Calibrating ModVege

I. INTRODUCTION

In the following we describe the approach that was used to identify optimal site and model parameters based on existing experimental data. This process is called *model calibration*. ModVege [1] is designed to require only three inputs pertaining to the simulated site:

- The nutritional index NI, a dimensionless value in the interval $[0.3, 1]$ which represents plant nutritional availability.
- The soil's water-holding capacity WHC in mm.
- The composition of plant functional groups in the site's plant community according to the categorization by Cruz *et al.* [2]. This is expressed using weights for the four different functional groups (w_i with $i \in \{A, B, C, D\}$), where

$$\sum_{i \in \{A,B,C,D\}} w_i = 1 \quad . \tag{1}$$

II. GENERAL CONSIDERATIONS

Different strategies are commonly employed for calibrating models and there are many decisions that can influence the calibration procedure [3].

A. Model Performance

Firstly, one has to define how model performance is measured and thus how the suitability of a set of parameters is evaluated. This begins with the choice of modeled variables to consider. Here, reasonable variables are those related to biomass growth, as these are the quantities for which experimental data is available: BM, cBM, dBM or harvested biomass hvBM. Other options could be to compare the predicted vegetations times, start of the growing season, or the peak height in daily biomass growth. All of these latter options are only indirectly available from the measured data, and only with rather large uncertainties. For this reason, we only consider these aspects secondary to the directly measured biomass quantities.

Once suitable variables are chosen, one also has to define in which way error is expressed. A common choice is the square-root of mean squared deviations σ_{RMSE}, sometimes called root-mean-squared error (RMSE):

$$\sigma_{\mathrm{RMSE}} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(y_i - y_i^{\mathrm{model}}\right)^2} \tag{2}$$

with the N observations y_i and model predictions y_i^{model}. Here, we are also going to make use of the mean absolute error σ_{MAE} and the bias b:

$$\sigma_{\mathrm{MAE}} = \frac{1}{N} \sum_{i=1}^{N} \left| y_i - y_i^{\mathrm{model}} \right| \tag{3}$$

$$b = \frac{1}{N} \sum_{i=1}^{N} y_i - y_i^{\mathrm{model}} \quad . \tag{4}$$

The available grass growth curves provide the average daily biomass growth over the measurement period T of two weeks. The variable we choose for model evaluation is therefore derived from of the model output variable dBM. We replicate the measurement situation by averaging dBM over the past two weeks:

$$y_i^{\mathrm{model}} = \overline{\mathrm{dBM}}_i = \frac{1}{T} \sum_{j=i-T}^{i} \mathrm{dBM}_j \quad . \tag{5}$$

B. Frequentist or Bayesian

In a frequentist approach, one typically carries out some sort of minimization of a defined objective function, often in the form of a gradient descent. In a Bayesian framework, one has to make assumptions about the forms of the prior distributions and the likelihood. Here, the typical implementation of a Bayesian calibration, the Metropolis or Markov-Chain Monte-Carlo (MCMC) algorithm, could not be employed in a straightforward manner due to interdependence of the parameters w_i through Equation 1. We have thus devised an efficient manner for sampling parameter space and extensively tested an MCMC approach to calibration. Despite intense efforts, we had to conclude that this strategy is unfit to reasonably calibrate ModVege with the existing data, see section V for details.

C. Choice of Calibration Parameters

It is not always necessary or meaningful to calibrate all model parameters. Often, one can limit the effort to a subset. In our case, WHC can be inferred from the Swiss soil suitability map [4] and does therefore not have to be calibrated on a per-site basis. Thus, we are left with the parameters NI, w_A, w_B, w_C and w_D that need to be calibrated.

III. CALIBRATION PROCEDURE

In line with the observation that fully-automated calibrations are not unequivocally advised [3] and the ap-

parent equifinality that hindered Bayesian calibration (cf. Appendix), we eventually defined a semi-automatic calibration procedure. In it, we iteratively sample subsequently smaller volumes in parameter space and evaluate the respective performance scores of the sampled parameter combinations. Based on these scores, the successive sampling volume in parameter space is defined and the procedure is re-iterated up to a defined convergence criterion.

In the following explanation of the procedure, we will use $p_{m,n}$ to refer to parameter m in iteration n and denote with $P_{m,n}$ the set of sampled parameter values for $p_{m,n}$. $p_{m,n}^{\min}$ and $p_{m,n}^{\max}$ then define the range which is considered for parameter m in iteration n. In the first iteration ($n = 1$), we usually allow the full parameter space. Therefore, $p_{m,1}^{\min} = 0$ and $p_{m,1}^{\max} = 1$ for $m \in \{w_A, w_B, w_C, w_D\}$ and $p_{\mathrm{NI},1}^{\min} = 0.3$, $p_{\mathrm{NI},1}^{\max} = 1$.

A. Step 1: Parameter Sampling

We choose for each parameter a number of values that lie within the allowed minimum and maximum values for that parameter in a given iteration n. The number of chosen values and their distribution might, in principle, depend on how much we expect that parameter to influence the results and with what precision we want to trace its influence. Typically, however, we choose N equidistantly spaced values between $p_{m,n}^{\min}$ and $p_{m,n}^{\max}$, i.e. $P_{m,n} = \{p_{m,n}^{\min}, p_{m,n}^{\min} + \Delta p_{m,n}, \ldots, p_{m,n}^{\max}\}$ with $\Delta p_{m,n} = (p_{m,n}^{\max} - p_{m,n}^{\min})/N$. N, or generally the number of elements in the $P_{m,n}$ should be chosen such that the computations in step 2 can be carried out in a reasonable timeframe.

We then generate all possible combinations C_n of the sampled parameters, taking care to remove those combinations I_n that do not fulfill the functional group sum criterion Equation 1:

$$C_n = \bigotimes_m P_{m,n} - I_n \quad . \tag{6}$$

The exclusion of combinations not fulfilling Equation 1 significantly reduces the number of elements in C_n, making this approach feasable despite the otherwise large numbers that arise combinatorially. We further note that care needs to be taken in the construction of the $P_{m,n}$ for the functional group weights w_i in that the resulting combinations actually need to be able to fulfill the sum rule (Equation 1). As an example, the following choices for $P_{w_i,n}$ would allow for only one valid combination, despite providing several values for each parameter:

$$\begin{aligned} P_{w_A,n} &= \{0.5, 0.75, 1\} \\ P_{w_B,n} &= \{0.5, 0.75, 1\} \\ P_{w_C,n} &= \{0.1, 0.2\} \\ P_{w_D,n} &= \{0.1, 0.2\} \quad . \end{aligned}$$

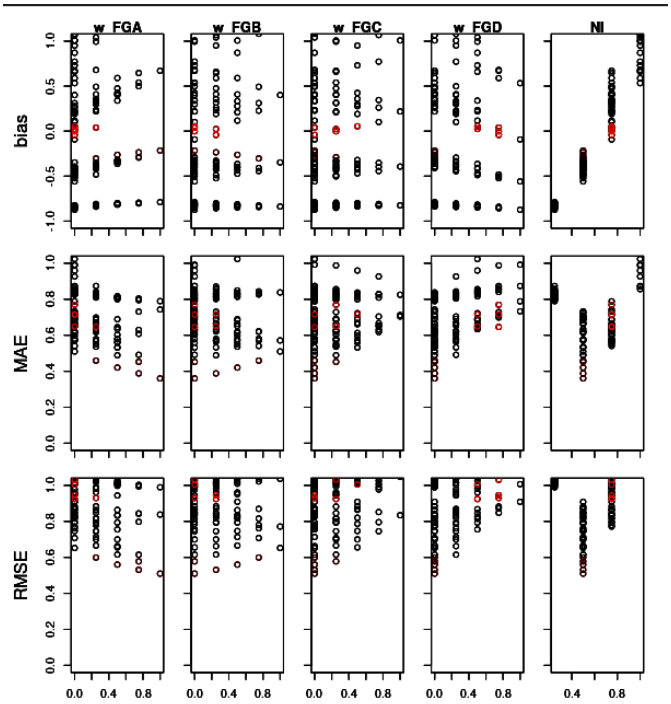

FIG. 1. Performance scores versus parameter value for different metrics for 140 combinations of parameters for an example site. The five top performing combinations in terms of bias b are highlighted in red. It is evident that these do not coincide with the top performers according to σ_{MAE} and σ_{RMSE}.

A general way for constructing suitable $P_{m,n}$ for the w_i is to (i) let the step size be equal for all sets and (ii) ensure that for every $p_{m,n}^{\max}$ there is at least one $p_{l,n}^{\min}$ with $p_{l,n}^{\min} = 1 - p_{m,n}^{\max}$. Even so, there is a danger of accidentally leaving out some volume of parameter space, even though it might look as if it was included on first glance. Human supervision of the process is therefore recommended.

B. Step 2: Performance Evaluation

We now run the model for each valid parameter combination $c \in C_n$, thus generating model outputs M_c. Performance scores for the metrics σ_{RMSE}^c, σ_{MAE}^c and b^c for each M_c are calculated using the measured data. These scores are carefully inspected and compared with each other in order to identify trends in the parameters. Generally, if a combination c scores well for all metrics, its parameter values should be within the parameter ranges for the next iteration. Likewise, parameter values of combinations that score badly in all metrics are candidates for exclusion. A way to visualize the results of this step is shown in Figure 1.

It might occur that there are less obvious situations: (i) Certain parameter combinations score very well for some metrics, but not for others. This case is showcased by the highlighted points in Figure 1. (ii) For one ore more pa-

rameters the whole range of its previous values appears in the top performing combinations. Case (i) is most easily addressed by requiring parameter combinations to score well across all metrics in order to be considered *good*. It is still advisable to take note of these cases and, potentially, decide whether the malperformance in one metric is made up for by the excellence in another. After all, we have already introduced an element of subjectivity through the choice of evaluated variable and used metrics. It may be that this subjectivity needs to be accounted for here. Case (ii) might hint towards the existence of separated optima in parameter space. One should carefully examine the different parameter combinations and potentially design the allowed parameter ranges for the next iteration with *holes*, i.e. with separated intervals. Alternatively, one could carry out separate runs towards the different optima and compare the converged results.

The above considerations implicitly assume the existence of one optimal, *correct* parameter set, an idea that is challenged by the concept of equifinality [5], confer also section V. In line with the validity of different parameter sets proposed by equifinality, one could also attempt to follow different optima and allow for multiple results instead of one. The final simulations should then be carried out with all identified optimal parameter sets and the variances in model output could be interpreted as a measure of model uncertainty.

C. Iteration Up To Convergence

Through the considerations in the previous step, one should be able to formulate the new parameter ranges $p^{\min}_{m,n+1}$, $p^{\max}_{m,n+1}$, possibly with *forbidden* intervals in between. Using this information, one can start over in step 1 and sample new sets of parameter values $P_{m,n+1}$ out of which to generate combinations C_{n+1}. This procedure is repeated up to the point where the parameter ranges are smaller than the desired precision ε_m for that parameter: $|p^{\max}_{m,\text{final}} - p^{\min}_{m,\text{final}}| < \varepsilon_m$.

The ε_m can be defined through considerations of how much a change Δp_m in a given parameter affects the model output M. If changes in model outputs are in the order of the measurement uncertainty or smaller, it is not possible to gain higher precision for the parameters. We find that ε_m is in the order of 0.1 for the w_i and $\varepsilon_{\text{NI}} \approx 0.05$.

D. Automation

We note that, despite the fact that we have chosen to manually inspect each step in the calibration, it is possible to employ the same routine in an unsupervised fashion. One would initiate the parameter scan with a sample of full parameter space. Criteria could then be formulated to automatically identify boundaries on the parameters for the next step. As an example, one could look at the distributions of parameters spanned by the top 30 % performing combinations for each metric. From these *top-performers*, one should exclude those that perform poorly in different metrics, e.g. there might be combinations that minimize the bias b but show large σ_{RMSE} and σ_{MAE} or vice versa. Given that only few (< 5) iterations are required, the reduction of workload gained by such an automation does not seem worth the loss of fine-grained control over the procedure.

IV. EVALUATION OF THE RESULTS

V. APPENDIX

Parameter correlations such as between the functional group weights, Equation 1, are not foreseen in typical MCMC implementations. The most straightforward way to incorporate them would be a rejection scheme, where parameters w_i are sampled until they fulfill Equation 1, up to a small error to account for numeric precision. This approach, however, leads to unfeasibly large sampling attempts. Is is also not possible to start by sampling w_A freely, then choosing w_B from the interval $[0, 1 - w_A]$, then w_C from $[0, 1 - w_A - w_B]$, etc. as this will skew the distributions for the parameters in favour of large values for the first sampled and low values for the last sampled. We hypothesize that the same approach should work, if the order in which the w_i are sampled is also randomized, but have not proven that this would lead to independent distributions when executed along a Markov-Chain.

The solution we have implemented instead relies on the fact that the sum rule (Equation 1) can be interpreted geometrically as the vector $\vec{p} = (w_A, w_B, w_C, w_D)$ lying on the four-dimensional hyperplane $E : x+y+z+w = 1$. If we start with a vector $\vec{p_0}$ already lying on that plane, we can ensure that we stay on that plane if our parameter step $\Delta\vec{p}$ is parallel to E, the sum $\vec{p_0} + \Delta\vec{p}$ is also in E and thus, the w_i fulfill the sum rule.

The problem now comes down to sampling a step vector $\Delta\vec{q}'$ in the xy plane and rotating it onto the plane E: $\Delta\vec{p} = M\Delta q$. The matrix M rotates the plane xy into E and can be found using singular value decomposition:

$$M_0 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \times \frac{1}{\sqrt{4}}(1, 1, 1, 1) \tag{7}$$

$$M_0 = U\Sigma V^T \tag{8}$$

$$M = U \times V^T \quad . \tag{9}$$

Equation 7 constructs a matrix from the normal vectors of the two planes we want to connect through rotation. The matrices U and V are used to construct the final rotation matrix M are found through singular value decomposition according to Equation 8 (the matrix Σ is not used).

A. Testing the Bayesian Approach

Despite the previously described way of efficiently sampling parameter space in a Markov-Chain, we have found after extensive testing that this Bayesian calibration approach is not able to reliable converge to the correct solution. We have tested this by generating model data with known parameter sets and treating them as our calibration data. If the calibration routine works, we should recover the known parameters that were used to generate the testing data. However, this was generally not the case. We suspect the reason in the fact that models with different sets of parameters are able to describe a given set of data points almost equally well, a phenomenon sometimes referred to as *equifinality* [5]. The Monte-Carlo nature of the Bayesian calibration means that the question of which optimum is found in a given run is a consequence of random number generation.

There are, however, still more constraints on the w_i than have been incorporated mathematically, thus far. The w_i reflect certain properties of the plant communities, some of which are, not strictly, but somewhat, mutually exclusive. E.g. functional group A defines plant species adapted to fertile sites and frequent defoliation, while group D describes plant species adapted to infertile sites and infrequent defoliation. It is therefore rather unlikely for a plant community to be composed of similar amounts of groups A and D. The same can be said about B and C. We have captured this fact algorithmically in the prior definition, by letting said combinations lead to lower scores. Even with this being accounted for, some equifinality remains rendering the MCMC approach impractical.

[1] M. Jouven, P. Carrère, and R. Baumont, Grass and Forage Science **61**, 112 (2006).

[2] P. Cruz, M. Duru, O. Therond, J. P. Theau, C. Ducourtieux, C. Jouany, R. Al Haj Khaled, and P. Ansquer, **172**, 335 (2002).

[3] D. Wallach, T. Palosuo, P. Thorburn, Z. Hochman, E. Gourdain, F. Andrianasolo, S. Asseng, B. Basso, S. Buis, N. Crout, C. Dibari, B. Dumont, R. Ferrise, T. Gaiser, C. Garcia, S. Gayler, A. Ghahramani, S. Hiremath, S. Hoek, H. Horan, G. Hoogenboom, M. Huang, M. Jabloun, P.-E. Jansson, Q. Jing, E. Justes, K. C. Kersebaum, A. Klosterhalfen, M. Launay, E. Lewan, Q. Luo, B. Maestrini, H. Mielenz, M. Moriondo, H. Nariman Zadeh, G. Padovan, J. E. Olesen, A. Poyda, E. Priesack, J. W. M. Pullens, B. Qian, N. Schütze, V. Shelia, A. Souissi, X. Specka, A. K. Srivastava, T. Stella, T. Streck, G. Trombi, E. Wallor, J. Wang, T. K. D. Weber, L. Weihermüller, A. de Wit, T. Wöhling, L. Xiao, C. Zhao, Y. Zhu, and S. J. Seidel, Environmental Modelling & Software **145**, 105206 (2021).

[4] "Bodeneignungskarte der Schweiz, Geodaten," (2002).

[5] K. Beven, Advances in Water Resources **16**, 41 (1993).