levels. The estimate of the residual variance within groups (one split of the data into a training and a test set) were $\hat{\sigma}^2 = 0.9 \cdot 10^{-4}$ and the variance of the random effect between groups $\hat{\sigma}_1^2 = 3.7 \cdot 10^{-4}$.
While asymmetry significantly improved the model according to RMSEP, the conservative option worsened it and significance was not interesting and therefore not calculated.

The three factor mixed model (Equation 14) used for only Models 4 and Model 5 gave strong significance for both model (*p<0.0001*) and updating T_{min} (*p<0.0001*), but weaker insignificant effect for asymmetry (*p=0.0571*) and more highly insignificant effect for the interaction terms (*p>0.1777*). The estimate of the residual variance within groups were $\hat{\sigma} = 1.2 \cdot 10^{-4}$ and the variance of the random effect between groups $\hat{\sigma}_1 = 2.7 \cdot 10^{-4}$. The same conclusions were found for $RMSEP_c$, but the variances were larger, still the variance of the random effect between groups was larger than the estimate of the residual variance within groups.

As mean value over the different splits of the data, the comparison criteria DIC and RMSEP gave two different models as the best. While RMSEP (including $RMSEP_c$) resulted in the simple asymmetric version of Model 3 (Model 3') as the best model, DIC resulted in the more complex asymmetric version of Model 5 when updating T_{min} (Model 5') as the best. Both models include the linear weight to slow down the maturity process in dry periods (Equation 4), while Model 5' additionally contains the weighting function for air temperatures (Equation 6). The one factor analysis of variance model gave according to Tukeys-test, insignificant differences between Model 3' and the NH model (*p=0.6818*) and between Model 5' and the NH model (*p=0.7782*)) according to RMSEP. According to $RMSEP_c$, both differences were significant, Model 3' and the NH model (*p<0.01*) and Model 5' and the NH model (*p<0.01*)). No significant differences between the two models were found for neither of the criteria (RMSEP (*p=1.0*) and $RMSEP_c$ (*p=1.0*)).

The asymmetric optimization alternative did improve the fit significantly, but the model did still not fit the main structure in the data very well (Fig. 1). The prior intervals turned out to be a limiting factor for the degree of asymmetry, and widened prior intervals were tested up to a level, giving 0.3 % improvements according to RMSEP and 0.1 % improvements according to DIC.

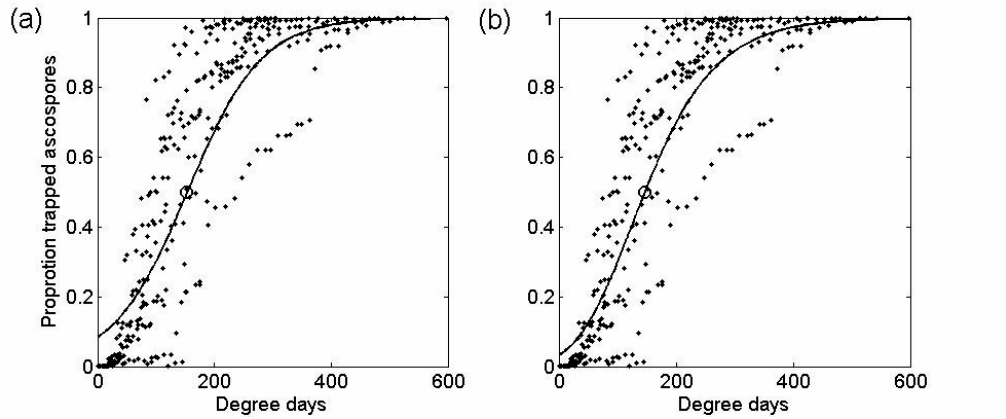


Fig. 1. Model 3 for one of the splits plotted with the belonging training data, (a) using GLM that makes it symmetric around the circle and (b) using the asymmetric version of GLM.

Compared to the NH model, Model 3' was 9.2 % improved according to RMSEP and 19.0 % improved according to $RMSEP_c$, while Model 5' was 8.6 % improved according to RMSEP and 18.2 % according to $RMSEP_c$ when looking at the mean value over the ten splits of the data into a training and a test set. Fig. 2 shows the NH model (Fig. 2a), Model 1 (Fig. 2b), Model 3' (Fig. 2c) and Model 5' (Fig. 2d) together with the test data for one of the splits of the data into one training and one test set.

Individually for each of the ten splits of the data, two different models were best according to DIC, and five and six, according to RMSEP and $RMSEP_{c,}$, respectively. The standard deviations for Model 3' and Model 5' were much reduced compared to the NH model (Table 2).

Both the two best models were calibrated using all observations, and point and interval estimates for the parameters found (Table 3). Both the mean ($\overline{\boldsymbol{\theta}}$) and maximum a' posterior ($\widehat{\boldsymbol{\theta}}_{MAP}$) estimates were given as point estimates for the parameters. The parameter reflecting when to stop the maturity process for dry periods after a slow down period (N_d) was close to one for both models and both point estimates. This gave a small weight for the air temperatures if one dry day, and zero if more than one consecutive dry days. The potential scale reduction factor at the end of the run (GR) was highest for β (GR=1.0044, Table 3) in Model 3', suggesting that additional simulation might reduce the posterior interval for this parameter by up to a factor of 1.0044. Parameter uncertainty reduced by collecting data was

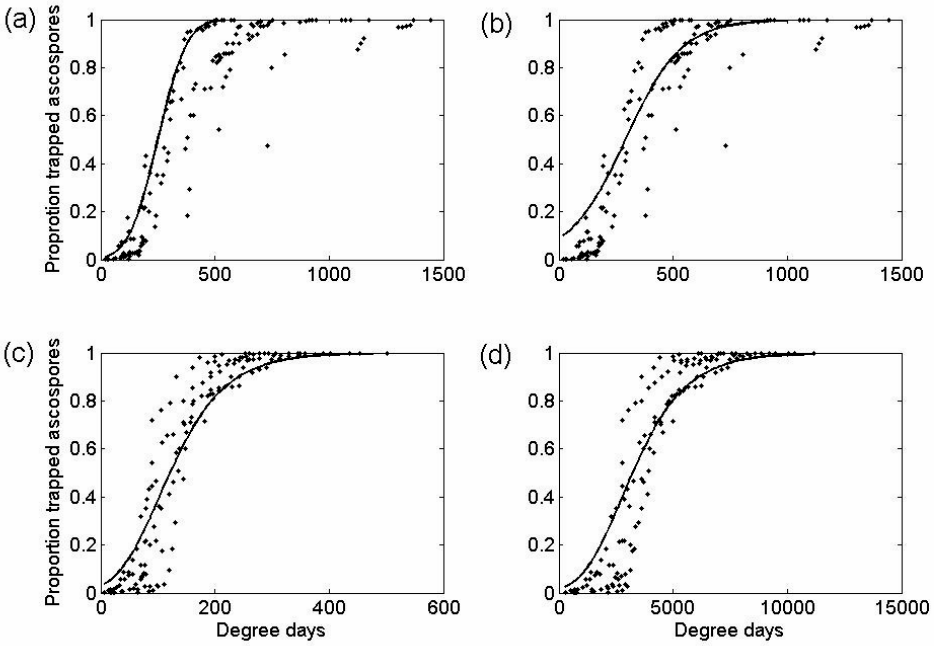


Fig. 2. Model results plotted together with the test data for one of the splits (a) the original New Hampshire model (Gadoury and MacHardy 1982), (b) Model 1, (c) Model 3' and (d) Model 5' (Kelvin degree days).

Table 3: The maximum posterior estimate ($\widehat{\boldsymbol{\theta}}_{\boldsymbol{MAP}}$), mean ($\overline{\boldsymbol{\theta}}$), the coefficient of variation (CV), the 95 % highest posterior density credible interval (HPD) and the potential shrink factor (GR) for the two best models found by treating all observations as training data.

Symbol	MAP	Mean	CV	95 % HPD	GR
Model 3'					
$\boldsymbol{\alpha}$	-0.0000	-0.0001	-1.0053	[-0.0003 0]	1.0000
$\boldsymbol{\beta}$	0.0163	0.0163	0.0012	[0.0163 0.0164]	1.0044
$\boldsymbol{N_d}$	1.0342	1.0330	0.0033	[1.0254 1.0386]	1.0029
$\boldsymbol{\lambda}$	5.4261	5.4268	0.0011	[5.4151 5.4385]	1.0018
Model 5'					
$\boldsymbol{\alpha}$	-0.0002	-0.0003	-0.9926	[-0.0008 0]	1.0008
$\boldsymbol{\beta}$	0.0007	0.0007	0.0013	[0.0006 0.0007]	1.0000
$\boldsymbol{N_d}$	1.1657	1.1659	0.0044	[1.1602 1.1760]	1.0009
$\boldsymbol{\lambda}$	6.5851	6.5812	0.0012	[6.5664 6.5967]	1.0027
$\boldsymbol{T_{opt}}$	16.0001	16.0001	0.0000	[16.0000 16.0003]	1.0002
$\boldsymbol{T_{min}}$	-2.0000	-1.9992	0.0000	[-2.0000 -1.9977]	1.0002
$\boldsymbol{T_{max}}$	32.9997	32.9996	0.0000	[32.9988 33.0000]	1.0000

determined by comparing the prior parameter intervals (Table 1) with the 95 % highest posterior density credible intervals (HPD) (Table 3).

DISCUSSION

The NH model generated to estimate the maturity of ascospores of *V. inaequalis* (Gadoury and MacHardy 1982) was further developed and optimized to better predict future system outcomes. The Bayesian statistical approach was used for model calibration, and ten different splits of the data were used in model development and testing. The split of the data turned out to be a crucial factor for the conclusions made both according to DIC and RMSEP (including $RMSP_c$,). Conclusions drawn based on only one split of the data did depend heavily on the split and would consequently not fit well to novel data from the same population. Conclusions were therefore based on the mean output and the standard deviation over the ten splits. The two model criteria for model comparison (DIC and RMSEP) gave two different models as the overall best model (Model 5' and Model 3', respectively). According to RMSEP and $RMSEP_c$, no significant differences in a Tukey-test were found between the two models. Compared to the NH model, significant differences in a Tukey-test were found for both models according to $RMSEP_c$, but not for RMSEP. The standard deviation over the ten splits was much larger for the NH model than for Model 3' and Model 5'. This indicated that both models gave more similar prediction error over the splits, while the NH model varied more with seasons/years.

Two criteria were used for model comparison, DIC and RMSEP. While RMSEP calculates the models prediction abilities by use of the test data, DIC makes use of the training data. By using the training data, it is always possible to construct a model with perfect fit, as long as the model is complex enough. But, when a model becomes more complex, it is also able to adapt more complicated underlying structures, and the prediction error will increase. In between there is an optimal model complexity that gives minimum test error (Hastie et al. 2001), and DIC is one measurement to estimate this optimum. Here, the two model comparison methods, gave two different conclusions, where DIC preferred a more complex model than RMSEP. Also the more known pairwise model selection criteria Bayes factor (Kass and Raftery 1995), providing the relative evidence of one model compared to another, could have been used, but the high number of models to be compared, made DIC much more efficient.

Model improvements were clearly seen while Model 3' (Fig. 2c) and Model 5' (Fig. 2d) concentrated the data closer together to fit the data better. Although the NH model (Fig. 2a) and Model 1 (Fig. 2b) are almost the same models, they looked very different, since the different training data used generated different parameter estimates for the models. The combination of high degree day accumulations in the data and the symmetry of GLM, gave larger gaps between the data and the model for small degree day accumulations in Model 1. This was not the case for the NH model, since the model was trained on data not including those high degree day accumulations.

The asymmetric version of GLM significantly improved the models, but still the resulting model did not seem to follow the main structure in the data. The prior intervals were a limiting factor for the degree of asymmetry. Both the prior intervals for the technical parameters α and β were chosen widely to give a shape of the model in the same direction as the data. For the third technical parameter λ, there were no available prior information other than it had to be higher than one to give asymmetry in the correct direction. Additionally, the new parameter had a direct influence on the other parameters and consequently α converged at its prior boundary, while an increased prior interval for α, made λ converged at its boundary. The prior interval for λ increased up to 100 for its maximum and still the parameter chain converged at its boundary, giving only 0.3 % improvements according to RMSEP and 0.1 % improvements according to DIC. Even more widened priors were not used because of increasing computational difficulties with increasing prior intervals combined with only small improvements according to RMSEP and DIC. For a better fit, the Gompertz function (Vieira and Hoffman 1977) which in contrast to the logit link function is asymmetric, could have been tested.

The cardinal temperatures in the weighting function for air temperatures (Equation 6) also converged at their prior boundaries (Table 3). Since the model was not a perfect representation of the real world system, the parameters will neither correspond perfectly with their corresponding physical values. Widened or other prior distributions could have improved the model, but would also give unreasonable physical interpretations of the parameters and was therefore not used.

Fig. 3 shows the development of the parameter distribution of N_d in Model 3', from only the prior knowledge through added collected data for all sites/years. The distribution of the parameter became narrower already when only two years of collected data were included,

long before the position of the distribution converged. This indicated a model fitting well to the specific data at hand and not to novel data from the same population. The parameter estimates will continue to change until the added data reflects the entire population (including variation in season dryness, number and length of dry periods, extreme weather). Two seasons were clearly not enough to include all those factors, but unfortunately the data were still treated as representing the whole population. According to the figure, the posterior parameter

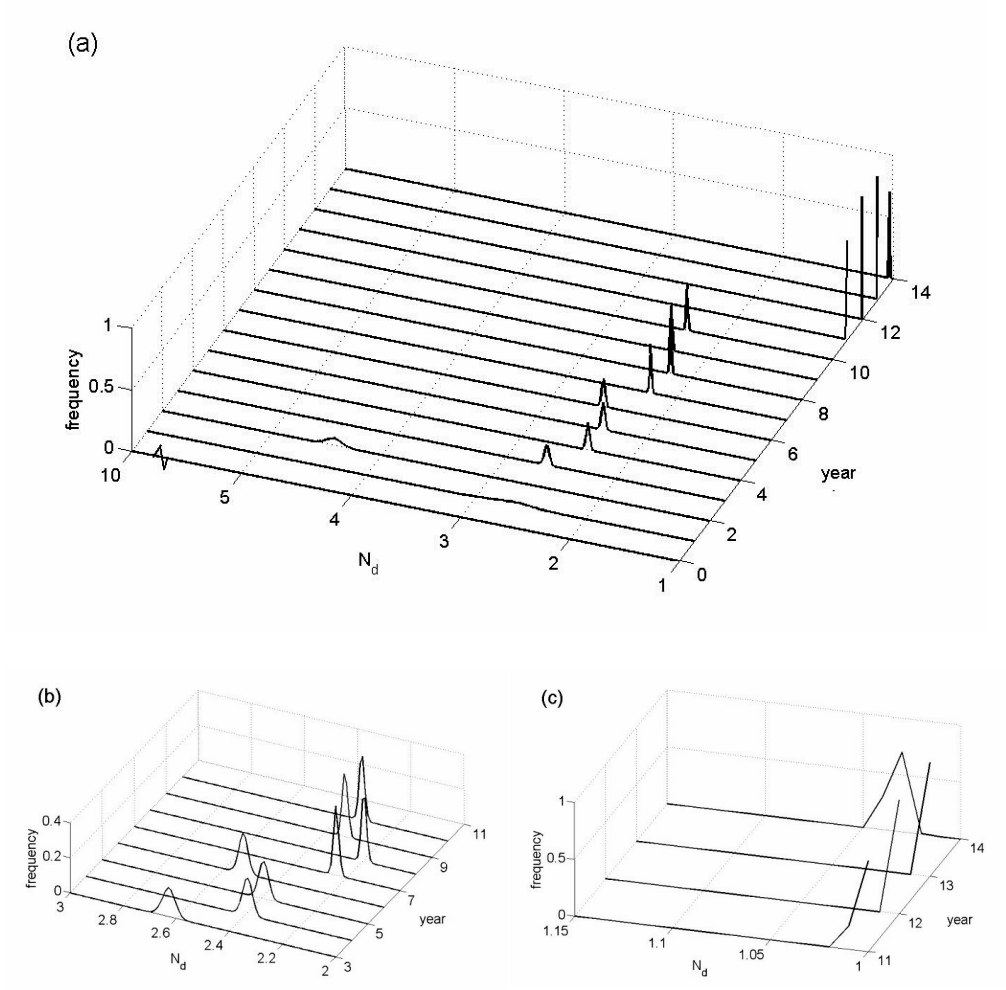


Fig. 3. Progress in parameter knowledge from adding data for each year is shown by plotting the estimated posterior parameter distribution for N_d in Model 3', (a) all collected data, (b) adding between 3 and 10 years of data and (c) adding 11 to 14 years of data.

distribution seemed to converge after only three years of incorporated data (Fig. 3b), but big changes happened after 11 years (Fig. 3c). In the first 10 years, only Norwegian data from Ås and Hjelmeland appeared. After 11 years, additionally German and Belgian data were included, and the posterior distribution became approximately exponential with modal value at zero. Then again, after 14 years, only data from Belgium were added, and smaller changes appeared in the posterior distribution, that become approximately Gaussian. This may indicate that the data from Germany affected the model differently from the other places, and this may be due to the different type of spore traps used. The leaves in the spore traps at Lake Konstanz were not kept in direct contact with the ground, and this may have lead to an earlier dry-up of the leaf litter in periods between rain events. Protracted dry periods will halt the ascospore maturation. (Keitt and Jones 1926, Wilson 1928, James and Sutton 1982a, James and Sutton 1982b, O'Leary and Sutton 1986, Schwabe et al. 1989, Stensvand 2005), and thus the German samples may have slowed down their maturation earlier than was accounted for by halting degree days from day 7 without rain (Stensvand et al. 2005). This also seems to be the case, because a slight delay in spore trapping in comparison to what was predicted by the models was observed (A.-G. Roer, unpublished data).

In the analysis of variance models used, the ten splits of the data were treated as a random block effect, which partitioned the total variability in the observations into one component that measured the variation between the splits and one component that measured the variation within splits. The estimated variances showed that most of the variability was attributed to differences between the splits of the data. Each split of data contains the same number of years and seasons in combination, but each combination do not contain the same number of trapped spores, which may cause the large variability between splits.

This study has shown the importance of using several splits of the data in model development and model assessment. Splits of the data may be of great importance, and conclusions made on calculations for only one split may depend heavily on the split and not be general for novel data. Additionally, by adding one year of data at a time to the calibration process, we saw that lots of data were needed to stabilize the model parameters. We have also seen that the two different statistical criteria, DIC and RMSEP gave different conclusions. In a Tukey-test, none of the two best models were significantly different from the NH model according to RMSEP, but both were significantly different according to $RMSEP_c$. Since differences were found when looking at the most important infection period between 5 and 95 % maturation, we will recommend the new models over the NH model. In a Tukey-test, no significant difference

between the two models was found, and plots of them (Fig. 2) were very similar. Over the two models we would therefore recommend the simplest model, Model 3'. Additionally the model is constructed for prediction, and Model 3' had the smallest prediction error both according to RMSEP and $RMSEP_c$.

ACKNOWLEDGEMENTS

We would like to thank Peter Triloff for providing us with three years of spore trapping data from Lake Konstanz in Germany.

REFRENCES

Agresti A. (2007) An introduction to categorical data analysis. Wiley

Berger O J. (1985) Statistical Decision Theory and Bayesian Analysis, New York, Springer-Verlag

Bøvelstad HM, Nygård S, Størvold HL, Aldrin M, Borgan Ø, Frigessi A, Lingjærde OC. (2007) Predicting survival from microarray data - a comparative study. Bioinformatics 23:2080-2087

Brook PJ. (1976) Seasonal pattern of maturation of *Venturia inaequalis* ascospores in New Zealand. N. Z. J. Agric. Res. 19:103-109

Childs L. (1917) New facts regarding the period of ascospore discharge of apple scab fungus. Oreg. Agric. Exp. Stn. Bull. 143

Fjelddalen J. (1948) Frukthagens fiende nr 1 – epleskurven. Frukt og Bær 1:66-80

Frey CN, Keitt GW. (1925) Studies of spore dissemination of *Venturia inaequalis* (Cke.) Wint. in relation to seasonal development of apple scab. J. Agric. Res.15:529-540

Gadoury DM, MacHardy WE. (1982) A Model to Estimate the Maturity of Ascospores of *Venturia inaequalis*. Phytopathology, 72:901-904

Gelman A. (2002) Prior distribution. Encyclopedia of Environmetrics 3:1634-1637

Gelman A, Rubin BC. (1992) Inference from iterative simulation using multiple sequences. Statistical Science 7:457-511

Gelman A, Carlin JB, Stern HS, Rubin DB. (1996) Bayesian Data Analysis. Chapman & Hall

Gilks WR, Richardson S, Spiegelhalter DJ. (1996) Markov Chain Monte Carlo in Practice. Chapman & Hall

Gjærum HB. (1954) Ascospore maturity, dissemination and infection by apple scab. Nor. Plant Prot. Inst. Bull. 13

Goldstein M, Rougier J. (2006) Bayesian linear calibrated prediction for complex systems. Journal of the American Statistical Association 101:1131-1143

Hårdh JE. (1955) Apple scab and its control in Finland. Fin. Agric. Res. Board Pub. 144

Hastie T, Tibshirani R, Friedman J. (2001) The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Springer

Hosmer DW, Hjort NL. (2002) Goodness-of-fit processes for logistic regression: simulation results. Statist. Med. 21:2723-2738

James JR, Sutton TB. (1982a) Environmental factors influencing pseudothecial development and ascospore maturation of *Venturia inaequalis*. Phytopathology 72:1073-1080

James JR, Sutton TB. (1982b) A model for predicting ascospore maturation of *Venturia inaequalis.* Phytopathology 72:1081-1085

Jeffreys H. (1961) Theory of probability. Oxford University Press

Kass RE, Raftery AE, (1995) Bayes factor, Journal of the American Statistical Association 90:773-795

Keitt GW, Jones LK. (1926) Studies of the epidemiology and control of apple scab. Wis. Agric. Exp. Stn. Res. Bull. 73

Kennedy MC, O'Hagan A. (2001) Bayesian Calibration of Computer Models. Journal of the Royal Statistical Society B63:425-464

Lagarde MP. (1988) Etudes sur la maturation des ascospores de *Venturia inaequalis* (Cke.) Wint. en vue de 1'élaboration d'un modele. Annales I I/ I I I 4:1093-1098

Liu JS. (2001), Monte Carlo Strategies in Scientific Computing. Springer.

MacHardy WE. (1996). Apple Scab, Biology, Epidemiology, and Management. APS Press

Massie LB, Szkolnik M. (1974) Prediction of ascospore maturity of *Venturia inaequalis* utilizing cumulative degree-days. Proc. Am. Phytopathol. Soc. 1:40

Moller WJ, Latorre BA, Docampo D. (1971) Liberación de inòculo primario de *Venturia inaequalis* (Cke.) Wint., en Chile. Agric.Tec. 31:27-33

Montgomery DC. (2005) Design and Analysis of Experiments. Wiley International Edition

Norin I. (1989) Skorvvarning – erfarenheter av 1988 års forsøk. Frukt- och Bärodling 1:42-45

O'Leary AL, Sutton TB. (1986) The influence of temperature and moisture on the quantitative production of pseudothecia of *Venturia inaequalis*. Phytopatology 76:199-204

Rossi V, Ponti I, Marinelli, M., Giosué S, Bugiani R. (1999). Field evaluation of some models estimating the seasonal pattern of air-borne ascospores of *Venturia inaequalis*. J. Phytophatol. 147:567-575

Schneiderhan FJ. (1925) Rainfall in relation to ascospore discharge and infection in *Venturia inaequalis*. (Abstr.) Phytopatology 15:56

Schwabe WFS, Jones AL, van Blerk E. (1989) Relation of degree-day accumulations to maturation of ascospores of *Venturia inaequalis* in South Africa. Phytophylactica 21: 13-16

Spiegelhalter DJ, Best NG, Carlin PB, Van der Linde A. (2002) Bayesian measures of model complexity and fit (discussion), Journal of the Royal Statistical Society 54, 583-616

St-Arnaud M, Neumann P. (1990) Évaluation au Québec d`un modèle de la prédiction de la fin de la période annuelle d`éjection des ascospores du *Venturia inaequalis*. Phytoprotection 71:17-23

Stensvand A, Eikemo H, Gadoury DM, Seem RC. (2005) Use of Rainfall Frequency Threshold to Adjust a Degree-Day Model of Ascospore Maturity of *Venturia inaequalis.* Plant Disease 89:198-202

Stover WG, Johnson HW. (1924) First progress report on the study of apple scab under Ohio conditions. (Abstr.) Phytopathology 14:60

Szkolnik M. (1969) Maturation and discharge of ascospores of *Venturia inaequalis*. Plant Dis. Rep. 53:534-537

Szkolnik M. (1974) Apple scab ascospore release as related to apple fruit bud development and calendar dates. Proc. Am. Phytopathol. Soc. 1:146.

Triloff P. (1997). Apple scab control with the simulation programme "RIMpro" at Lake Constance, Germany: results and experiences in the past three years. Proceedings of 4the workshop on integrated control of pome fruit diseases. . IOBC wprs Bulletin 20 (9):229-240

Vieira S, Hoffman R. (1977) Comparison of the Logistic and the Gompertz Growth Functions Considering Additive and Multiplicative Error terms. Applied Statistics. 23:143-148

Weber A. (1934-35) Undersøgelser over æble-skurvens (*Venturia inaequalis*) overvintring. Tidsskr. Planteavl. 40:754-758

Weber A, Jørgensen HA. (1953) Forsøk med bekæmpelse af æbleskurv efter løvfald samt undersøkelse over skurvens modningstid. Tidsskr. Planteavl 56:443-468

Wiesman R. (1935) Untersuchungen über die Bedeutung der Ascosporen (Wintersporen) und die Konidien and den schorfigen Trieben für die entstehung der Primärinfektionen des Apfelschofpilzes *Fusicladium dendriticum*. Separatabdruck aus dem Landwirtschaftlichen Jahrbuch der Schweiz 1953.

Wilson EE. (1928). Studies of the development of the ascigerous stage of *Venturia inaequalis*. Phytopathology 18:375-420

Wolfson LJ, Kadane JB, Small MJ. (1996) Bayesian environmental policy decisions: two case studies. Ecological Applications 6:1056-1066

Yan W, Hunt LA. (1999) An Equation for Modelling the Temperature Response of Plants using only the Cardinal Temperatures. Annals of Botany Company 84: 607-614

Ylämäki A. (1989) Kokemuksia Biomatruivaroituslaitteesta. Puutarha 92:410-411

IV

The influence of the likelihood function in Bayesian calibration to a snow depth model

Anne-Grete Roer*[ab], Trygve Almøy[b], Trond Rafoss[a]

[a] Norwegian Institute for Agricultural and Environmental Research, Plant Health and Plant Protection Division. Høgskoleveien 7, N-1432 Ås, Norway

[b] Norwegian University of Life Science, Department of Chemistry, Biotechnology and Food Science, N-1432 Ås, Norway

*Corresponding author. Tel.: +47 922 83 427; fax: +47 649 46 110.

Email address: anne-grete.roer@bioforsk.no, (Anne-Grete Roer)

Abstract

The Bayesian statistical framework is used to calibrate a process based model of snow depth using the random walk Metropolis algorithm. Different versions of the likelihood function; the Gaussian and the fat tailed Gaussian (Sivias') using fixed standard deviations of 20, 30 and 40 % of the observed value, estimation from data and estimation as one constant standard deviation was used. The calculations were time consuming, and the likelihoods using fixed standard deviations gave the most time effective calibration simply because they had one less parameter to estimate. Another result was that the fatter tails of the likelihood were, the faster the calibration process went. Impact on the estimates was small for different likelihood functions; whereas the fat tailed with fixed standard deviations of 40 % gave the overall smallest prediction error.

Key words: The Gaussian likelihood, Sivias' likelihood, estimated covariance matrix, fixed covariance matrix

INTRODUCTION

Computer models are built in many fields of science to simulate real world systems. Unfortunately, outputs from the model representation of the system will generally not correspond perfectly with the behavior of the real world phenomena being modeled, but will be hampered by uncertainties. First, the model is a simplification of the real world system, second, the model contains unknown parameter values, and third, data used in the model are related with measurement error (Goldstein and Rougier 2006). In process based models, the model parameters usually have a physical meaning, but they may be difficult to measure and therefore hard to obtain precise information about. In situations with substantial uncertainty but where prior parameter knowledge is available, the Bayesian calibration approach is natural to use (Kennedy and O'Hagan 2001).

The Bayesian statistical approach differs from the classical approach by regarding parameters as random variables instead of as fixed values. Prior knowledge about the parameters expressed by a prior probability distribution is combined with information from the available data through the likelihood function (the joint probability density function of the data, considered as a function of the model parameters for the fixed observed data) to produce posterior probability distributions of the parameters, describing the parameter uncertainty after new information is incorporated.

The collected data only affect the posterior distribution through the likelihood function. Commonly, the likelihood function is determined by the distribution of the model errors assumed to be additive and Gaussian (Marshall et al. 2004, Marcel van Oijen et al. 2005a, Hue et al. 2008, Hassan et al. 2009). Outliers can be accommodated by assuming a fat-tailed distribution, as the t-distribution (Gilks et al. 1996) or Sivias constrained Gaussian distribution (Sivia 2006). The covariance matrix of model error is usually unknown. Estimation of it can easily be done for small data sets (Hue et al. 2008), but high dimensional parameter spaces (i.e. higher dimensions on the unknown covariance matrix) will follow from larger amounts of data and make the computation difficult. The problem is therefore often simplified by assuming a constant error variance for all data points (Marshall et al. 2004, Hassan et al. 2009) or by using a fixed covariance matrix (Van Oijen et al. 2005a).

For complex models with high dimensional parameter spaces and large amounts of training data, convergence is generally hard to detect and a long period by trial and error to form a reasonable proposal parameter distribution is needed.

This study compares two different likelihood functions of the data, namely the Gaussian distribution and a fat tailed Gaussian distribution (Sivias'). Both distributions applied diagonal covariance matrixes that was either estimated as a constant error variance or as a percent of the measurements, or fixed at 20, 30 or 40 % of the measurements.

The main objective of this paper was to learn about the effect on model output caused by the choice of likelihood by quantify the effect of the chosen likelihood and the connected covariance matrix. It is of great importance to know whether the likelihood distribution and its covariance matrix should be carefully chosen or whether a reasonable quantification of the joint distribution of the data, giving easier convergence, is adequate. In this paper, a complex dynamic model of snow cover (Thorsen and Haugen 2007) was used as case.

MATERIALS AND METHOD

The snow depth model

A plant model for timothy and perennial ryegrass was developed to forecast winter climate impacts on forage crops (van Oijen et al. 2005b). The model is driven by climate data and includes a sub model to estimate snow depth (Thorsen and Haugen 2007). The sub model is a process based model using a set of different equations to describe the physical dynamic of the system, taking air temperature and precipitation as input to the model containing eight parameters. The model was denoted

$$D_t = M_t(\boldsymbol{\theta}, \boldsymbol{X}_t) + \varepsilon_t \quad (1)$$

where D_t is the observed snow depth at time *t*, $M_t(\boldsymbol{\theta}, \boldsymbol{X}_t)$ is the corresponding model output at time *t*, taking the state variables $\boldsymbol{X}_t$ and parameter set $\boldsymbol{\theta}$ as inputs, and $\varepsilon_t$is model errors reflecting both measurement error and model inadequacy. Further, the error were assumed independent Gaussian distributed

$$\varepsilon_t \sim N(0, \sigma_t^2) \quad (2)$$

where σ_t is the standard deviation of model error at time *t*, assumed independent.

The snow dept data used to calibrate and test the model was collected daily data from Kise in Norway, situated 60.77N, 10.8E, 127 meters above sea level in the time period 1988 to 2003.

The first 10 years were used for model development (training data), and the remaining 5 years for model assessment (test data).

The Bayesian approach

The posterior distribution of interest ($\pi(\boldsymbol{\theta}|\boldsymbol{D})$) was found as a combination of prior parameter knowledge ($\pi(\boldsymbol{\theta})$) and new incorporated information through the likelihood function ($L_{\boldsymbol{D}}(\boldsymbol{\theta})$) of the measurement data, according to Bayes theorem (Berger 1985)

$$\pi(\boldsymbol{\theta}|\boldsymbol{D}) \propto \pi(\boldsymbol{\theta}) \cdot L_{\boldsymbol{D}}(\boldsymbol{\theta}) \quad (3)$$

Exact calculations were impossible because of integral problems in high dimensional spaces, and the Markov chain Monte Carlo (MCMC) algorithm random walk Metropolis (Liu 2001) was therefore used. The algorithm was implemented in Matlab. It is an iterative algorithm that starts with an initial guessed parameter set $\boldsymbol{\theta}^0$. A proposal parameter set $\boldsymbol{\theta}'$ is generated, from the proposal distribution, which is an independent and identically distributed function describing the step length added to the current state. The acceptance rate

$$\alpha = \min\left(1, \frac{\pi(\boldsymbol{\theta}') \cdot L_{\boldsymbol{D}}(\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}^i) \cdot L_{\boldsymbol{D}}(\boldsymbol{\theta}^i)}\right) \quad (4)$$

is calculated and the proposed parameter set accepted ($\boldsymbol{\theta}^{i+1} = \boldsymbol{\theta}'$) if the acceptance rate α (Equation 4) is greater of equal to a random number simulated from the standard uniform distribution. Otherwise, the previous parameter set is repeated ($\boldsymbol{\theta}^{i+1} = \boldsymbol{\theta}^i$).

The proposal distribution controls the effectiveness of the Metropolis algorithm. It is here used as a Gaussian function, centered at the current state and with standard deviations tuned by trial and error to give an acceptable acceptance rate between 0.15 and 0.5 (Gilks et al. 1996). The idea of MCMC is that the resulting Markov chains of the parameters will inn the long run converge to the posterior parameter distributions. To detect the state of convergence (burn-in), four chains are run in parallel and the scale reduction factor proposed by Gelman and Rubin (1992) used.

The prior parameter distributions represent our knowledge about the parameters before new data are incorporated. Usually, the prior distributions are based on existing data, expert opinions or literature. If no prior information is available, non informative priors (Jeffrey 1961) may be used and the inferences will only be affected by the data. Here, the prior

Table 1: Upper and lower limits used to construct the prior distributions. Modal values are present for the beta distributed prior parameters distributions. Uniform distributions were used for the rest of the parameters.

Parameter	Lower limit	Upper limit	Modal value
T_{rs}	-10	10	0.5
T_{mf}	-10	10	0.5
ξ	0	1	-
ΔK_{max}	0	10	1.25
K_{min}	0	10	2
SW_{rf}	0	10	0.01
ρ_{ns}	10	250	-
SW_{ret}	0	1	0.1
β	0	1	-
λ	0	1	-

distributions used are uniform and beta distributed (Table 1). The upper and lower limits are chosen relatively wide according to expert opinions, and the modal values used in the beta distributions are based on results by Engseth et al. (2000).

Through collected data, the likelihood function modifies the prior information into posterior parameter distributions. The more experimental data added the greater effects on the posterior distribution will the likelihood function have. The prior information remains important if the available data provide only indirect information about the parameters of interest, or if the sample size is small (Gelman 2002). We use the likelihood function after some simplifications determined by the distribution of the model errors (including both measurement error and model inadequacy) according to Rougier (2007).

$$L_{\boldsymbol{D}}(\boldsymbol{\theta}) = f(\boldsymbol{D}|\boldsymbol{\theta}, \boldsymbol{X}_t) = \prod_{t=1}^{T} \varphi(D_t - M_t(\boldsymbol{\theta}, \boldsymbol{X}_t); 0, {\sigma_t}^2) \quad (5)$$

were T is the number of training data and φ denotes the Gaussian or Sivias' probability density function with given mean and standard deviation (assumed independent).

Gaussian Likelihood

The original Gaussian probability density distribution (Miller and Miller 1999) used in Equation 5 gives the following likelihood function

$$L_{\boldsymbol{D}}(\boldsymbol{\theta}^*) = \prod_{t=1}^{T} \frac{1}{\sigma_t (2\pi)^{1/2}} \exp((D_t - M_t(\boldsymbol{\theta}^*, \boldsymbol{X}_t))^2 / 2\sigma_t^2) \quad (6)$$

where $M_t(\boldsymbol{\theta}^* | \boldsymbol{X}_t)$ determines model output at time *t* using the specific parameter set $\boldsymbol{\theta}^*$ and input variables $\boldsymbol{X}_t$.

Sivias' Likelihood

Sivias' probability density distribution (Sivia 2006) is constructed to handle outliers in the collected data. A variant of Jeffreys' prior is used to specify a lower boundary of the standard deviation. Sivias' distribution used in Equation 5 gives the following likelihood function

$$L_{\boldsymbol{D}}(\boldsymbol{\theta}^*) = \prod_{R_t \neq 0} \frac{1}{\sigma_{0t} (2\pi)^{1/2}} \left(1 - \exp(-R_t^2/2)\right) / R_t^2 \cdot \prod_{R_t = 0} \frac{1}{2\sigma_{0t} (2\pi)^{1/2}} \quad (7)$$

where σ_{0t} is the lower bound of the standard deviation and $R_t = (D_t - M_t(\boldsymbol{\theta}^*, \boldsymbol{X}_t))/\sigma_{0_t}$.

The standard deviation

The covariance matrix was defined diagonal with variances (σ_t^2) at the diagonal. With no information about the precision of the measurements or the models inadequacy, the standard deviations can be chosen in different ways. According to van Oijen et al. (2005a) it is appropriate to assume a standard deviation of 30 % of the measured value and a negligible model inadequacy. The standard deviation can generally be defined by

$$\sigma_t = \min(0.1, \beta \cdot D_t) \quad (8)$$

where the minimum of 0.1 is included to avoid a standard deviation of zero, and D_t is the observed value at time *t*. To see the effect of the standard deviation, the calculations were done when fixing β at 20, 30 and 40 % of the measured value. Also, β was treated as an unknown parameter and estimated. Additionally, the standard deviation was estimated as a constant

$$\sigma_t = \lambda \quad (9)$$

Model assessment

Root Mean Square Error of Prediction (RMSEP) (Hastie et al. 2001) calculates the prediction error of a model. It is here defined as an average ($\overline{RMSEP}$) over 10.000 parameter sets randomly drawn from the Markov chains

$$\overline{RMSEP} = \frac{1}{E}\sum_{e=1}^{E}\sqrt{\frac{1}{K}\sum_{t=1}^{K}(D_t - M_t(\boldsymbol{\theta}_e, \boldsymbol{X}_t))^2} \quad (10)$$

where E is the number of parameter sets (ensembles), K is the total number of test data, D_t is the observed value at time t and $M_t(\boldsymbol{\theta}_e, \boldsymbol{X}_t)$ is estimated model output at time t using parameter set e.

Root Mean Square Deviation (RMSD) (Iizumi et al. 2009) quantifies the uncertainty in model outputs derived from the parameters. Here 10.000 model outputs (ensembles) are estimated for each day, using parameter sets randomly drawn from the resulting Markov chains. M_{et} is the model output at time t using parameter set number e, $\overline{M_{\cdot t}}$ is an ensemble mean model estimate at time t, K is the number of data in the test set and E is the number of ensembles used.

$$\overline{RMSD} = \frac{1}{E}\sum_{e=1}^{E}\sqrt{\frac{1}{K}\sum_{t=1}^{K}\left(M_{et}(\boldsymbol{\theta}_e, \boldsymbol{X}_t) - \overline{M_{\cdot t}(\boldsymbol{\theta}_{\cdot}, \boldsymbol{X}_t)}\right)^2} \quad (11)$$

Analysis of variance (Montgomery 2005) was used to determine significant differences in estimated $\overline{RMSEP}$ and $\overline{RMSD}$ for the different combinations of likelihood function and covariance matrix.

$$y_{ij} = \mu + L_i + C_j + \varepsilon_{ij} \begin{cases} i = 0,1 \\ j = 1, \ldots, 4 \end{cases} \quad (12)$$

where y_{ij} is the response (either $\overline{RMSEP}$ of $\overline{RMSD}$) using likelihood function i and covariance matrix j, μ is the general effect, L_i is the effect of the choice of likelihood function, C_j is the effect of the chosen covariance matrix and ε_{ij} is he error ($\varepsilon_{ij} \sim N(0,\sigma^2)$).

RESULTS

Nine combinations of the choice of likelihood function and covariance matrix (Equation 5) were used to calibrate the snow depth model. Four Markov chains were run in parallel for 100.000 iterations and the shrink factor introduced by Gelman and Rubin (1992) detected burn-in for all combinations after less than 50.000 iterations. The last 50.000 iterations were kept and regarded as samples from the posterior parameter distributions. Summary statistics gave point parameter estimates for each combination of likelihood function and covariance matrix, found as the parameter set with maximum posterior probability ($\widehat{\boldsymbol{\theta}}_{MAP} = \arg\max(\pi(\boldsymbol{\theta}|\boldsymbol{D}))$) (Berger 1985).

For all choices, the covariance matrix in the likelihood function was assumed to be diagonal (that means no correlations). In six combinations the standard deviations at the diagonal in the covariance matrix were assumed fixed and in three of the combinations they were estimated. For both the Gaussian and Sivias' likelihood function they were estimated as a percent of the observed value, defined to be at least 0.1 (Equation 8). Additionally for the Gaussian likelihood function, the standard deviation was estimated as a constant (Equation 9).

Histograms of the Markov chains after burn-in were plotted in Fig. 1 and shows small estimated quantities for the standard deviations. When using both the Gaussian (Fig. 1a) and Sivias' (Fig. 1b) likelihood function, the posterior parameter distribution of the percent of the observed value become relatively uniform up to approximately 18 % (0.18) for both functions, but with a peak between 15 and 20 % for the Gaussian. The maximum posterior parameter (MAP) estimates become respectively 15.93 % (0.1593) and 9.87 % (0.0987) for the Gaussian and Sivias' likelihood function. Both estimates become smaller than the fixed percents of 20, 30 and 40 % used. The constant standard deviation was only estimated for the Gaussian likelihood function (Fig. 1c), and the Markov chain formed approximately a Gaussian distribution ($\hat{\sigma} \sim N(0.0433, 0.0005)$) for the constant with MAP estimate of 4.33 % (0.0433).

Snow depths were predicted daily in the test period from 1998 to 2003 at Kvithamar for each of the nine combination of likelihood function and covariance matrix, when using the MAP parameter estimates. For each day in the test period, the largest difference between the nine predictions of snow depth was calculated and plotted in Fig.2. The largest differences were of about 25 cm, one the winter 1998/1999 and one the winter 2002/2003, both lasted for only a few days. Also, a longer period for almost a month the winter 1998/1999 contains larger

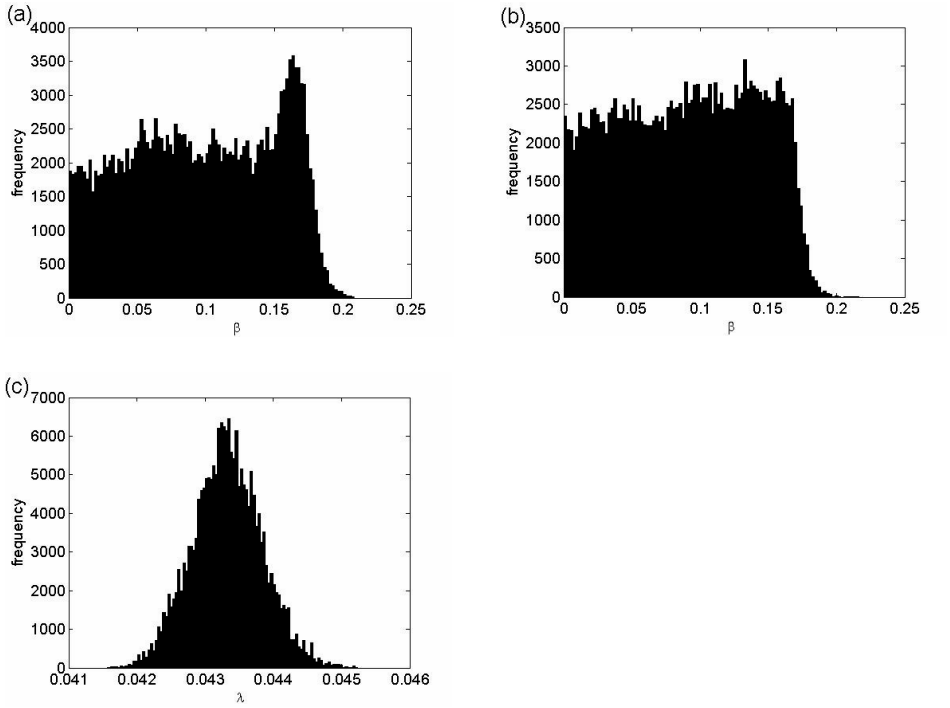


Fig. 1: Histograms of the posterior parameter chains for the estimated standard deviations (a) as a proportion of the observed value using the Gaussian likelihood, (b) as a proportion of the observed value using Sivias' likelihood and (c) as a constant using the Gaussian likelihood.

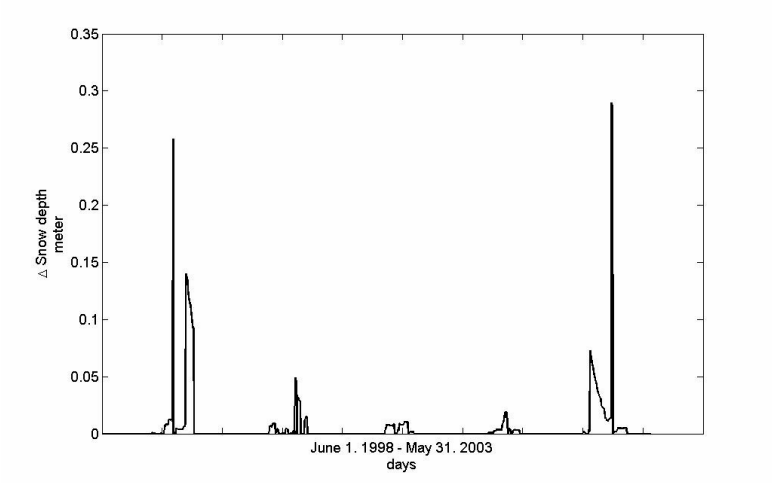


Fig. 2: The maximum difference in estimated snow depth each day for the test years, caused by using the different likelihood versions.

differences of approximately 14 cm at its most, and one period the winter 2002/2003, with differences at approximately 7 cm at its most, all with relatively fast reducing differences. The rest of the test set includes differences of less than 5 cm, and mostly much smaller.

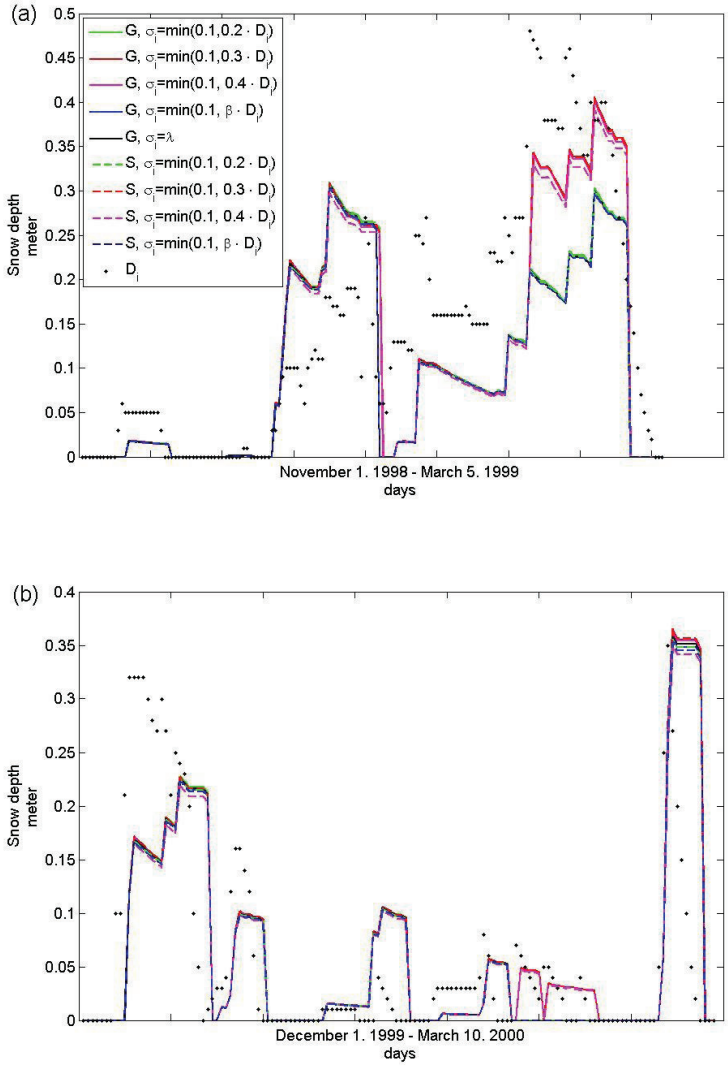


Fig. 3: Uncertainty in Snow depth estimates caused by using the different likelihood versions (a) the winter 1998/1999 and (b) the winter 1999/2000.

Fig. 3 shows estimated snow depth using the MAP parameter estimates for all nine combinations of the likelihood function and covariance matrix calibrated for the winter 1998/1999 (Fig. 3a) and 1999/2000 (Fig. 3a). The dispersion of the estimates indicates the uncertainty of estimated snow depths derived from the choice of combination of likelihood function and covariance matrix used. The predictions fitted partly well to the observations, but lots of variations in the snow depth observations were not covered by the predictions. The differences between observed and predicted snow depth was large compared to the uncertainty caused by the choice of likelihood function and covariance matrix used.

In the two winters, 1998 to 2000 (Fig. 3), a period of approximately one month the first winter and two weeks the second winter, respectively, stands out with larger dispersions. In both cases, the fixed standard deviations of 30 and 40 % of the observed value using both the Gaussian and Sivias' likelihood function, underestimated snow depth less compared to the other combinations. The first high peak in Fig. 2 of more than 25 cm was recognized in Fig. 3a as a delay in the melting process for the same combinations and additionally for Sivias' likelihood function using a standard deviation of 20 % of the observed value. A closer look at the underlying data for the winter 2002/2003 (not plotted), showed that the high peak of more than 25 cm this winter was caused by a delay of 4 days in the melting process when using Sivias' likelihood function with 40 % of the observed value as standard deviation, compared to the Gaussian likelihood function using a constant standard deviation. The other combinations gave predictions in between. Finally, the period of differences of approximately 7 cm the same winter was caused by a smaller underestimation when using Sivias' likelihood function compared to all other combinations.

The different predicted snow depths that derived from the nine combinations were assessed according to $\overline{RMSEP}$ which indicates the accuracy of model prediction, and according to $\overline{RMSD}$ which indicates the uncertainty of estimated snow depth derived from the parameters (Table 2). Analysis of variance was used in a two factor fixed effect model, to detect the significance level of the effect of likelihood function and the effect of the covariance matrix used in $\overline{RMSEP}$ and $\overline{RMSD}$. All combinations were included, except for the fixed standard deviations that were only run for the Gaussian likelihood function. In a 5 % level of significance, a significant effect of the likelihood function *($p=0.0147$)*, and a smaller insignificant effects of the covariance matrix *($p=0.0757$)* was found according to $\overline{RMSEP}$.

Table 2: Calculated root mean square error of prediction ($\overline{RMSEP}$) and root mean square deviation ($\overline{RMSD}$) for the model output, when calibrated using the different likelihood versions.

Model version	$\overline{RMSEP}$	$\overline{RMSD}$
Gaussian likelihood		
$\boldsymbol{\sigma_i = \max(0.1, 0.4 \cdot D_i)}$	0.0795	0.0118
$\boldsymbol{\sigma_i = \max(0.1, 0.3 \cdot D_i)}$	0.0798	0.0130
$\boldsymbol{\sigma_i = \max(0.1, 0.2 \cdot D_i)}$	0.0809	0.0127
$\boldsymbol{\sigma_i = \max(0.1, \beta \cdot D_i)}$	0.0807	0.0127
$\boldsymbol{\sigma_i = \lambda}$	0.0792	0.0169
Sivias' likelihood		
$\boldsymbol{\sigma_i = \max(0.1, 0.4 \cdot D_i)}$	0.0787	0.0169
$\boldsymbol{\sigma_i = \max(0.1, 0.3 \cdot D_i)}$	0.0792	0.0166
$\boldsymbol{\sigma_i = \max(0.1, 0.2 \cdot D_i)}$	0.0797	0.0160
$\boldsymbol{\sigma_i = \max(0.1, \beta \cdot D_i)}$	0.0792	0.0170

According to $\overline{RMSD}$, a higher significant effect of the likelihood function *(p=0.0054)*, and an insignificant effect the covariance matrix *(p=0.6238)* were found. While Sivias' likelihood function gave significantly smaller prediction errors than the Gaussian, it gave significantly larger uncertainty of the estimates derived from the parameters. Interaction effects were not included in the two factor model because of limitations in the degrees of freedom. By assessing plots of the factors for both responses, any interaction was found. Sivias' likelihood function with a standard deviation of 40 % of the observations gave the smallest prediction error while the Gaussian likelihood function with 40 % of the observed value gave the smallest uncertainty in the estimates derived from the parameters.

Histograms of the obtained Markov chains after burn in are plotted in Fig. 4 for two parameters (Fig. 4a for the density of fresh snow (ρ_{ns}) and Fig. 4b for the retention capacity of snow cover (SW_{ret})). The plot is three-dimensional, including all nine combinations of likelihood function and covariance matrix used. The figure shows how the posterior parameter uncertainty changes depending on the combination of likelihood function and covariance matrix used. For both parameters, the constant covariance matrix (Equation 9, which was only calculated for the Gaussian likelihood function) gave a narrower distribution with a higher peak. This indicates a smaller parameter uncertainty derived from the constant covariance matrix compared to all other combinations used.

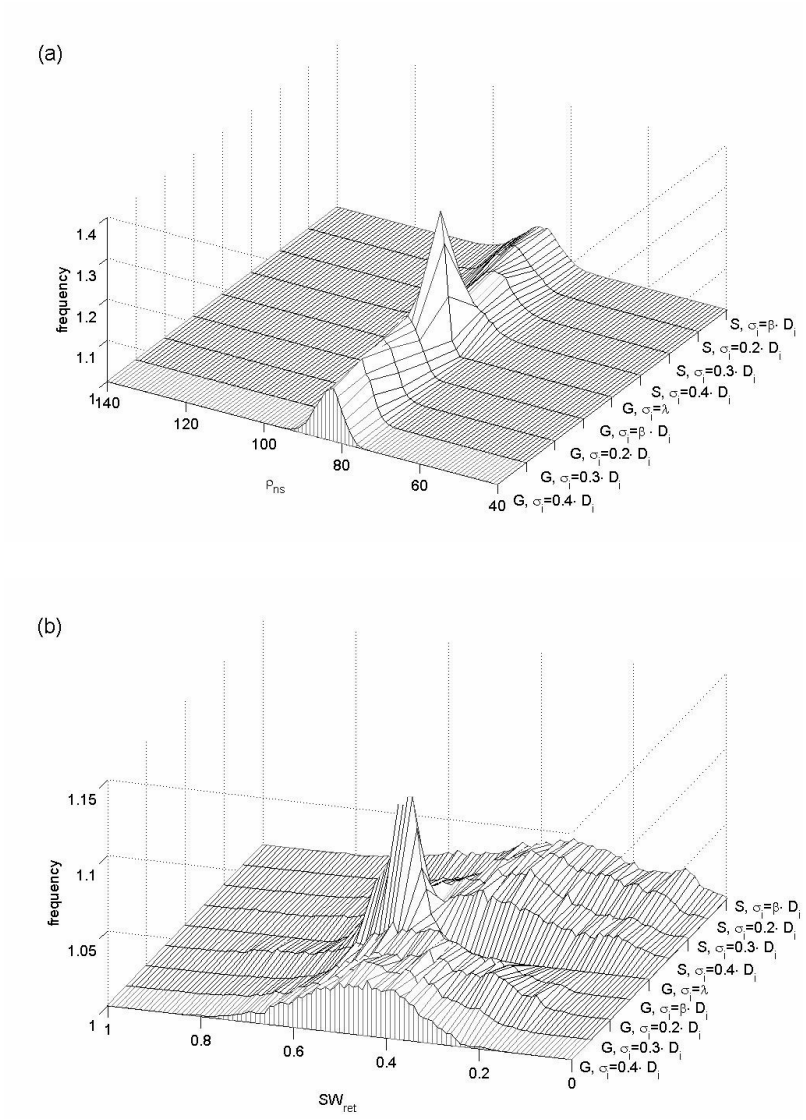


Fig. 4: Parameter uncertainty as a histogram for each likelihood version for (a) the density of fresh snow ρ_{ns} and (b) the retention capacity of snow cover SW_{ret}.

DISCUSSION AND CONCLISION

The Bayesian approach was used to calibrate a model simulating snow depths using different combinations of likelihood function and covariance matrix. The uncertainty in model output

derived from the different combinations of likelihoods was small compared to the prediction error. For the five test years, only four periods gave larger than 5 cm differences in estimated snow depth, whereas two of them were caused by a delay in the melting process for only one and four days. The prediction error was significantly smaller when using Sivias' likelihood function compared to the Gaussian, but the uncertainty in the estimates derived from parameters was significantly higher. The effect of the choice of covariance matrix was not significant for none of the assessment criteria. Over all the widest likelihood function being Sivias' likelihood function with a standard deviation of 40 % of the observations gave the best model fit to the observed data, but, on the other hand one of the largest uncertainties in the estimates derived from the parameters. The smallest uncertainty was found for the Gaussian likelihood function with a standard deviation of 40 % of the observations.

The likelihood function is a probability density function conditional on the parameters, considered as a function of the second argument with its first argument held fixed. The likelihood function was in this paper determined by the distribution of model errors. Since model error is not known accurately, only a reasonable quantification was used, where the quantification was a thought of how good we believed the model error was. The nine combinations used relied differently on the data and Sivias' likelihood with a standard deviation of 40 % of the observed value was the widest function while the Gaussian likelihood with a constant standard deviation was the narrowest.

The searched posterior distribution is a combination of prior knowledge and new incorporated information through the likelihood function. Prior information will dominate if the size of the collected data is small or if the data only provide indirect information about the parameters. Since the likelihood function is not known, our quantification may be either too certain (when using a too narrow probability distribution) or too uncertain (when using a too wide probability distribution). When using a too widely distributed likelihood function compared to the actual distribution of model error, details in the data will disappear and the model will only be fitted to the main structure in the data. On the opposite, by using a too narrow likelihood distribution, sporadic variation in the data (representing diversity within the population) will be treated as detailed information.

The smallest prediction error was found for the widest likelihood (Sivias' likelihood with a standard deviation of 40 % of the observed value) used. Sivias' likelihood (which is a wider probability distribution than the Gaussian), gave smaller prediction error than the Gaussian.

For the fixed standard deviations at 20, 30 and 40 % of the observed value, the prediction error decreased with an increased standard deviation. This shows that the prediction error is smallest when looking at only the main structure in the data, not focusing on details. This is reasonable since the measurements are point values, while the model predicts average snow depth in a large homogeneous field. The second smallest prediction error was found for Sivias' likelihood both when estimating the standard deviation as percent of the observed values and fixed as 30 % of the observed value, and for the Gaussian when estimating the standard deviation as a constant. The latter choice was the narrowest likelihood version used, but also the only one not assuming the standard deviation to be a percent of the observations. It is reasonable to believe that the standard deviation increases as the observed snow depth measurements increases, but maybe the increase is not linear with the measurements.

The Gaussian likelihood function gave significantly smaller uncertainty compared to Sivias' likelihood, but the effect of the choice of the standard deviation was insignificant. When fixing a parameter that is not known accurately, the parameter uncertainty will be underestimated (Gelman et al. 1996a). The quantified output uncertainty derived from the parameters ($\overline{RMSD}$) calculated from the fixed and the estimated standard deviations are therefore not comparable. Among the fixed standard deviations, $\overline{RMSD}$ seemed to increase with an increased standard deviation, except for the Gaussian likelihood with a standard deviation of 40 % of the observed value, which happened to have the all over smallest $\overline{RMSD}$.

The structure of the likelihood including its covariance matrix had an impact on the length of the trial and error period to obtain an acceptable proposal distribution in order to reach convergence for the chains. By increasing the dimension of the parameter space, convergence was harder to reach. Caused by the difficulty of reaching convergence within a limited amount of time, the posterior distribution was not found for the combination of Sivias' likelihood and the constant standard deviation. Also, we experienced that the wider likelihood functions needed shorter periods of trial and error. One explanation of this can be the fact that the data were very uncertain or that the details in the data were variations that we would not fit the model to.

In this study we have seen that the choice of likelihood did affect the model results, but to a very small extent that much. The smallest prediction error obtained was 2.7 % smaller than the highest. All combinations of likelihood function and covariance matrix gave results close to each other and in both longer periods with higher differences the wider likelihood versions

underestimated least. The effect of the covariance matrix was not significant. When estimating the covariance matrix, convergence of the Markov chains were harder to reach and longer periods of trial and error was needed. At the same time, the effect on the prediction error and the uncertainty in the estimates caused by the parameters were small. Sivias' likelihood gave significant improvements in $\overline{RMSEP}$ and was therefore preferred compared to the Gaussian in this case, although it also increased $\overline{RMSD}$ significantly.

Based on this study we would prefer using wide likelihood functions with wide fixed standard deviations instead of estimating it. But it is important to notice that this was only based on one case and similar studies should have been repeated on different cases to generalize the conclusion.

REFRENCES

Berger, O. J., 1985. Statistical Decision Theory and Bayesian Analysis. New York: Springer-Verlag

Engseth, R. V., Sorteberg, H. K., Udnæs, H., 2000. NOSiT, utvikling av NVEs operasjonelle snøinformasjonstjeneste. Norges vassdrags- og energidirektorat.

Gelman A. (2002) Prior distribution. Encyclopedia of Environmetrics 3:1634-1637

Gelman, A., Rubin, D. B., 1992. Inference from Iterative Simulation Using Multiple Sequences. Statistical Science 7 457-511

Gilks, W. R., Richardson, S., Spiegelhalter, D. J., 1996. Markov chain Monte Carlo in practice. London: Chapman and Hall

Goldstein, M., Rougier, J., 2006. Bayes Linear Calibrated Prediction for Complex Systems. Journal f the Americal Statistical Association 101 1131-1143

Hassan, A. E., Bekit, H. M., Chapman, J. B., 2009. Using Marcov Chain Monte Carlo to quantify parameter uncertainty and its effect on predictions of a groundwater flow model. Environmental Modelling & Software 24 749-763

Hastie, T., Tibshirani, R., Friedman, J. (2001). The elements of Statistical Learning. Data Mining. Inference, and Prediction. Springer

Hue, C., Tremblay, M., Wallach, D. J., 2001. A Bayesian Approach to Crop Model Calibration Under Unknown Covariance. Journal of Agricultural, Biological, and Environmental Statistics 13 355-365

Iizumi, T., Yokozawa, M., Nishimori, M., 2009. Parameter estimation and uncertainty analysis of a large scale crop model for paddy rice: Application of a Bayesian approach. Agricultural and Forest Meteorology 149 333-348

Jeffreys H. (1961) Theory of probability. Oxford University Press

Kennedy, M. C., O'Hagan, A., 2001. Bayesian Calibration for Computer Models. Journal of the Royal Statistical Society 63 425-464

Liu JS. (2001), Monte Carlo Strategies in Scientific Computing. Springer.

Marshall, L., Nott, D., Sharma, A., 2004. A comparative study of Markov Chain Monte Carlo methods for conceptuall rainfall-runoff modeling. Water Resource Research. 40

Miller, I., Miller, M., 1999. Freund's Mathematical Statistics. Prentice-Hall, Upper Saddle River.

Montgomery D. C. 2005 Design and Analysis of Experiments. Wiley International Edition

Rougier, J., 2007. Probabilistic Inference for Future Climate Using an Ensemble of Climate Model Evaluations. Climate Change 81 247-264

Sivia, D. S., 2006. Data Analysis, a Bayesian Tutorial. Oxford University press

Thorsen, S. M., Haugen, L. E., 2007. Development of the SnowFrost modell for the simulation of Snow Fall and Soil Frost. Bioforsk FOKUS 2 1-23

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

Van Oijen, M., Höglind, M., Hanslin, H. M., Caldwell, N., 2005b. Process-Based Modelling of Thimothy Regrowth. Agronomy Journal 97 1295-1303