Sensitivity analysis and Bayesian calibration for testing robustness of the BASGRA model in different environments

Anne-Grete Roer Hjelkrem[*a], Mats Höglind[a], Marcel van Oijen[c], Jürgen Schellberg[b], Thomas Gaiser[b] and Frank Ewert[b]

[a] Norwegian Institute of Bioeconomy Research (NIBIO), 1431 Ås, Norway

[b] Institute of Crop Science and Resource Conservation, University of Bonn, Katsenburgerweg 5, D-53115 Bonn, Germany

[c] Centre for Ecology & Hydrology, CEH-Edinburgh. Bush Estate, Penicuik EH26 0QB, United Kingdom

*Corresponding author. Tel.: +47 971 18 993

Email address: anne-grete.hjelkrem@nibio.no, (Anne-Grete Roer Hjelkrem)

Abstract

Proper parameterisation and quantification of model uncertainty are two essential tasks in improvement and assessment of model performance. Bayesian calibration is a method that combines both tasks by quantifying probability distributions for model parameters and outputs. However, the method is rarely applied to complex models because of its high computational demand when used with high-dimensional parameter spaces. We therefore combined Bayesian calibration with sensitivity analysis, using the screening method by Morris (1991), in order to reduce model complexity by fixing parameters to which model output was only weakly sensitive to a nominal value. Further, the effect on the error term and the parametric uncertainty when fixing parameters were investigated in order to achieve a robust model. The process-based grassland model BASGRA was examined in the present study on two sites in Norway and in Germany, for two grass species (*Phleum pratense* and *Arrhenatherum elatius*). According to this study, a reduction of free model parameters from 66 to 45 was possible. The sensitivity analysis showed that the parameters to be fixed were

consistent across sites (which differed in climate and soil conditions), while model calibration had to be performed separately for each combination of site and species. The output uncertainty decreased slightly, but still covered the field observations of aboveground biomass. A detailed analysis of the mean square error was included, and the error term for both the 66 and the 45 parameter model was dominated by errors in timing (phase shift) when considering the training data, whereas no general pattern was found in errors when using the validation data. Stronger model reduction should be avoided, as the error term increased and output uncertainty was underestimated.

Key words: Metropolis-Hasting, Morris method, reducing complexity, robustness

1 INTRODUCTION

Grassland covers about 70% of the world's agricultural area (FAO). It has a central role in feeding ruminants and other herbivores, and the growing demand for meat may induce an even more intensive use in the future.

Complex dynamic growth models are increasingly used to simulate the interactions between vegetation and environment. Such models are useful in order to forecast yield, study the effect of climate change on yield, optimize management and to better understand the system. It is common to apply the same model in different regions and for different species and cultivars, and it should work well in all the situations for which it is applied. This requires that it is properly parameterised, and that parameters and output uncertainty are well quantified. Among parameter estimation methods, Bayesian calibration (Berger, 1985) has the advantage that it, in addition to calibrating the parameter values, simultaneously quantifies parameter uncertainty (Campbell, 2006). It achieves this by calculating posterior parameter distributions as a function of the original parameter uncertainty (prior knowledge) and new information incorporated through the conditional probability distribution of the collected data (likelihood function). The method is still rarely used for complex models, but its application has been increasing in recent years (Gouache et al., 2013; Minunno et al., 2013; Thorsen and Höglind, 2010; van Oijen et al., 2005a,b; Kennedy and O'Hagan, 2001).

To estimate all the parameters of complex, parameter rich models simultaneously is often challenging. A major problem is the large computational effort required to investigate a high

dimensional parameter space. As a result, predictive performance may be poor suggesting a need for model simplification (Cox et al. 2006). A study by Crout et al. (2014) identified several redundant variables in the Sirius wheat model. Here we focus on a different form of model simplification: reducing the number of free parameters in the model. Sensitivity analysis, or parameter screening, is a useful tool for model reduction that can make it easier and less time requiring to parameterise models by detecting the least sensitive parameters. These are parameters that can be fixed within their prior parameter boundaries without strongly affecting model robustness. Robustness is here referred to as the extent the model results are affected with when reducing the number of free parameters, where model results include the uncertainty in model outputs caused by parameter uncertainty. A simplification of a model by fixing the poorly sensitive parameters to nominal values will increase the efficiency of model calibration, but also result in underestimation of parameter uncertainty, since the parameter values that are fixed are not known for certain. A combination of sensitivity analysis and Bayesian calibration of a complex model was given by Raj et al. (2016), whereas the effect of model reduction on model uncertainty was not covered.

Study of the mismatch (error term) between observed and simulated model output is a widely used procedure for model evaluation. A detailed analysis of the error term, decomposing it into the three components of bias, variance error and phase shift, was proposed by Kabayashi and Salam (2000). Their method is still rarely used (but see van Oijen et al. 2011; Ewert et al. 2002), yet it adds valuable information about model behaviour.

The process-based BASGRA (BASic GRAssland) model is used in this study. It is a model that simulates growth of *Phleum pratense* (L.) (Höglind et al., 2001; Thorsen and Höglind, 2010; Thorsen et al., 2010; van Oijen et al., 2005a). BASGRA contains 66 parameters and is driven by the environmental variables air temperature, precipitation, relative humidity, global radiation and wind-speed at a daily resolution. It calculates 23 state variables of which 13 quantify the state of the plant and 10 represent the above- and belowground environment. Only one output variable, aboveground biomass, is the focus of this study. This is one of the most often measured variables in grassland research.

The general objective of this study was to examine the robustness of aboveground biomass predictions by the grassland model BASGRA. The impact of parameter screening and subsequent parameter reduction on aboveground biomass predictions were quantified in order to allow efficient quantification of output uncertainty. The specific objective of this study was to identify a minimum number of parameters required for the BASGRA model in order to

estimate both the value of aboveground biomass and its uncertainty with sufficient accuracy, consistent between sites and species.

Four sets of data were used: (1) total aboveground biomass of *Phleum pratense (P. pratense*) grown at Særheim, Norway, observed at intervals of 1-2 weeks throughout the growing season including at the agricultural harvests, (2) observations (two per year) of biomass yield from the same experiment, (3) observations (three per year) of biomass yield from a mixed sward dominated by *P. pratense* grown at Rengen, Germany and (4) observations (two per year) of biomass yield from a mixed sward dominated by *Arrhenatherum elatius (A. elatius*) grown at Rengen, Germany. Model performance had been tested thoroughly for *P. pratense* growth at Særheim by (van Oijen et al., 2005a) and the full dataset of that study was used here for model training. The datasets from Rengen were further split up into one training and one test dataset.

2 MATERIAL AND METHODS

2.1 Grassland Growth Model

The BASGRA (BASic GRAssland) model simulates the growth of grassland swards for any period of time (a short growing cycle, a sequence of growing cycles, a winter period, a sequence of whole years etc.). The model is based on the LINGRA model (Schapendonk et al., 1998), but differs in that it simulates the dynamics of both vegetative and reproductive tillers (Höglind et al., 2001; van Oijen et al., 2005a) and that it includes processes which occur during winter (Thorsen and Höglind, 2010; Thorsen et al., 2010), thus allowing for whole year simulations. The model was constructed with the aim to make it widely applicable by simulating the impact of a wide range of environmental drivers and with the intention to represent processes in a simple, yet realistic way.

The model considers the effects of weather, soil type (water storage capacity) and grassland management (timing and frequency of harvest). It calculates 35 different output variables, including 23 state variables (13 for the state of the plant and 10 for the above- and belowground environment). Only one output variable, aboveground biomass, is focused on in this study. The model is parameter rich, containing 62 parameters (51 plant specific and 11 site specific), and it requires time series of daily weather data (air temperature, precipitation, relative humidity, global radiation and wind speed).

BASGRA was originally adapted for simulating *P. pratense* growth, but was in this study additionally used for simulation of *A. elatius* dominated species-rich grassland plots. The BASGRA model was therefore generalised in the present study by including four additional parameters. Specifically, the changes dealt with the linear equations for the elongation rate of leaves on vegetative tillers (LERV) and for leaf elongation rate per leaf of reproductive tillers (LERG). In both equations, constants for the y-intercept (denoted by a) and the slope (denoted by b) were replaced by unknown parameters (LERVa, LERVb, LERGa and LERGb). The generalised BASGRA model that was used in this study contains therefore a total of 66 unknown parameters.

In addition to unknown parameters, the BASGRA model contains 14 fixed values of which seven are considered as known and universal and the remaining seven are known site specific values, including latitude and constants for soil properties (Table S1). Both the fixed values (Table S1) and the nominal values for the parameters (Table S2) were derived from earlier literature studies (Höglind et al., 2001; Thorsen and Höglind, 2010; van Oijen et al., 2005a), whereas the site specific values for Rengen and the nominal plant specific values for *A. elatius* were obtained in the present study.

The simulations were initiated in the year of establishment, except for the long-term experiment in Rengen for which the simulations were initiated in the autumn of the year prior to first harvest included in the analysis, and the model was in each case run for multiple years.

BASGRA is implemented in FORTRAN and simulations are run from script-files in R. The most recent model version, BASGRA 2014, can be downloaded from the internet together with a user manual (Van Oijen et al., 2015). This is a slightly updated version of BASGRA 2012 which was used in the present study. The major difference is that BASGRA 2014 distinguishes three tiller categories instead of the two distinguished in BASGRA 2012.

2.2 Field data

Data from three different field experiments were used, including two sites (Særheim in Norway and Rengen in Germany) and two grass species (*P. pratense* and *A. elatius*).

The first experiment was conducted at Særheim Research Centre at Klepp, about 8 km from the coast (58º46’N lat; 5º38’E long; 90 m above sea level) in Southwestern Norway. The composition of the soil is 60% sand, 27% silt, 6% clay, and 7% organic matter. The mean

annual temperature is 7.1 ºC and the mean annual precipitation is 1280 mm (1961-1990). The experiment was carried out for *P. pratense*, the most widely grown forage grass species in Scandinavia, with the cultivar Grindstad, which is the most commonly grown timothy cultivar in Norway. The data were collected from two different fields, established in 1999 (measurements for 2000) and 2000 (measurements for 2001 and 2002). The full dataset includes measurements at intervals of one to two weeks of a large number of variables during the first and second regrowth cycles in 2000–2002 (Höglind et al., 2005). Only the total aboveground biomass data from these experiments were used in the present study.

The other two experiments were conducted at Rengen Grassland Farm of the University of Bonn, which is located in the Eifel Mountains, about 60 km west of the Rhine (50º13'N lat; 6º51'E long; 490 m above sea level) in West Germany. The soil is an intermittently wet Pseudogley (Stagnic Luvisol). The mean annual temperature is 6.9 ºC and the mean annual precipitation 811 mm. The first experiment at Rengen consisted of *P. pratense* dominated grassland that has been established in 1988. Data of biomass yield were collected between 1989 and 1994, with three harvests each year. The 1989 to 1991 data from this experiment were used as training data for sensitivity analysis and model calibration, while the remaining data from 1992-1994 were used as an independent test data set for model validation. The second experiment conducted at Rengen was established on an extensively grazed heathland in 1941 (Chytrý et al., 2009; Schellberg et al., 1999), naturally dominated by *Calluna vulgaris* L. and *Nardus stricta* L. In 1941, the turf layer was grubbed and reseeded with a grass/legume mixture. From this long-term experiment, we extracted data from one fertilizer treatment (Ca/N/P2O5/KCl) in the years 2000-2005. Long-term data up to 2014 on floristic composition show that the sward in this particular treatment is now dominated by *A. elatius*. Data of biomass yield, with two harvests each year, was collected and used in the present study. The data collected between 2000 and 2002 were used as training data, while the remaining data from 2003 to 2005 were used for model validation.

2.3 Weather data

Weather data were automatically collected from on-site weather stations, provided by Agrometeorology Norway (Agrometeorology Norway, 2015) and Rengen meteorological station. At both stations, the daily weather records included air temperature (ºC), precipitation (mm) and relative humidity (%). Wind speed (m/s) and global radiation (W/m^2) was

191 additionally recorded at Særheim. At Rengen, wind speed data were not available and
192 averaged data over all Germany was used instead of local data, while global radiation was
193 estimated according to Angstrom (Angstrom, 1924), based on observed sunshine hours (h) at
194 Rengen.
195

196 **2.4 Sensitivity Analysis**

197 Sensitivity analysis determines the parameters that are the key drivers of a model, by
198 investigating to what extent the variation in model output is influenced by different sources of
199 variation in the model parameters (Saltelli et al., 2004). It is a suitable tool for model
200 simplification in that the parameters that are detected to have minor impact on model output
201 can be fixed to a nominal value. The sensitivity method introduced by Morris (Morris, 1991)
202 is a screening method that is suitable for complex models where the number of parameters or
203 the computational cost limit the possibility of numerical calculation.

204 In the screening method by Morris, the parameter space is defined by a *p*-level grid within the
205 parameter boundaries, and the parameter θ_i, where i=1,…,k, is assumed to vary across the p
206 selected levels. Elementary effects (EE_i) of the model output are calculated from two
207 consecutive model runs according to Equation 1.

208 $$EE_i(\boldsymbol{\theta}) = \left(\frac{y(\theta_1, \dots, \theta_{i-1}, \theta_i + \Delta, \theta_{i+1}, \dots, \theta_k) - y(\boldsymbol{\theta})}{\Delta}\right) \quad (1)$$

209 Here, Δ is in the range of [1/(p-1), 1-1/(p-1)], p is the number of levels, $\boldsymbol{\theta}$ is any selected
210 parameter vector in the parameter space such that the transformed point ($\boldsymbol{\theta}$+$\mathbf{e}_i\Delta$) remains
211 within the parameter space for each index i=1,2,…,k and $\mathbf{e}_i$ is a vector of zeros with a unit
212 corresponding to its i'th component.

213 The finite distribution (F_i) of elementary effects (EE_i), denoted $EE_i(\boldsymbol{\theta})\sim F_i$, is constructed by r
214 elementary effects that are sampled using an efficient design that constructs r trajectories of
215 (k+1) points in the parameter space. Two sensitivity measures can then be calculated from
216 $\boldsymbol{EE}$: (1) $\boldsymbol{\mu}$ (the mean value), which evaluates the overall influence of the parameters on model
217 output, and (2) $\boldsymbol{\sigma}$ (the standard deviation), which is used to detect parameters involved in
218 interaction with other parameters or whose effect is nonlinear. To avoid the problem of effects
219 of opposite signs which occur when the model is non-monotonic, we will in this study use $\boldsymbol{\mu}^*$
220 (the mean of the absolute value of $\boldsymbol{EE}$) that was introduced by Campolongo et al. (2007).

For dynamic models that simulate daily outputs, the sensitivity of model parameters may change with time. It is consequently most appropriate to consider the outputs over the whole time series (Lamboni et al., 2009), but the large number of responses that need to be evaluated makes this approach challenging for parameter rich models. In this study, the total aboveground biomass over all harvests was selected as the response.

In this study, the screening method of Morris was first applied to the dataset from Særheim 2000-2002, evaluating the total aboveground biomass summed over the individual harvests of *P. pratense*. Secondly, the method was applied to the dataset from Rengen 2000-2002, evaluating the total aboveground biomass summed over the individual harvest dates of *A. elatius*. The ranking order of the parameters with respect to sensitivity was determined, and groups consisting of the 45, 9 and 4 most sensitive parameters were defined.

2.5 Bayesian calibration

The Bayesian framework is based on Bayes theorem (Berger, 1985) and is given in Equation 2.

$$\pi(\boldsymbol{\theta}|\boldsymbol{D}) = \frac{\pi(\boldsymbol{\theta}) \cdot f(\boldsymbol{D}|\boldsymbol{\theta})}{f(\boldsymbol{D})} \propto \pi(\boldsymbol{\theta}) \cdot f(\boldsymbol{D}|\boldsymbol{\theta}) \qquad (2)$$

Here, $\boldsymbol{\theta}$ is the vector of the model parameters and $\boldsymbol{D}$ is the observed data. The resulting posterior parameter distribution ($\pi(\boldsymbol{\theta}|\boldsymbol{D})$) is the probability distribution for the parameters conditional on the data, determined as a combination of our prior knowledge of the parameters before new data are included ($\pi(\boldsymbol{\theta})$, the prior parameter distribution) and the distribution of the new data conditional on model parameterisation ($f(\boldsymbol{D}|\boldsymbol{\theta})$, likelihood function). The integrated likelihood ($f(\boldsymbol{D})$) is the marginal probability of the data, which is a constant. With only few experimental data, the prior parameter distribution will highly affect the posterior probability distribution, but the more such data are added to the calibration, the smaller will be the impact of the prior parameter distribution.

Integration problems make exact calculations impossible when the parameter space is highly dimensional. In this study, calculations were done using the Markov chain Monte Carlo (MCMC) algorithm Random walk Metropolis (Liu, 2001). The prior probability distributions were described by beta distributions with minimum, maximum and nominal value given in Table S2. Prior independence was assumed, and the joint distribution was thus determined as

the product of the marginal parameter distributions. The likelihood function was determined by the distribution of measurement error, following van Oijen et al. (2005b). As specific information of the precision of the measurements was not available, the standard deviation of each measurement was set to 5% of its observed value. The model was calibrated separately to each of the four datasets described above, i.e. two datasets from Særheim and two datasets from Rengen. For each of the four datasets, the full model as well as the reduced models consisting of the 45, 9 and 4 most sensitive parameters were calibrated. The non-calibrated parameters in the reduced models were fixed to their nominal values.

2.6 Model fit and validation

Model performance was evaluated on the basis of the root mean square error for the mismatch between simulated and observed biomass yields normalised by the mean of the observed aboveground biomass (NRMSE). In addition, the mean square error (MSE) for the mismatch between simulated and observed biomass yield was calculated and decomposed into three components (Kobayashi and Salam 2000), given in Equation 3.

$$MSE = (\bar{\boldsymbol{M}} - \bar{\boldsymbol{D}})^2 + (SD_M - SD_D)^2 + 2SD_M SD_D(1 - r) \qquad (3)$$

Here, $\boldsymbol{M}$ is the vector of model simulations, $\boldsymbol{D}$ is the vector of observed data, SD_M and SD_D are the standard deviation of respectively model simulations and observed data, while r is the correlation between them. The three components of the right-hand side of Equation 3 are the squared bias (henceforth referred to as 'bias'), squared difference between the standard deviations ('variance error') and lack of correlation weighted by the standard deviations ('phase shift') (Kobayashi and Salam 2000).

The error terms were calculated for the full model as well as for the reduced models consisting of the 45, 9, and 4 most sensitive parameters. For each model, it was calculated for all four training datasets (the data used for sensitivity analysis and Bayesian calibration), in order to show the effect of model reduction on how well the simulations fitted to the observations of the training data. Additionally, NRMSE was calculated for the separate test datasets from Rengen in order to validate the model's ability to make predictions. In this case, the normalised root mean square error of prediction (NRMSEP) was calculated.

Model uncertainty was calculated daily as minimum and maximum predicted aboveground biomass from 100,000 samples, sampled randomly from the posterior distributions.

Additionally, uncertainty was calculated as the posterior coefficient of variation (CV) based on summed aboveground harvests over 100,000 samples, sampled randomly from the posterior distributions. The CV is a normalised measure of discrepancy of the probability distribution defined as the ratio of the standard deviation to the mean.

3 RESULTS

3.1 Sensitivity Analysis

The sensitivity analysis explored the space within the prior parameter boundaries (Table S2), and was performed using the Morris method with 2000 trajectories and 4 levels. It was applied separately to the dataset at Særheim and Rengen.

3.1.1 Dataset from Særheim, 2000-2002

The sensitivity analysis was run for the BASGRA model, using the 2000-2002 weather data from Særheim and site specific harvest dates. The summary statistics of the elementary effects of each parameter were calculated and plotted in Figure 1a. The points in the upper right corner, with both high μ^* and σ, indicate parameters to which the model is highly sensitive.

[FIGURE 1]

Twenty-one parameters stood clearly out as the least sensitive ones according to aboveground biomass (Figure 1a, points in the lower left corner). These poorly sensitive parameters consisted of 10 plant specific (LERGa, RRDMAX, LOG10CRTI, RATEDMX, LDT50A, LDT50B, KRDRANAER, TRANCO, HAGERE and CLAIV) and 11 site specific (FGAS, FO2MX, gamma, KRTOTAER, KSNOW, LAMBDAsoil, RHOnewSnow, RHOpack, Swret, SWrf and TrainSnow) parameters. By fixing the poorly sensitive parameters identified above to their nominal values (Table S2), a reduced version of the BASGRA model was constructed with 45 (plant specific) parameters.

The reduced BASGRA model with 45 parameters is still parameter rich, and two even simpler models were constructed by fixing all parameters except for the nine and four most sensitive

ones to their nominal values. The four parameters to which the model was most sensitive (Figure 1a) were a constant in the logistic curve for frost survival (KRSR3H), day length below which DAYLGE (day length effect on allocation, tillering, leaf appearance, leaf elongation) becomes less than 1 (DLMXGE), the initial and maximum value of rooting depth (ROOTDM) and day length below which phenological stage is reset to zero (DAYLB), where KRSR3H was by far the most sensitive. The group of the nine most sensitive parameters additionally included maximum SLA of new leaves (SLAMAX), day length below which phenological development slows down (DAYLP), the minimum SLA of new leaves as a fraction of maximum possible SLA (FSLAMIN), the maximum ratio of tiller and leaf appearance at low leaf area index (LAITIL) and the rate of elongation of leaves on non-elongating tillers (LERVb).

3.1.2 Dataset from Rengen, 2000-2002

Summary statistics from the sensitivity analysis using the dataset from Rengen 2000-2002 with site specific harvest dates are plotted in Figure 1b. Exactly the same parameters were detected in the groups of the four and nine most sensitive parameters as when using the dataset from Særheim. Also, the group of the 45 most sensitive parameters were quite similar, with the exception of three plant specific parameters. These parameters were: (1) the slope of linear dependence of duration of anaerobic conditions at which death rate is half the maximum and the temperature that kills half the plants in a day (LDT50B), (2) the maximum relative death rate due to anaerobic conditions (KRDRANAER) and (3) the maximum leaf area index remaining after harvest, when no tillers elongate (CLAIV) that were included for Særheim. Oppositely, the parameters: (1) log10 of initial value of reserves (LOG10CRESI), (2) phenological stage above which elongation and appearance of leaves on elongation tillers decreases (PHENCR) and (3) maximum relative death rate of leaves and non elongating tillers due to shading (RDRSMX) stood out as sensitive at Rengen.

3.2 Bayesian Calibration

Bayesian calibration was performed for the model with the full parameter set and the reduced parameter sets of 45, 9 and 4 parameters. Two Markov chains were run in parallel

for 500,000 iterations and convergence occurred within the first 100,000 iterations for all cases.

Point estimates were calculated from the Markov chains of the posterior probability distributions, as maximum *a posteriori* (MAP) estimates, and given in Table 1, for both the 4 and the 45 parameter sets. Many estimates differed strongly from the nominal values for both the 4 and 45 parameter sets. Only two MAP values were similar to the nominal value; this was the case for the maximum surface temperature at which hardening is possible (THARDMX) and LUE-increase with increasing fraction of elongating tillers (KLUETILG) for the field data of *P. pratense* grass growth at Særheim, respectively, using the complete dataset and the harvest data only. The largest difference was found for the common logarithm of the initial value of reserves (LOG10CRESI) that was reduced by 213% when the field data of *P. pratense* dominated grass growth at Rengen was used.

The most sensitive parameter according to the sensitivity analysis, KRSR3H, was in all cases (Table 1) found to have been overestimated in the prior given the lower MAP values for this parameter compared to its nominal value. When field data of *P. pratense* grass growth were used, the value of the parameter was reduced by 18% or less, with the highest reduction for the field data from Rengen. A much higher decrease was estimated for *A. elatius* dominated grass growth data from Rengen, with a 58% reduction in the 4 parameter set. According to the 45 parameter set, smaller decreases were found for all cases, with a maximum decrease of 15% at Særheim, using only the harvest observations of *P. pratense* grass growth.

Also the value of DLMXGE was reduced after model calibration. For the 4 parameter set, the highest reduction (59%) was detected when the field data of *A. elatius* dominated grass growth at Rengen were used, while also the two datasets from Særheim exhibited a high reduction (30 and 31%). According to the 45 parameter sets, a 35% decrease was found for the complete dataset at Særheim, while only smaller reductions (<18%) were detected for the other datasets.

ROOTDM increased by about 40% for both the complete dataset of *P. pratense* grass growth data from Særheim and for the harvest observations of *A. elatius* dominated grass growth data from Rengen, while a decrease of 24 and 18% was detected for the *P. pratense* aboveground biomass from both Særheim and Rengen, respectively, when the 4 parameter set was calibrated. According to the 45 parameter set, opposite results were detected, with a decreased value for the complete *P. pratense* grass growth dataset from Særheim and the *A. elatius*

dominated grass growth data from Rengen of respectively 21 and 70% and an increase for *P. pratense* dominated grass growth at Rengen of 18%.

For DAYLB, generally increased values appeared after calibration, except for a reduction of 36% for *P. pratense* grass growth data using only harvest observations at Særheim and a reduction of 79% for *A. elatius* dominated grass growth according to the 4 parameter set.

[TABLE 1]

3.3 Model outputs and validation

3.3.1. Training dataset

Model outputs were calculated for the four cases of field data and the four parameter sets, using the MAP parameter estimates. The NRMSE and MSE (Table 2) and the percentage decomposition of the MSE (Figure 2) were calculated for the training data to identify the model's ability to adapt to the underlying structure in the data. As a mean value over the four different cases of field data, the 66 parameter set had the lowest NRMSE with only 0.08, whereas the 45 parameter set gave only slightly poorer fit (NRMSE = 0.09) (Table 2). For all four different cases of field data, the decomposition of MSE for both the 45 and 66 parameter set (Figure 2a-d) were dominated by the phase shift component (more than 64-100% of the total MSE), followed by the bias (2-32%) and the variance error (< 10%). Individually, the two sets of parameters (45 and 66) gave best fit for two cases of field data each. While the 45 parameter set gave the overall best fit for the harvest observations of *P. pratense* dominated grass growth data from Rengen and the *A.elatius* dominated grass growth data from Rengen, the 66 parameter set gave best fit for the complete set of field data of *P. pratense* grass growth at Særheim and for the *P. pratense* grass growth field data with only harvest observations from Særheim. The 4 and 9 parameter sets gave the worst and the second worst fit for all cases according to NRMSE (Table 2). The decomposition of MSE (Figure 2a-d) showed high variability between the cases of field data. Both the *P. pratense* grass growth field data with only harvest observations from Særheim and the *A. elatius* dominated grass growth data from Rengen were dominated by the bias (73-91%) component for both the 4 and 9 parameter sets. For the complete set of field data of *P. pratense* grass growth at Særheim with 9 parameters, the phase shift component

dominated, whereas the effect was more equally spread out between the components for the remaining cases.

[TABLE 2]

[FIGURE 2]

Robustness of predicted aboveground biomass was evaluated for the four parameter sets at Særheim and Rengen. Model error (NRMSE) was used as a measure of model fit while the posterior coefficient of variation (CV) summed over the harvests, was used as a measure of model uncertainty. In Figure 3a and b, NRMSE and CV are plotted as functions of the fraction of parameters determined in the calibration (0.06 (4 parameters), 0.14 (9 parameters), 0.68 (45 parameters) and 1 (66 parameters)). For both Særheim (Figure 3a) and Rengen (Figure 3b), model discrepancy (NRMSE) decreases clearly when increasing the faction of parameters from 0.06 toward 0.14 and to 0.68, whereas no improvement was detected when increasing the fraction of parameters from 0.68 to 1. As model discrepancy decreases, model output uncertainty (CV) increases with the fraction of parameters determined in the calibration. Clearly, a higher increase was determined until the fraction of parameters was 0.68, whereas no increase was detected when increasing the fraction of parameters from 0.68 to 1. In Figure 3c and d, NRMSE and CV are plotted as functions of the highest normalised μ^* among the parameters left out from the calibration (3.2 (4 parameter set), 2.1 (9 parameter set), 0.2 (45 parameter set), 0 (66 parameter set)). For both Særheim (Figure 3c) and Rengen (Figure 3d), model discrepancy (NRMSE) increases and model uncertainty (CV) decreased clearly when increasing the highest normalised μ^* among the parameters left out from the calibration, but for parameters having a lower normalised μ^* than 0.2, no or smaller effects were detected on model discrepancy and model uncertainty.

[Figure 3]

In Figure 4, model outputs calculated from the estimated MAP values of the 45 parameter model and the 4 parameter model are plotted together with the training data. For *P. pratense*,

grass growth at Særheim in 2000 (Figure 4a), an almost perfect fit was found for all the field observations when the 45 parameter set was used. For the 4 parameter model, on the other hand, a poor estimation was found, with much lower estimations compared to the observations.

Also for 2001 and 2002 (Figure 4b), the 45 parameter model estimated aboveground biomass adequately. For the first cut in 2001, the estimated aboveground biomass decreased slightly before the harvest, whereas for the second cut in 2002 an underestimation occurred. For the 4 parameter set model, large underestimations appeared for all field observations, which is in line with the results from 2000 (Figure 4a). However, the second year actually estimated the first part of both the first and second regrowth well, but the growing stopped too early and caused an underestimation in aboveground biomass for the last part of both re-growing periods in the second year.

For the *P. pratense* swards at Særheim, only including harvest observations, the 45 parameter model fitted perfectly to the observations in both the years 2000 (Figure 4c) and 2001-2002 (Figure 4d). The 4 parameter model underestimated aboveground biomass for all the observations. The model fitted to the *P. pratense* dominated grass growth in Rengen (a three cut system) provided a quite nice fit with the 45 parameter model (Figure 4e). For the first and third cut of the first year, the second cut in the second year and the second cut in the third year, the estimated aboveground biomass decreased slightly before the harvest. The remaining observations fitted well to the observations, except for the third cut in the second year, where almost no regrowth was estimated, thus causing strong underestimation of aboveground biomass. The 4 parameter model generally underestimated the observed aboveground biomass. For the first year, the observed aboveground biomass was low, and fitted quite well to the estimated results, as was also the case for the last cut in the third year.

All other observations were highly underestimated. For *A. elatius* dominated grass growth in Rengen (Figure 4f), the fit between model outputs from the 45 parameter model and observed data was good, but the periods in between the observations seemed to be highly incorrect. Several drops in aboveground biomass were estimated between the cuts. The second cut in the second and third year seemed to be perfectly estimated. The 4 parameter model highly underestimated aboveground biomass for all the field observations, except that the last cut in the last year gave a perfect fit.

[FIGURE 4]

3.3.2. Validation dataset

The NRMSEP and MSEP (Table 3) and the percentage decomposition of the MSE (Figure 5) was calculated for the *P. pratense* and the *A. elatius* dominated grass growth at Rengen, using the validation data. For both datasets, the model with the 45 parameter set predicted grass growth best, with a mean NRMSEP over species of 0.65. The second best model was the fully parameterised model with an average NRMSEP of 0.67. The 9 parameter model gave mean NRMSEP of 0.75 while the worst prediction was made by the 4 parameter model with 0.79 as mean NRMSEP. The *A. elatius* dominated grass growth data from Rengen were dominated by the bias (63-90%), followed by phase shift (5-37%) and variance error (< 5%) (Figure 5b). For the *P. pratense* dominated grass growth data from Rengen on the other hand, the effect was more spread out between the components, except for the error with the 45 parameter set that was dominated by the phase shift component (92%).

[TABLE 3]

[FIGURE 5]

Predictive uncertainty in model outputs induced by parameter uncertainty is shown in Figure 6, together with field observations. The uncertainties are given as prior parameter knowledge and posterior parameter knowledge, both according to the 45 and the 4 parameter model, calculated by sampling randomly 100,000 samples from the prior distribution and from the posterior chains. Model output are then calculated for each parameter set, and uncertainty plotted as minimum and maximum model output for each day. Figure 6a gives the results for the *P. pratense* dominated sward at Rengen 1991-1994. The output uncertainty for this situation decreased slightly when using the results from the 45 parameter model calibration compared to our prior probability distributions of the parameters. According to the 4 parameter model, a much clearer decrease was found in predictive uncertainty, compared to both the prior and the posterior uncertainty from the 45 parameter model. All observations fall within both the prior and the posterior uncertainties, except the first observation of each year, which did not fall within the posterior uncertainty from the 4 parameter model. Figure 5b

gives the results for the *A. elatius* dominated sward at Rengen 2003-2005. Also here, the same pattern of decreased uncertainties for the posterior uncertainties compared to the prior was found. All observations did fall within the prior uncertainty and the posterior uncertainty for the 45 parameter model, but only one of the observations fell within the posterior predictive uncertainty for the 4 parameter model.

[FIGURE 6]

4. DISCUSSION

Process-based growth models, as the BASGRA model, are usually parameter rich. Satisfactory simplification of such models has previously been shown (Oomen et al. 2016; Raj et al. 2016). Based on the sensitivity analysis performed in this study, reduction of the number of model parameters seems possible for the BASGRA model as well. Results from the analysis showed large differences between the impact of parameters on model output, which is consistent with similar studies of other parameter rich crop models (Confalonieri, 2010; Confalonieri et al., 2010a; Confalonieri et al., 2010b; Richter et al., 2010; Thorsen et al., 2010). The screening method by Morris was easy to interpret and suitable in order to range the parameters according to their influence on model outputs. From the Morris method, we found some parameters that could be safely ignored in the calibration of our model, for our data. However, by calibrating the model with only a subset of parameters, rather than all of them, the error term (NRMSE) increased and the parametric uncertainty (CV) incorrectly decreased. When considering the error term and the parametric uncertainty as a function of the fraction of parameters kept in the calibration, a threshold of 0.7 of the parameters seemed to be required in order to have an acceptable model fit (NRMSE ≤ 0.1) and not a too highly underestimated uncertainty (CV $\geq$ 0.4) for both Særheim and Rengen. Considering the error term and the parametric uncertainty as a function of the highest normalized μ^* among the parameters left out from the calibration, this study showed that parameters with a lower normalized μ^* than 0.2 could be left out of a Bayesian calibration. This was valid for our simulations at both Særheim and Rengen, but need to be checked with other models and data as well for a generalisation. Consequently, the Morris method identified those parameters with such a small contribution to model output that they could be set equal to any value within their range without affecting model output considerably.

The importance of site specific sensitivity analysis, in order to investigate the stability of the sensitivity by the variety of climatic conditions for which the model is used, was highlighted by Confalonieri et al. (2010b). Accordingly, in this study the sensitivity analysis was performed using weather variables from two different sites in order to investigate the consistency of the ranking order of parameters according to their sensitivity to the model output across sites. The ranking order was not exactly the same across sites, but the same pattern was identified with the groups of sensitive and poorly sensitive parameters being very similar at both sites. This underlines the generality of the results and suggests that site specific sensitivity analysis is not needed for the range of variation in climate and soil conditions covered in the present study.

Three reduced models were developed, based on the sensitivity analysis of the BASGRA model, containing only the 4, 9 and 45 most sensitive parameters. The models were calibrated and error terms calculated based on model output from the new parameter values. According to both the training and the validation data, the error term for the differences between observed values and the estimated model output was similar for the fully parameterised model (66 parameters) and the reduced model consisting of 45 parameters. The 45 parameter model fitted well to all observations in all cases with training data, except one regrowth for the *P. pratense* dominated sward at Rengen that was not estimated properly. This small difference in the error term indicates that model reduction is possible without affecting model performance and that the response of aboveground biomass could be explained using fewer or simpler relationships. For the 45 and the fully parameterised model, the error term for differences between observed and simulated values was dominated by differences in timing (phase shift) considering the training data, whereas no general pattern was found in the decomposition of MSE for the validation data. However, a too strong model reduction should be avoided, as can be seen from the highly increased error term when considering the 4 and 9 parameter model. According to the training dataset, the 4 parameter model generally underestimated aboveground biomass highly.

Model calibration depends highly on the variability in the calibration data, and it is important to include as much valuable information about the processes as possible. A successful calibration requires appropriate data for model calibration (Yapo et al. 1996). In this study, we only had harvest observations for the aboveground biomass at Rengen. Therefore, the Bayesian calibration was performed twice for *P. pratense* grass growth in Særheim, firstly by using the full time series of aboveground biomass observations and secondly by only

including the aboveground biomass observations at regular harvests as calibration data, in order to visualise the effect. The model predictions fitted the output at harvest best when only the regular harvest observations were included in the analysis, but at the same time the growth and regrowth periods became unrealistic with high regrowth during winter when in practice growth is severely restricted by low temperature and solar radiation. In order to estimate grass growth through time, and not only make predictions of yield at harvest, it is consequently important to include data from the regrowth period into the calibration.

For Rengen, only harvest observations were included in the calibration. For the 45 parameter model, the *P. pratense* swards seemed to have a nice and realistic estimation of grass growth also between harvests, according to the training dataset. For the *A. elatius* dominated sward on the other side, highly unrealistic values were estimated. Although the harvest observations fitted almost perfectly, an unrealistic decline in aboveground biomass was estimated prior to several of the harvests. These poorer results for *A. elatius* may be due to physiological or morphological differences between this species and the better studied *P. pratense* that was not covered in the model. In addition, by including more observation points for biomass between harvests in the calibration dataset, the result would have looked different, as illustrated with the two calibrations for Særheim, with time series of biomass growth observations.

Output uncertainty caused by parameter uncertainty was included in the study, and a high uncertainty on model outputs was estimated based on prior parameter knowledge of all 66 parameters. The uncertainty was largest at harvest, and declined thereafter, with an estimated lower boundary of zero aboveground biomass for all days included. The posterior uncertainty of the 45 parameter model was based on the posterior parameter knowledge of the 45 most sensitive parameters, with the remaining parameters fixed at a certain value. The uncertainty was slightly reduced through the analysed period, caused by a combination of fixing uncertain values and by updating the remaining parameter uncertainty with new knowledge through observed data. All observed validation values for Rengen were within both the prior and the posterior uncertainty for the 45 parameter model. The 4 parameter model highly reduced the output uncertainty. The same data were included in the calibration of the 4 and the 45 parameter models, but the much higher number of uncertain parameters to be fixed at a certain value in the 4 parameter model, highly affected the uncertainty. Several of the observations fell outside the posterior uncertainty, which clearly demonstrated the danger of fixing uncertain parameter values. However, fixing only the 21 most uncertain values did not seem to affect significantly, thus confirming their weak effect on model output.

In the present study, the estimated MAP values for the parameters highly depended on the dataset used in the calibration. It is therefore important to calibrate the model for the specific case for which the model will be used i.e. specific sites and species as well as the target output variable (harvest only versus biomass growth dynamics).

Additionally, estimated MAP values depended on the model calibration procedure (45 or 4 parameter model). High variations were detected for the four most sensitive parameters when comparing the MAP values estimated from the calibration of the 4 parameter model compared to the 45 parameter model. As several of these 45 parameters proved to have an influence on model output, covariances between the parameters led to different estimates for the 4 most sensitive parameters depending on the values used for the remaining 41 (fixed to their nominal value in the 4 parameter model, while they are fixed to their MAP value in the 45 parameter model).

Winter observations were not included in this study, and the unimportant parameters identified included several “winter” processes governing the dynamics of water in the forms of snow cover and ice layer. Still, a parameter governing the rate of death due to frost (KRSR3H) appeared to be the most sensitive parameter. This unexpected result may be explained by the operations in the Morris method and its use of prior boundaries instead of prior probabilities. As the prior probability of the parameter was given by a beta distribution with its maximum value being the most probable and almost zero probability for the lower part of its interval, the sensitivity analysis includes these values as well. Since these values of negligible probability highly impacted model output, the parameter was regarded as sensitive. Similarly, the width of the parameter boundary will impact parameter sensitivity to model output, since model output will be less impacted when a parameter is only allowed to be varied within a narrow boundary compared to a wider boundary. Consequently, a parameter to which the model is very sensitive may only be so because it has a wide prior boundary. Once we know the value of a parameter very well (e.g. after a calibration leading to a narrow marginal posterior distribution for that parameter), the sensitivity can disappear. The Morris method applied to the prior gives higher sensitivity than Morris applied to the posterior. Sensitivity is consequently not purely a function of model, parameter and environmental conditions at the simulation site, but depends on our knowledge as well.

The BASGRA model was built in order to estimate 21 different output variables, but only the aboveground biomass was considered in this study. Consequently, the simplifications made here based on sensitivity analysis, and the parameterisation done by Bayesian calibration are

only relevant for this single output variable. In order to retain the complexity of the model and the ability to estimate several model outputs, more output variables should be considered in the analysis and more observed data included. Biomass yield is the most commonly used model output in practice, and this study is therefore highly important and relevant.

The validation of well calibrated simulation models is often limited because of insufficient data. Long-term experiments, from where the data in this study were derived, are an excellent source especially because management is kept constant and environmental conditions are well-known. In order to validate the model and to test its suitability also at broader scales, however, other sources need to be explored. Remote sensing may contribute to calibration as well as validation of such models through the provisioning of crop parameters and variables such as leaf area index (Darvishzadeh et a., 2011), crop phenometrics (Parplies et al., 2016) and dry matter yield (Quan et al., 2017), the latter by coupling remote sensing information with a radiative transfer model. That way, simulations could also be supported through coupling such data with spatially explicit site information on e.g. soil properties in a Geographic Information System.

5 CONCLUSION

The objective of this study was to examine the impact of parameter screening and subsequent parameter reduction on aboveground biomass predictions by the grassland model BASGRA, in order to efficiently be able to include uncertainty in model outputs. According to this study, a reduction of model parameters from 66 to 45 was possible. The error term, for both the 45 and the fully parameterised model was characterised by the timing (phase shift) when considering the training data, while no general pattern was found in the decomposition of the MSE for the validation data. The sensitivity analysis showed that the parameters to be fixed were consistent across sites (variation in climate and soil conditions), while model calibration had to be performed separately for each specific case (site and species) for which the model was used. The output uncertainty decreased slightly, but still covered the field observations of aboveground biomass. Strong model reductions to 9 or 4 parameters should be avoided because they lead to highly increased error terms and underestimated model output uncertainties. In order to estimate the periods between the regular harvests adequately, it proved to be important to include data from the regrowth period as well, in addition to the aboveground biomass at the regular harvests, especially for the *A. elatius* sward at Rengen.

658 The model has originally been built for *P. pratense* grass growth, and physiological or
659 morphological differences between the two species may have been neglected. Better
660 predictions could possibly have been identified in the model by including regrowth data
661 during calibration.

662

663 **ACKNOWLEDGEMENT**

664 This work was funded by the Research Council of Norway and was conducted within the
665 framework of the Modelling European Agriculture with Climate Change for Food Security
666 (MACSUR) knowledge hub within the Joint Programming Initiative for Agriculture, Climate
667 Change, and Food Security (FACCE-JPI).

668 Jürgen Schellberg acknowledges funding from the SPACES project "Limpopo Living
669 Landscapes" (01LLL1304C) funded by Federal Ministry of Education and Research.

670

671 **REFRENCES**

672 Agrometeorology Norway, 2015. http://lmt.nibio.no/.

673 Angstrom, A., 1924. Solar and terrestrial radiation. Report to the international commission for
674 solar research on actinometric investigations of solar and atmospheric radiation. Quarterly
675 Journal of the Royal Meteorological Society 50, 121-126.

676 Berger, O.J., 1985. Statistical Decision Theory and Bayesian Analysis, New York.

677 Campbell, K., 2006. Statistical calibration of computer simulations. Reliability Engineering &
678 System Safety 91, 1358-1363.

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

681 Chytrý, M., Hejcman, M., Hennekens, S.M., Schellberg, J., 2009. Changes in vegetation types
682 and Ellenberg indicator values after 65 years of fertilizer application in the Rengen Grassland
683 Experiment, Germany. Applied Vegetation Science 12, 167-176.

684 Confalonieri, R., 2010. Monte Carlo based sensitivity analysis of two crop simulators and
685 considerations on model balance. European Journal of Agronomy 33, 89-93.

686 Confalonieri, R., Bellocchi, G., Bregaglio, S., Donatelli, M., Acutis, M., 2010a. Comparison
687 of sensitivity analysis techniques: A case study with the rice model WARM. Ecological
688 Modelling 221, 1897-1906.

Confalonieri, R., Bellocchi, G., Tarantola, S., Acutis, M., Donatelli, M., Genovese, G., 2010b. Sensitivity analysis of the rice model WARM in Europe: Exploring the effects of different locations, climates and methods of analysis on model sensitivity to crop parameters. Environmental Modelling & Software 25, 479-488.

Cox, G.M., Gibbons, J.M., Wood, A.T.A., Craigon, J., Ramsden, S.J., Crout, N.M.J. 2006. Toward the systematic simplification of mechanistic models. Ecological Modelling 198, 240-246.

Crout, N.M.J., Craigon, J., Cox, G.M., Jao, Y., Tarsitano, D., Wood, A.T.A., Semenov, M. 2014. An objective apporach to model reduction: Application to the Sirius wheat model. Agricultural and Forest Meteorology 189-190, 211-219.

Darvishzadeh, R., Atzberger, C., Skidmore, A., Schlerf, M. 2011. Mapping grassland leaf area index with airborne hyperspectral imagery: A comparison study of statistical approaches and inversion of radiative transfer models. ISPR Journal of Photogrammetry and Remote Sensing, 66, 894-906.

Ewert, F., Rodriguez, D., Semenov, M.A., Mitchell, R.A.C., Goudriaan, J., Porter, J.R., Kimball, B.A., Pintr Jr, P.J., Manderscheid, R., Weigel, H.J., Fangmeier, A., Fereres, E., Villalobos, F. 2002. Effects of elevated CO_2 and drought on wheat: testing crop simulation models for different experimental and climatic conditions. Agriculture, Ecosystems and Environment 93: 249.266.

Gouache, D., Bensadoun, A., Brun, F., Pagé, C., Makowski, D., Wallach, D., 2013. Modelling climate change impact on Septoria tritici blotch (STB) in France: Accounting for climate model and disease model uncertainty. Agricultural and Forest Meteorology 170, 242-252.

Höglind, M., Hanslin, H.M., Van Oijen, M., 2005. Timothy regrowth, tillering and leaf area dynamics following spring harvest at two growth stages. Field Crops Research 93, 51-63.

Höglind, M., Schapendonk, A.H.C.M., Van Oijen, M., 2001. Timothy growth in Scandinavia: combining quantitative information and simulation modelling. New Phytologist 151, 355-367.

Kennedy, M.C., O'Hagan, A. 2001. Bayesian calibration of computer models. Journal of the Royal Statistical Society. Series B 63: 425-464.

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

Lamboni, M., Makowski, D., Lehuger, S., Gabrielle, B., Monod, H., 2009. Multivariate global sensitivity analysis for dynamic crop models. Field Crops Research 113, 312-320.

Liu, J.S., 2001. Monte Carlo Strategies in Scientific Computation, New York.

Minunno, F., Oijen, M.v., Cameron, D.R., Pereira, J.S., 2013. Selecting Parameters for Bayesian Calibration of a Process-Based Model: A Methodology Based on Canonical Correlation Analysis. SIAM/ASA Journal on Uncertainty Quantification 1, 370-385.

Morris, M.D., 1991. Factorial Sampling Plans for Preliminary Computational Experiments. Technometrics 33, 14.

Oomen, R.J., Ewert, F., Snyman, H.A. 2016. Modelling rangeland productivity in response to degredation in a semi-arid-climate. Ecological Modelling 322, 54-70.Parplies, A., Dubovyk, O., Tewes, A., Mund, J.P., Schellberg, J. 2016. Phenomapping of rangelands in South Africa using time series of RapidEye. International Journal of Applied Earth Observation and Geoinformation, 53, 90 - 102. .http://dx.doi.org/10.1016/j.jag.2016.08.001

Quan, X., He, B., Yebra, M., Yin, C., Liao, Z., Zhang, X., Li,X. 2017. A radiative transfer model-based method for the estimation of grassland aboveground biomass. International Journal of Earth Observation and Geoinformation, 54, 159-168.

Raj, R., Hamm, N.A.S., van der Tol, C., Stein, A. 2016. Bayesian integration of flux tower data into process-based simulator for quantifying uncertainty in simulated output. Geoscientific Model Development. doi: 10.5194/gmd-2016-216.

Richter, G.M., Acutis, M., Trevisiol, P., Latiri, K., Confalonieri, R., 2010. Sensitivity analysis for a complex crop model applied to Durum wheat in the Mediterranean. European Journal of Agronomy 32, 127-136.

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

Saltelli, A., Tarantola, S., Campolongo, F., Ratto, M., 2004. Sensitivity Analysis in Practice, New York.

Schapendonk, A.H.C.M., Stol, W., van Kraalingen, D.W.G., Bouman, B.A.M., 1998. LINGRA, a sink/source model to simulate grassland productivity in Europe. European Journal of Agronomy 9, 87-100.

Schellberg, Möseler, Kühbauch, Rademacher, 1999. Long-term effects of fertilizer on soil nutrient concentration, yield, forage quality and floristic composition of a hay meadow in the Eifel mountains, Germany. Grass and Forage Science 54, 195-207.

Thorsen, S.M., Höglind, M., 2010. Assessing winter survival of forage grasses in Norway under future climate scenarios by simulating potential frost tolerance in combination with simple agroclimatic indices. Agricultural and Forest Meteorology 150, 1272-1282.

Thorsen, S.M., Roer, A.-G., Van Oijen, M., 2010. Modelling the dynamics of snow cover, soil frost and surface ice in Norwegian grasslands. Polar Research 29, 110-126.

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

van Oijen, M., Höglind, M., Cameron, D.R., Thorsen, S.M., 2015. BASGRA 2014. http://dx.doi.org/10.5281/zenodo.27867.

van Oijen, M., Höglind, M., Hanslin, H.M., Caldwell, N., 2005a. Process-Based Modeling of Timothy Regrowth. Agronomy Journal 97, 1295-1303.

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

Yapo, P.O., Gupta, H.V., Sorooshian, S. 1996. Automatic calibration of conceptual rainfall-runoff between models: sensitivity to calibration data. Journal of Hydrology 181(1-4): 23-48.

Table 1: Nominal values and maximum posterior (MAP) estimates from Bayesian calibration of the 4 and 45 most sensitive parameters from (1) *P. pratense* growth at Særheim, (2) harvest observations of *P. pratense* growth at Særheim, (3) harvest observations of *P. pratense* dominated grass growth at Rengen and (4) harvest observations of *A. elatius* dominated grass growth at Rengen.

Parameter	Nominal value	*P. pratense* Særheim All data	*P. pratense* Særheim Harvest data	*P. pratense* Rengen Harvest data	*A. elatius* Rengen Harvest data
	The reduced parameter set of 4 parameters				
DAYLB	0.392	0.479	0.250	0.508	0.655
DLMXGE	0.992	0.688	0.693	0.964	0.406
KRSR3H	1.00	0.982	0.946	0.821	0.417
ROOTDM	0.761	1.26	0.621	0.579	1.23
	The reduced parameter set of 45 parameters				
COCRESMX	0.141	0.0989	0.185	0.0967	0.179
CSTAVM	0.230	0.130	0.331	0.218	0.194
DAYLB	0.392	0.530	0.438	0.447	0.0811
DAYLP	0.632	0.683	0.617	0.604	0.487
DLMXGE	0.992	0.649	0.823	0.891	0.816
Dparam	0.00320	0.0029	0.0032	0.0029	0.0036
FSLAMIN	0.466	0.641	0.329	0.161	0.355
Hparam	0.00560	0.0069	0.0086	0.0061	0.007
K	0.500	0.513	0.584	0.630	0.509
KLUETILG	0.500	0.413	0.500	0.476	0.145
KRESPHARD	0.0100	0.0234	0.0212	0.00783	0.0295
KRSR3H	1.00	0.911	0.847	0.911	0.889
LAICR	3.80	2.14	6.30	2.55	3.09
LAIEFT	0.200	0.226	0.189	0.172	0.220
LAITIL	0.567	0.776	0.391	0.593	0.974
LERGb	2.89	0.736	5.86	7.67	8.38
LERVa	-2.76	-6.42	-2.83	-7.27	-4.47
LERVb	0.520	0.685	0.470	0.472	1.05
LFWIDG	(0.00850/0.00600)*	0.006	0.0102	0.00987	0.0104
LFWIDV	(0.00490/0.00300)*	0.005	0.005	0.0046	0.0027
LOG10CLVI	1.50	1.98	1.56	1.03	1.67
LOG10CRESI	0.500	0.393	0.673	-0.565	0.766
LOG10LAII	0.00	-0.767	-0.530	-0.290	0.204
LT50MN	-(26.7/16.0)*	-20.1	-21.9	-20.1	-20.9
LT50MX	-4.79	-5.04	-4.49	-4.44	-5.08
NELLVM	2.0918\1	1.14	1.10	1.35	1.99
PHENCR	0.495	0.713	0.823	0.636	0.373
PHY	(63.1/110.0)*	57.3	75.3	74.8	88.2
RDRSCO	0.0712	0.0604	0.0559	0.0969	0.0797
RDRSMX	0.0600	0.0566	0.0504	0.0502	0.0854
RDRTEM	0.00100	0.0013	0.0009	0.0008	0.0009
reHardRedDay	145	98.5	155	142	114
RGENMX	0.0109	0.0197	0.0158	0.00847	0.0147
ROOTDM	0.761	0.660	0.553	0.937	0.231
RUBISC	5.78	4.31	5.87	4.83	3.94
SHAPE	0.539	0.866	0.489	0.484	0.440
SINMAX1T	0.00450	0.0046	0.0058	0.0040	0.0049
SLAMAX	0.0600	0.0612	0.0476	0.0714	0.0792
TBASE	3.61	3.23	3.90	4.25	4.12
TCRES	1.89	2.69	2.10	2.44	3.27
THARDMX	14.7	14.7	13.7	14.8	14.8
TILTOTI400	1600	1030	1410	1290	897
TOPTGE	12.6	9.21	11.5	8.89	7.58
TsurfDiff	0.623	2.59	0.939	3.30	3.15
YG	0.842	0.823	0.732	0.798	0.849

* The first number is nominal value for *P. pratense* and the second number is for *A. elatius*

Table 2: Normalised root mean square errors (NRMSE) and the mean square errors (MSE) in parentheses calculated between model outputs and observed aboveground biomass of *P. pratense* grass growth at Særheim, harvest observations of *P. pratense* grass growth at Særheim harvest observations of *P. pratense* dominated grass growth at Rengen and *A. elatius* dominated grass growth at Rengen (training data). The model outputs are calculated for the models constructed by the fully parameterised model and the models with the 4, 9 and 45 most important parameters according to sensitivity analysis, using MAP values on the calibrated parameters and their nominal values for the others and for the total parameter set.

Dataset	4 par	9 par	45 par	66 par
***P. pratense* Særheim**	0.8129	0.4423	0.1105	0.0940
All data	(131,090)	(38,806)	(2,423)	(1,751)
***P. pratense* Særheim**	0.8277	0.3919	0.0211	0.0195
Harvest data	(467,450)	(104,800)	(302)	(259)
***P. pratense* dominated Rengen**	0.7855	0.7059	0.1578	0.1614
Harvest data	(66,801)	(53,939)	(2,696)	(2,821)
***A. elatius* dominated Rengen**	0.7483	0.8119	0.0578	0.0606
Harvest data	(84,887)	(99,934)	(505)	(557)

Table 3: Normalised root mean square errors of prediction (NRMSEP) and the mean square errors (MSE) in parentheses calculated for the difference between calculated and observed aboveground biomass at harvest of *P. pratense* dominated grass growth at Rengen and *A. elatius* dominated grass growth at Rengen (validation data). The model outputs are calculated for the models constructed by the fully parameterised model and for the models with the 4, 9 and 45 most important parameters according to sensitivity analysis, using MAP values on the calibrated parameters and their nominal values for the others and for the total parameter set.

Dataset	4 par	9 par	45 par	66 par
***P. pratense* dominated Rengen**	0.5911	0.6153	0.5764	0.6414
Harvest data	(32,575)	(35,292)	(30,972)	(38,356)
***A. elatius* dominated Rengen**	0.9945	0.8839	0.7325	0.6804
Harvest data	(189,520)	(147,690)	(102,810)	(88,719)

Figure 1: Results from sensitivity analysis of the BASGRA model, using the Morris method for a) Særheim weather data, soil data and harvest dates in 2000-2001, b) Rengen weather data, soil data and harvest dates in 2000-2001. Only the nine most important parameters according to the sensitivity analysis are named. These are: a constant in the logistic curve for frost survival (KRSR3H), day length below which the day length effect on allocation, tillering, leaf appearance and leaf elongation becomes less than 1 (DLMXGE), the initial and maximum value of rooting depth (ROOTDM), day length below which phenological stage is reset to zero (DAYLB), maximum SLA of new leaves (SLAMAX), the minimum SLA of new leaves as a fraction of maximum possible SLA (FSLAMIN), the maximum ratio of tiller and leaf appearance at low leaf area index (LAITIL) and the rate of elongation of leaves on non-elongating tillers (LERVb).

Figure 2: The percentage decomposition of mean square error (MSE) into bias, variance error and phase shift for the 4, 9, 45 and 66 parameter model for the training data with a) *P. pratense* grass growth at Særheim, b) harvest observations of *P. pratense* grass growth at Særheim, c) harvest observations of *P. pratense* dominated grass growth at Rengen and d) harvest observations of *A. elatius* dominated grass growth at Rengen.

Figure 3: Normalised root mean square error (NRMSE) and parametric uncertainty (CV) as a function of the fraction of parameters included for a) Særheim and b) Rengen, and NRMSE and CV as a function of the highest normalised μ^* (mean from the Morris method) among the parameters left out from the calibration for c) Særheim and d) Rengen.

Figure 4: Observed values and model outputs of the BASGRA model using MAP estimates of the 45 and 4 most sensitive parameters according to sensitivity analysis for a) *P. pratense* grass growth at Særheim in 2000 b) *P. pratense* grass growth at Særheim in 2001-2002 c) harvest observations of *P. pratense* grass growth at Særheim for 2000 d) harvest observations of *P. pratense* grass growth at Særheim for 2001-2002 e) harvest observations of *P. pratense* dominated grass growth at Rengen 1989-1992 and f) harvest observations of *A. elatius* dominated grass growth at Rengen for 2000-2002. Time is equivalent to the number of days, starting at sowing day.

Figure 5: The percentage decomposition of mean square error (MSE) into bias, variance error and phase shift for the 4, 9, 45 and 66 parameter model for the validation data with a) harvest observations of *P. pratense* dominated grass growth at Rengen and b) harvest observations of *A. elatius* dominated grass growth at Rengen.

Figure 6: Prior and posterior (for the 45 and 4 parameter models) output uncertainty and observed values for (a) *P. pratense* dominated grass growth in Rengen 1991-1994 and (b) *A. elatius* dominated grass growth in Rengen 2003-2005.

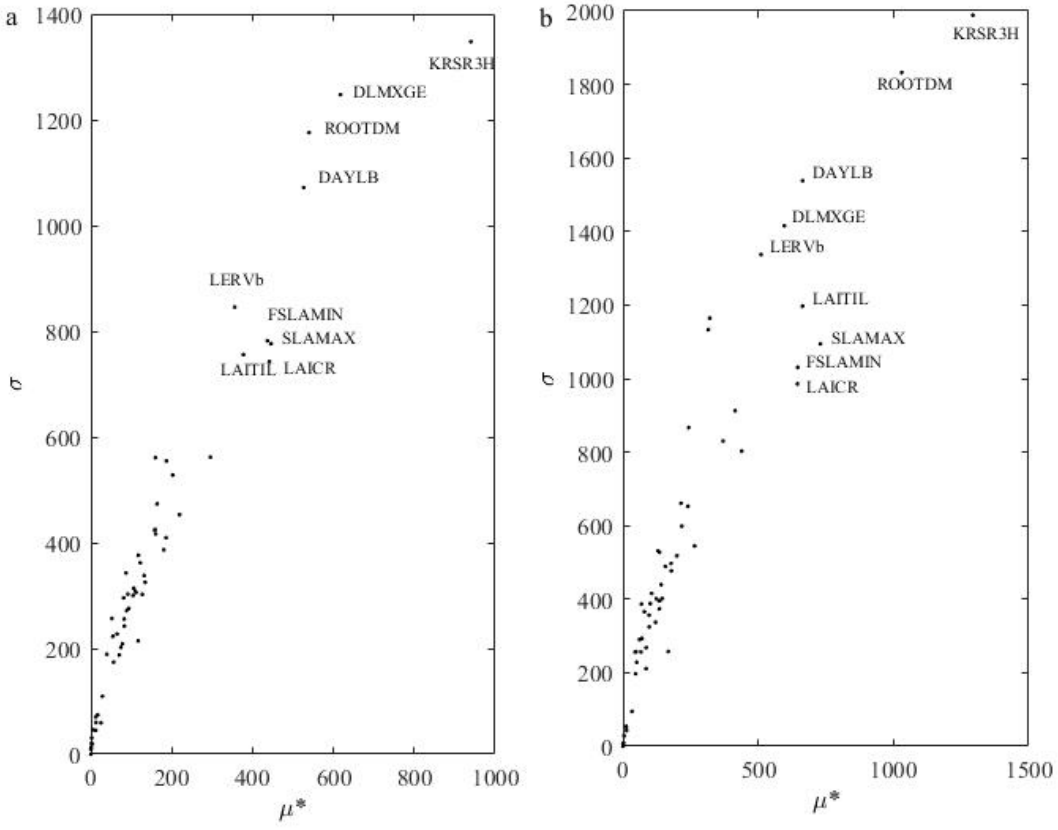

a
b
σ
μ*
KRSR3H
DLMXGE
ROOTDM
DAYLB
LERVb
FSLAMIN
SLAMAX
LAITIL
LAICR

Figure 1

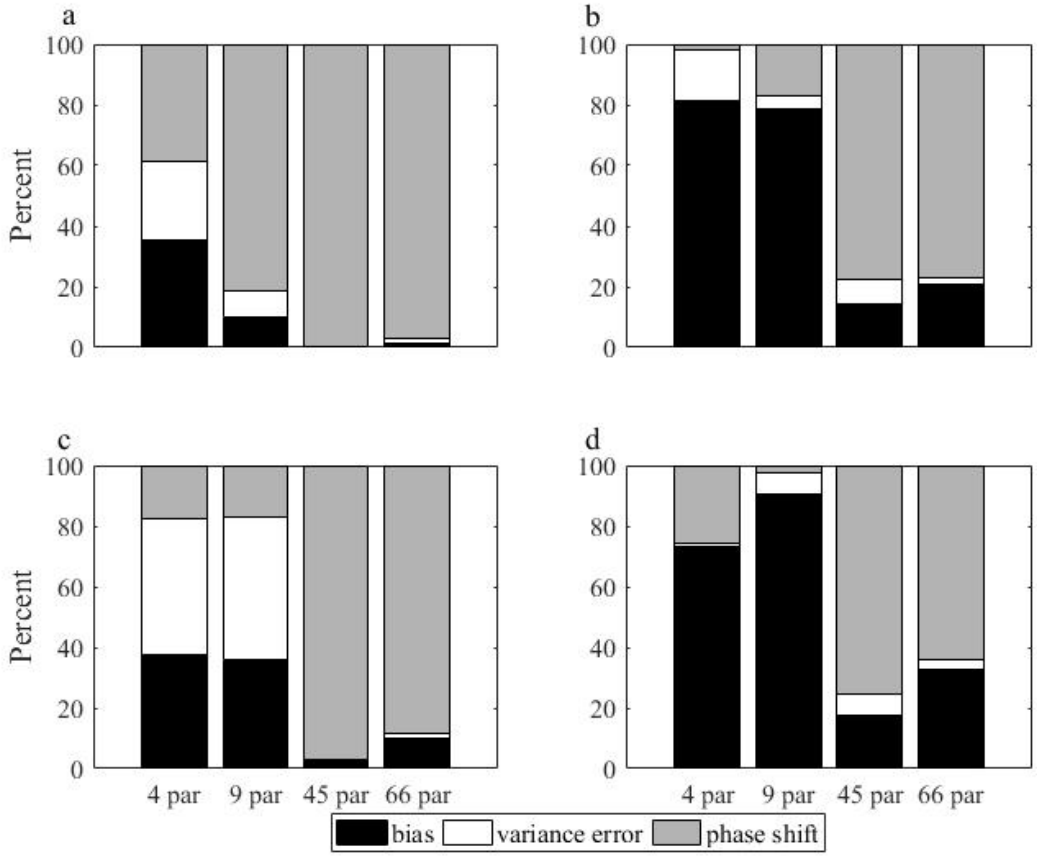
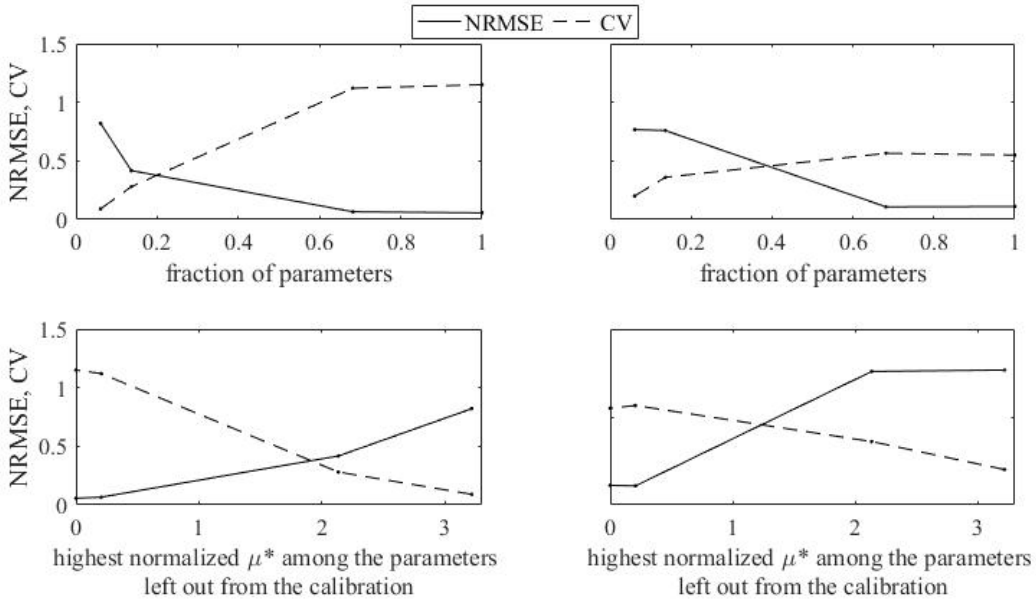
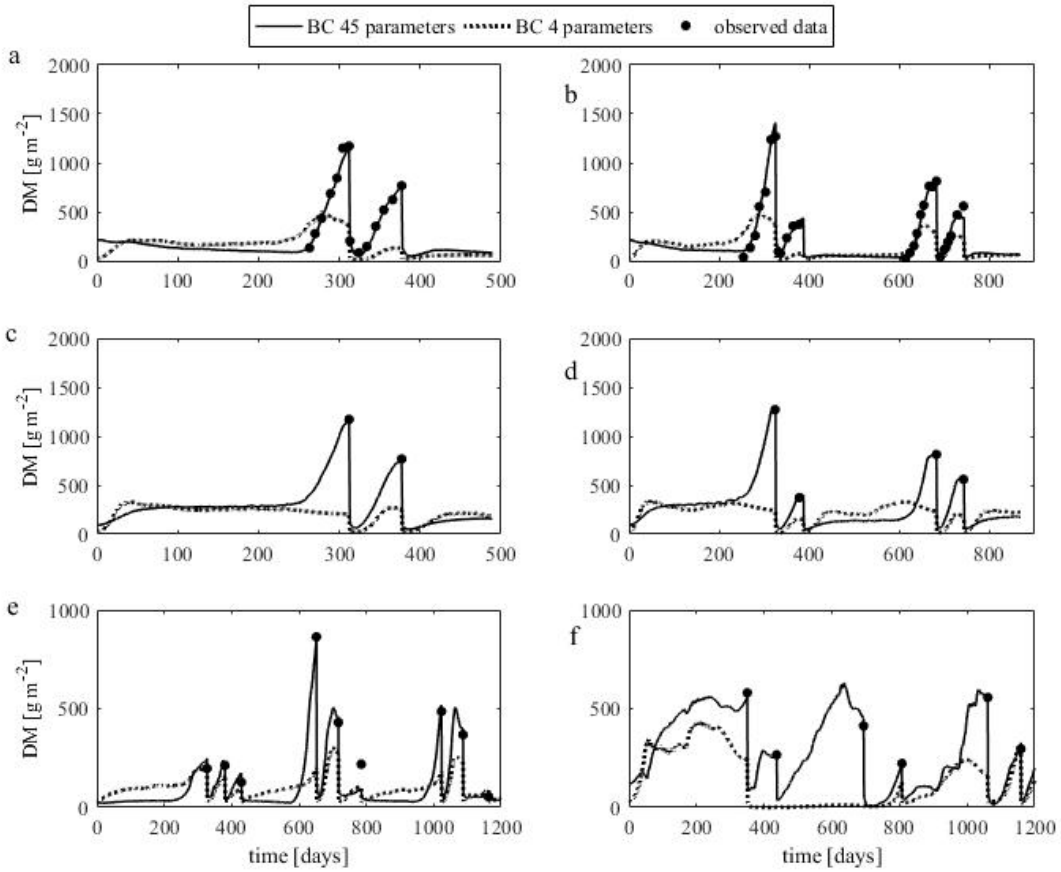

a
b
c
d
Percent
100
80
60
40
20
0
4 par
9 par
45 par
66 par
bias
variance error
phase shift

Figure 2


NRMSE
CV
NRMSE, CV
fraction of parameters
highest normalized μ^* among the parameters left out from the calibration

858 Figure 3:

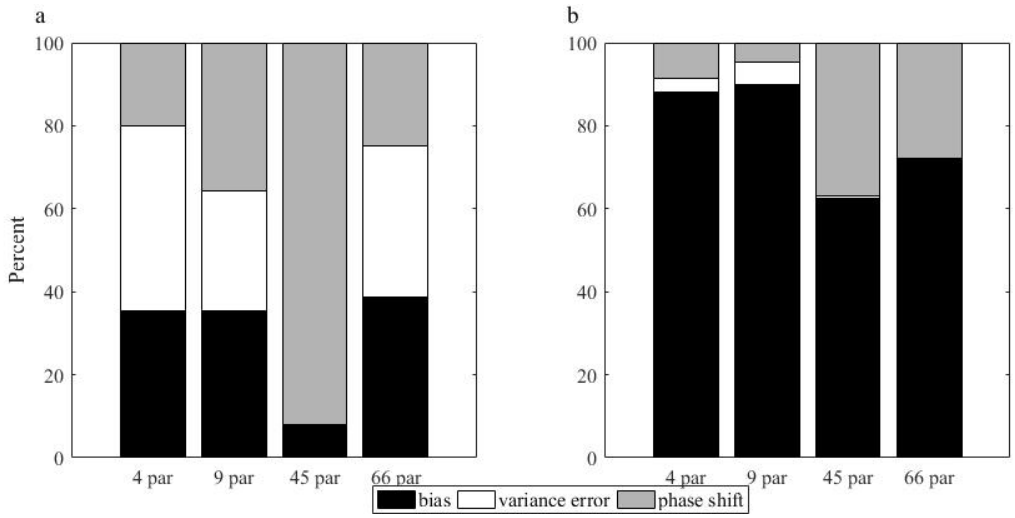


859

860 Figure 4

861

862



863

864 Figure 5

865

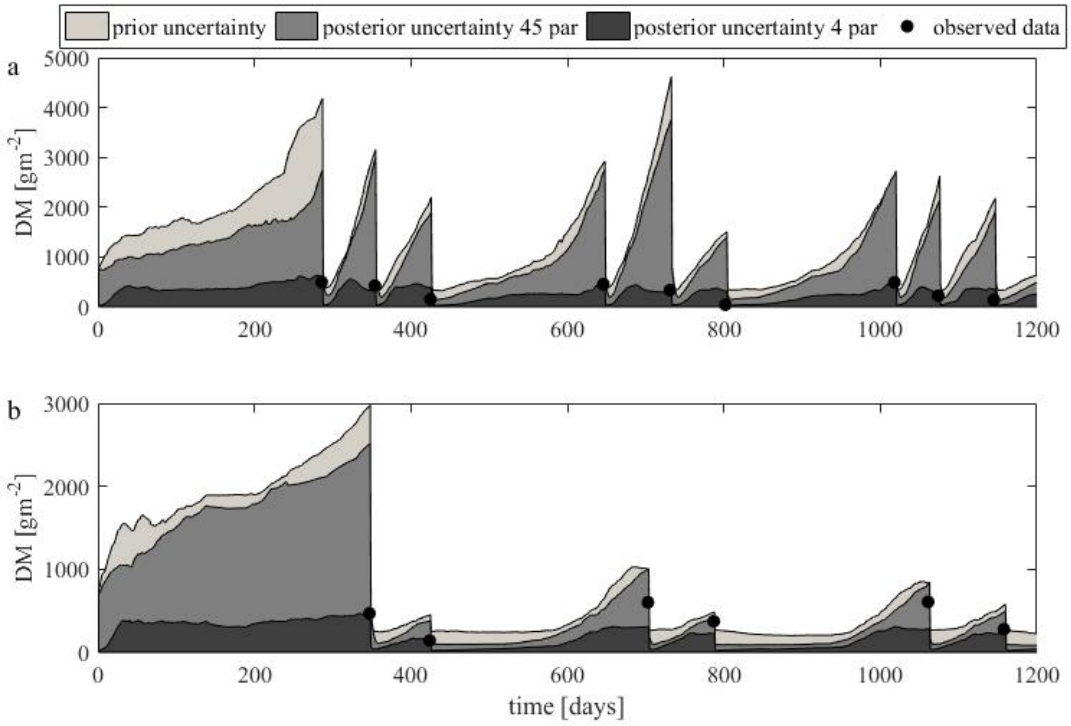

prior uncertainty
posterior uncertainty 45 par
posterior uncertainty 4 par
observed data
a
b
DM [gm-2]
time [days]
0
200
400
600
800
1000
1200
1000
2000
3000
4000
5000

Figure 6