Multi-model simulation of soil temperature, soil water content and biomass in Euro-Mediterranean grasslands: uncertainties and ensemble performance

Sándor R.[a,b], Barcza Z.[c], Acutis M.[d], Doro L.[e], Hidy D.[f], Köchy M.[g], Minet J.[h], Lellei-Kovács E.[i], Ma S.[a,†], Perego A.[d††], Rolinksi S.[j], Ruget F.[k], Sanna M.[d], Seddaiu G.[e], Wu L.[l], Bellocchi G.[a,*]

[a] *Grassland Ecosystem Research Unit, French National Institute for Agricultural Research, Clermont-Ferrand, France*

[b] *MTA Centre for Agricultural Research, Institute for Soil Sciences and Agricultural Chemistry, Budapest, Hungary*

[c] *Eötvös Loránd University, Department of Meteorology, Budapest, Hungary*

[d] *University of Milan, Department of Agricultural and Environmental Sciences - Production, Landscape, Agroenergy, Milan, Italy*

[e] *University of Sassari, Desertification Research Centre, Sassari, Italy*

[f] *Szent István University, MTA-SZIE Plant Ecology Research Group, Gödöllő, Hungary*

[g] *Thünen Institute of Market Analysis, Braunschweig, Germany*

[h] *University of Liège, Arlon Environment Campus, Arlon, Belgium*

[i] *MTA Centre for Ecological Research, Institute of Ecology and Botany, Vácrátót, Hungary*

[j] *Potsdam Institute for Climate Impact Research, Potsdam, Germany*

[k] *French National Institute for Agricultural Research, Modelling Agricultural and Hydrological Systems in the Mediterranean Environment, Avignon, France*

[l] *Rothamsted Research, North Wyke, Okehampton, United Kingdom*

[†] *Currently at: University of New South Wales, Climate Change Research Center, Sydney, Australia*

[††] *Currently at: Catholic University of the Sacred Heart, Department of Sustainable Food Production, Piacenza, Italy*

[*] Corresponding author. 5 chemin de Beaulieu, 63039 Clermont-Ferrand (France); gianni.bellocchi@clermont.inra.fr

Abstract

This study presents results from a major grassland model intercomparison exercise, and highlights the main challenges faced in the implementation of a multi-model ensemble prediction system in grasslands. Nine, independently developed simulation models linking climate, soil, vegetation and management to grassland biogeochemical cycles and production were compared in a simulation of soil water content (SWC) and soil temperature (ST) in the topsoil, and of biomass production. The results were assessed against SWC and ST data from five observational grassland sites representing a range of conditions - Grillenburg in Germany, Laqueuille in France with both extensive and intensive management, Monte Bondone in Italy and Oensingen in Switzerland - and against yield measurements from the same sites and other experimental grassland sites in Europe and Israel. We present a comparison of model estimates from individual models to the multi-model ensemble (represented by multi-model median: MMM). With calibration (seven out of nine models), the performances were acceptable for weekly-aggregated ST (R^2 >0.7 with individual models and >0.8-0.9 with MMM), but less satisfactory with SWC (R^2 <0.6 with individual models and <~0.5 with MMM) and biomass (R^2 <~0.3 with both individual models and MMM). With individual models, maximum biases of about -5 °C for ST, -0.3 $m^3 m^{-3}$ for SWC and 360 g DM m^{-2} for yield, as well as negative modelling efficiencies and some high relative root mean square errors indicate low model performance, especially for biomass. We also found substantial discrepancies across different models, indicating considerable uncertainties regarding the simulation of grassland processes. The multi-model approach allowed for improved performance, but further progress is strongly needed in the way models represent processes in managed grassland systems.

Keywords: biomass, grasslands, modelling, multi-model ensemble, soil processes

1. Introduction

Grasslands are widespread vegetation types worldwide (about 40.5% of the Earth's landmass; Suttie et al., 2005), covering a large proportion of the European continent (67 million ha in the EU-27 that is 40% of agricultural land, 15% of total area, 85% of which being occupied by permanent grasslands, Peeters, 2012; Peyraud, 2013). Pastoral lands contribute to agricultural production and ecosystem services, including the provisioning of forage and, hence, of milk and meat (Huyghe, 2008). In addition, permanent grasslands are often hotspots of biodiversity (Marriott et al., 2004), which contributes to the temporal stability of their services.

Considering the role played by grasslands in maintaining food production, grassland biomass yield is an important agro-technical indicator to evaluate the economic viability of grassland-based milk and meat production systems as compared to concentrate feeding (e.g. Schader et al., 2013). In a climate-change context, for instance, adaptation of grasslands to climate change necessarily includes minimizing fluctuations in biomass produced (Collins, 1995). Considering the viability of grassland-based systems depending on their ability to produce meat from forage harvested on-farm, it is critical to examine the dynamics of grassland biomass production, where management plays a role by influencing the temporal forage availability and the interactions between herd and grassland.

Grassland ecosystem models have become important tools for extrapolating local observations and testing hypotheses on grassland ecosystem functioning (Chang et al., 2013; Graux et al., 2013; Vital et al., 2013; Ma et al., 2015). Under the auspices of the FACCE MACSUR knowledge hub (http://macsur.eu), a model intercomparison was conducted using datasets from an observational and experimental network of nine multi-year flux and production sites spread across Europe (France, Italy, Germany, Switzerland, The Netherlands, and United Kingdom) and Israel, and engaging a modelling community using a suite of different models to understand grassland functioning. In particular, the collected datasets of meteorological data, C, energy and water fluxes were used to drive and evaluate the performance of nine grassland models.

The identified models are an inventory of modelling approaches made available through the MACSUR consortium and applied worldwide. Grassland-specific approaches were used together with other approaches, mainly conceived to simulate crops and plant functional types. The primary goal of this study is to synthesize and compare the participating grassland models to assess current understanding of soil processes (soil temperature and soil water content, which are fundamental drivers of ecosystem-scale processes) and aboveground/harvested biomass (which is the output of major significance in agricultural production) in Europe and Israel. To achieve this goal, model evaluation against actual measurements was performed before and after model calibration. To the best of authors' knowledge, this is the first model intercomparison performed specifically on permanent grasslands. The present study, focused on grassland sites across Europe and a neighbour country (Israel), extends preliminary analyses (Ma et al., 2014; Sándor et al., 2015), and parallels other initiatives on the comparison of grassland models worldwide, such as the Agricultural Model Intercomparison and Improvement Project (AgMIP, Rosenzweig et al., 2013) and other international projects (Soussana et al., 2015).

The present grassland model intercomparison tries to answer five fundamental questions in a multi-site, multi-model framework: (1) are the main drivers of grassland processes represented well by state-of-the-art grassland models?, (2) what is the skill of the studied models considering the different processes?, (3) can calibration improve the models in terms of quality of simulation of different processes?, (4) can the ensemble of model results be used to estimate soil properties and grassland biomass in the study sites?, and (5) what uncertainties are associated with the different models, and how can uncertainty be quantified in a multi-model framework? In addition, areas are identified where structural changes in models may be needed to improve performances and decrease uncertainty of process representation.

2. Material and methods

2.1. Study sites

The nine long-term grassland sites used for the modelling exercise (Table 1) cover a broad range of geographic and climatic conditions (Fig. 1; see also Fig. A and Table A1 in the Supplementary material) as well as a variety of management practices (Table A2 in the Supplementary material).

Fig. 1. Geographic location (left) and classification (right) of grassland sites (black squares: grassland sites equipped with eddy covariance system; green circles: other grassland sites) with respect to De Martonne-Gottmann aridity index (De Martonne, 1942) and heat wave days frequency.

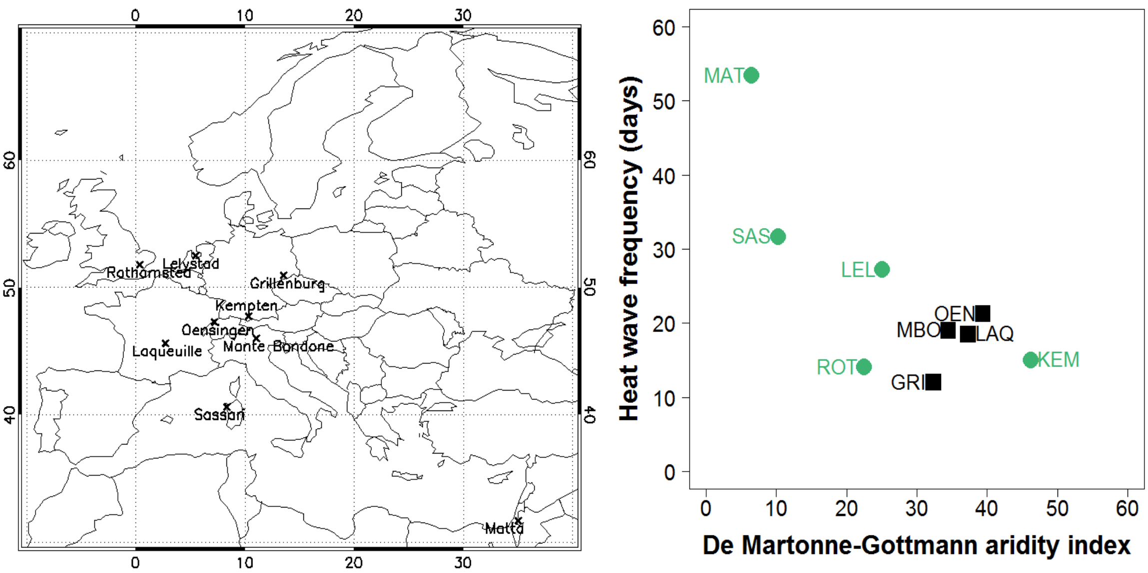


Table 1. List of permanent grassland sites.

Site	Latitude	Longitude	Elevation (m a.s.l.)	Years of available data	Notes	Source
Laqueuille (LAQ1, LAQ2), France	45° 38’ N	02° 44′ E	1040	2004-2010	Flux-tower grazed site, either intensively (LAQ1) or extensively (LAQ2) managed.	Klumpp et al. (2011)
Oensingen (OEN), Switzerland	47° 17′ N	07° 44′ E	450	2002-2008	Flux-tower mowed site, established on a ley-arable rotation.	Ammann et al. (2007)
Monte Bondone (MBO), Italy	46° 00′ N	11° 02′ E	1500	2003-2010	Flux tower Alpine hay meadow with occasional grazing in late autumn.	Wohlfahrt et al. (2008)
Grillenburg (GRI), Germany	50° 57° N	13° 30’ E	380	2004-2008	Flux-tower mowed, extensively managed site.	Prescher et al. (2010)
Kempten (KEM1, KEM2), Germany	47° 43° N	10° 20’ E	730	2004-2009	Experimental sward with different levels of N and cutting management (KEM1: four cuts per year; KEM2: two cuts per year).	Schröpel and Diepolder (2003)
Lelystad (LEL), The Netherlands	52° 30’ N	05° 28’ E	-4	1994-1998	Experimental sward with N management options.	Schils and Snijders (2004)
Matta (MAT), Israel	31° 42’ N	35° 03’ E	620	2007-2011	Dwarf shrubland in association with herbaceous annual species.	Golodets et al. (2013)
Rothamsted (ROT1; ROT2), United Kingdom	51° 48° N	00° 21’ E	128	1981-2011	Experimental sward with alternative N management options (ROT1: N-NH_4; ROT2: N-NO_3).	Silvertown et al. (2006)
Sassari (SAS), Italy	40° 39’ N	08° 21’ E	68	1983-1988	Mediterranean grassland dominated by annual self-seeding species.	Cavallero et al. (1992)

Four of the study sites (Laqueuille, Monte Bondone, Grillenburg, Oensingen) are equipped with an eddy covariance system to determine the net ecosystem exchange (NEE) of CO_2 and automated weather stations for hourly weather reports. They are essentially old semi-natural grasslands including vegetation types representative of the zone (with the exception of OEN, which was established in 2001). The flux-tower sites are the most data-rich grasslands in Europe, covering a variety of components of grassland ecosystem, including gross primary production (GPP), that is an estimate of the plant production of organic compounds from atmospheric CO_2, and ecosystem respiration (RECO), the latter playing an important role to estimate global C balances of terrestrial ecosystems (by definition NEE = RECO - GPP, with positive values indicating the system is a source of C, and negative values indicating that the system takes up C from the atmosphere). The flux-tower sites also record actual evapotranspiration, soil temperature (top 0.1 m) and soil water content (top 0.1 m). The eddy covariance system consists of a fast response 3D sonic anemometer coupled with fast CO_2-H_2O analysers measuring fluxes of CO_2, latent and sensible heat, and momentum fluxes at a 30-min time step. The basic data used in this study are at daily resolution to fit the temporal resolution of models. They are the result of a filtering process, quality check and gap filling according to European flux database guidelines (Aubinet et al., 2012). Data are also available on the standing aboveground biomass at given dates. Biomass was measured destructively at given dates in all the study sites (at ground level at Laqueuille, at site-specific canopy heights as part of regular mowing in the other sites).

Other grassland sites (Kempten, Lelystad, Matta, Rothamsted, Sassari) are from experimental research, with focus on forage production under a range of conditions, and for which weather inputs are available on a daily time step. These sites provide forage yields, i.e. the amount of dry matter biomass that is removed from the field at each cutting event that corresponds to removal of C and nitrogen (N) from these grassland systems. Each of these sites offer the possibility to model different grassland systems while expanding geographical coverage and the variety of management options tested.

2.2. Models description

The first phase of the study was to identify a wide selection of grassland models to be able to represent processes controlling energy, water and C cycle dynamics. The selection phase allowed identifying nine models in which processes are represented with different levels of detail. Whereas some models are empirically based with relatively simple relationships between driver variables and fluxes, others are more complex, simulating the coupled C, nutrient, and water cycles (process-based models). Models also differ in their representation of soil properties, vegetation type, farming practices, and environmental forcing, as well as the initialization of C pools.

Here we divide the models into three categories based on their feature sets. Three models - AnnuGrow, PaSim and SPACSYS - were specifically developed to simulate grasslands. Three models - EPIC, STICS and ARMOSA - were originally developed to simulate annual crops and include options for grassland systems. Other three models - Biome-BGC MuSo, CARAIB and LPJmL - that simulate different vegetation (or biome) types, including grasslands, were also included in the exercise. Supplementary material contains a brief description of the models and a synoptic table (Table B1) of the main processes implemented. The types of outputs generated by the models are in Table B2 (Supplementary material). The model results are presented anonymously in the paper, as the identification of models providing a specific performance is out of scope.

2.3. Simulation study design

Model simulations were carried out independently by the modelling groups (which included developers, expert users or end-users) using their own infrastructure and technical background, as harmonizing the calibration techniques was out of scope of the intercomparison. Models were evaluated with data from the study sites before and after calibration.

For the uncalibrated (blind) simulations, the models were run at each site using the available data of weather, soil and management, with no parameter adjustment. After the blind simulations were completed, additional plant and soil information from a sub-set of flux-tower site data was supplied to each modelling group, i.e. the first half of the whole series of available data or the first half plus one in the case of an uneven number of years (Table 1). The information provided were daily time series of GPP, RECO, soil water content, soil temperature, and actual evapotranspiration (some groups only used a subset of observations for calibration). For the same output variables, calibrated simulation results were evaluated against observations from the validation sub-set of years. Biomass data were not used for calibration and held back for validation purpose.

It was requested that each modelling group adjusts model parameters (especially vegetation parameters) to improve the simulations based on the observed data, using whatever techniques they normally use, and documenting the changes. Summary of the model parameters that were considered for calibration is presented in Table C of the Supplementary material.

Seven groups completed the full assessment of that step. Simulation results from the blind tests over the calibration time period were compared with the measured data over the same period. For both tests, model outputs including biomass (measured at given dates in all the sites), soil temperature and soil water content at 0.1 m depth (both measured continuously on a daily basis at flux sites) were compared against observed values, since other output variables were not common to all the models. The agreement between simulation and observations was evaluated by the inspection of time series graphs and, numerically, through a set of performance metrics (Table D in Supplementary material).

Performance metrics were calculated for four time series: uncalibrated (*U1*, *U2*), calibrated (*C*) and validated (*V*) years. *U1* and *C* refer to the first half of the whole series of available data (or the first half plus one in the case of an uneven number of years) which was used for calibration, while *U2* and *V* refer to the years which were excluded from calibration. Possible improvement of model performance due to calibration was evaluated using the metrics from the *U2* and *V* years. This logic was used because validation implies that model performance is assessed with calibration-independent data. Thus, possible improvement of model performance can be most clearly judged by comparing error measures from *U2* and *V*. Multi-site mean (i.e. average data from all sites) error statistics were analysed to quantify the overall effect of model calibration on the simulated processes.

2.4. Uncertainty assessment

We assessed the models in terms of quality of simulations, by first focussing on the quantification of model errors with statistical indicators, and then using these errors to assess the uncertainty of the individual models in comparison with the multi-model ensemble. The modelling groups provided deterministic model simulation results according to the protocol established, which means that one run was provided for one site. It also means that the spread of model results due to parameter uncertainty was not specifically addressed as it would have dramatically increased the model output database used within the study. As uncertainty cannot be associated to any of individual simulations, we focussed on model residuals to quantify uncertainty. Residuals (simulation-measurement differences) were used in a standardized form (divided by standard deviation) to estimate variability for the individual models, and for the multi-model ensemble. Here we tried to assess whether the multi-model error has smaller variability than the individual models or not. The spread (maximum minus minimum) of simulation results (uncertainty with the ensemble spread) was also standardized (divided by standard deviation) to obtain a metric comparable with the standardized residuals of each model. Given the internal logic of biophysical and biogeochemical grassland models, errors in the estimation of internal processes propagate to the estimation of biomass and related output. Thus, we also quantified the relationship between standardized model residuals of ST, SWC and biomass, based on the calibrated simulations. ST and SWC residuals were calculated by averaging the residuals of two weeks preceding biomass sampling events. Moreover, we quantified the relationship between the residuals and mean maximum temperature and precipitation sum values of the preceding two weeks relative to the biomass sampling.

3. Results

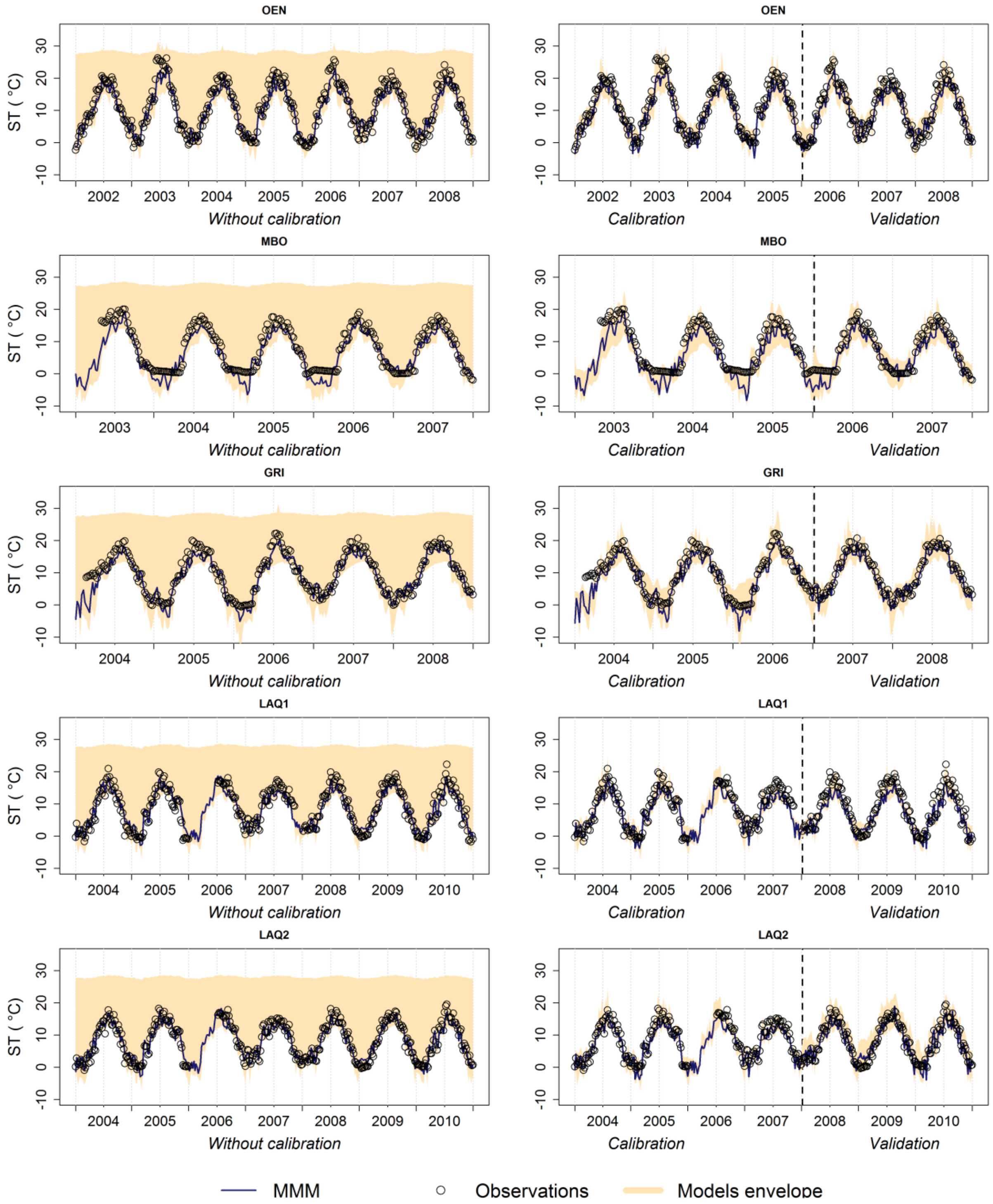
3.1. Analysis of individual model performance

Performance of the individual models is discussed according to the simulated output of interest. In order to assess the utility of using multi-model ensemble for the simulation of grassland functioning, performance of the multi-model simulation range and median is also assessed against measurement data. We used median instead of mean values in order to reduce the impact of outliers in the multi-model ensemble construction. For easier interpretation, weekly-aggregated data were used to quantify the overall measurement-model agreement (Supplementary material, section 3, provides additional information in daily and monthly resolutions). The identities of models were kept anonymous by using model codes from 1 to 9 (the order of models being not identical with the one used in Table B2, Supplementary material).

3.1.1. Evaluation of soil temperature (ST) estimates (flux sites)

Fig. 2 shows the range of model results (represented by the shaded area) and the multi-model median (MMM hereinafter) together with the measured values at weekly resolution (see also Figs. B and C of Supplementary material with daily and monthly time resolutions, respectively).

Fig. 2. Comparison of weekly averaged simulated and measured soil temperature (ST) at the flux sites (ID as in Table 1). The shaded area represents the range of estimations provided by the individual models while solid line shows the multi-model median (MMM). Open circles show the weekly averaged measured values. The dashed vertical line divides the measurement period into calibration and validation time series.



The figure suggests that the range of model results decreased drastically after calibration. However, it is worth noting that the upper bound in Fig. 2 (left) (almost constant ST around

28 °C) is caused by model 8 only, which did not provide results for the calibrated simulations. The rest of the models provided ST values in a more realistic fashion (not shown here).

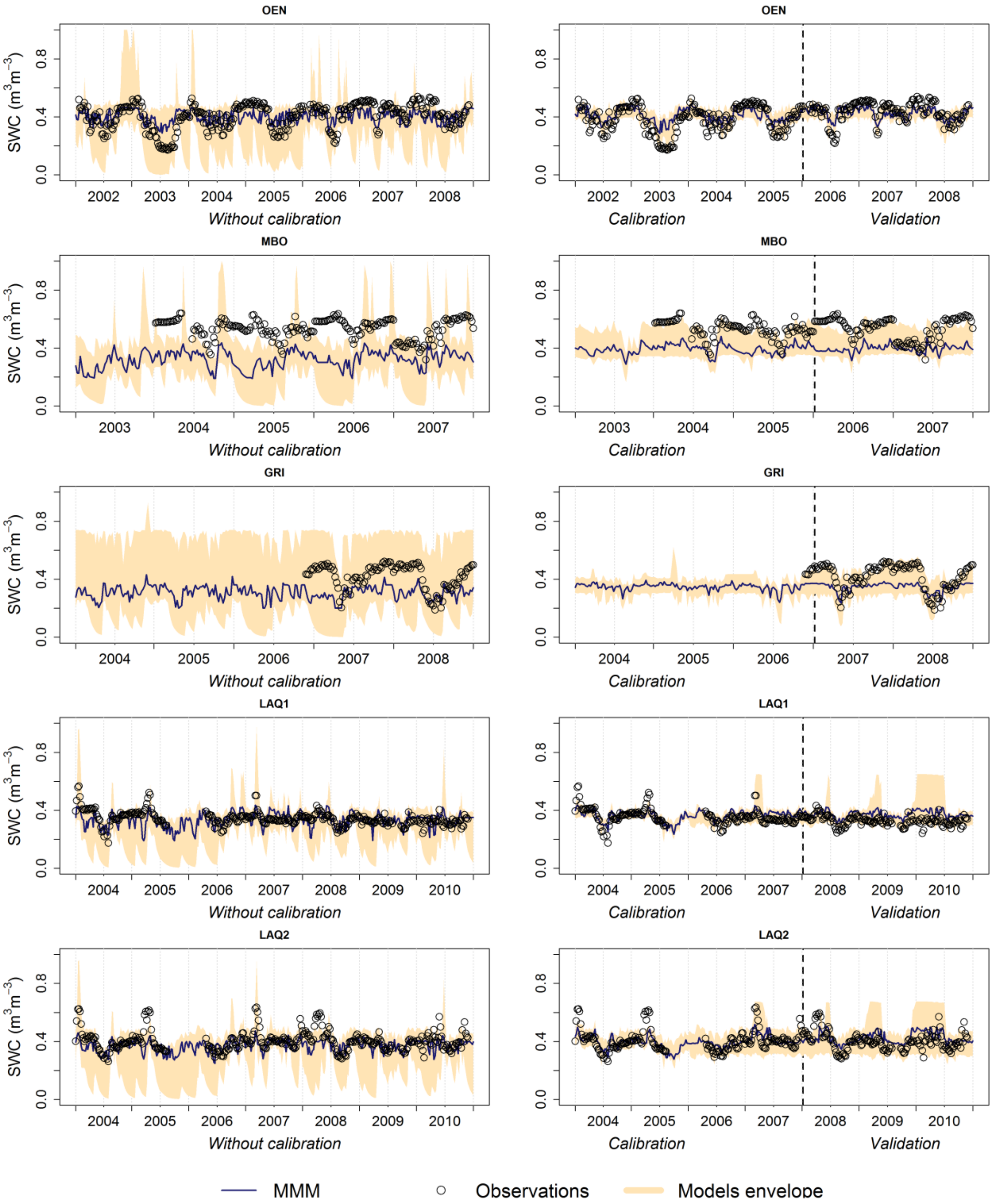
Scatterplots with weekly resolution (Figs. D-H in Supplementary material) show the improvements obtained with calibration, with a similar pattern across flux sites. Appendix 1 shows the statistical assessment of the model results for GRI and LAQ1, Grillenburg and Laqueuille being the driest and the wettest of the flux sites investigated, respectively (see other sites in Tables E-G of Supplementary material with weekly resolution).

Overall, calibration improved the quality of the ST simulation in terms of explained variance though the improvement is only marginal in some cases. In general, model performance was similar for calibration and validation periods for the seven models that provided both blind and calibrated results.

3.1.2. Evaluation of soil water content (SWC) estimates (flux sites)

Fig. 3 shows the comparison of measured and simulated SWC at weekly aggregation, for all five flux measurement sites (see Figs. I and J with daily and monthly time resolutions, respectively, in Supplementary material). The grey area provides information on the range of model results (nine models for the blind tests, seven of them for the calibrated tests), and the black line represents the MMM.

Fig. 3. Comparison of weekly averaged simulated and measured soil water content (SWC) at the flux sites (ID as in Table 1). The shaded area represents the range of estimations provided by the individual models while solid line shows the multi-model median (MMM). Open circles show the weekly averaged measured values. The dashed vertical line divides the measurement period into calibration and validation time series.



Blind simulation results indicate that some of the models gave unrealistically high and/or low SWC values. Given the soil texture at the sites, saturated SWC was not expected to stretch beyond ~0.52 $m^3 m^{-3}$ at any of the sites (as estimated by the SOILarium software from pedotransfer functions; Wösten et al., 1999; Fodor and Rajkai, 2011). The range of uncalibrated results had unrealistically high values of SWC. This was true at each site, but especially at GRI, characterized by the lowest clay and highest silt contents (Table 1). The lowest expected SWC (wilting point) is around 0.3 $m^3 m^{-3}$ at OEN and about 0.10-0.16 $m^3 m^{-3}$ at the other sites. Though the actual SWC can drop well below the wilting point in the upper soil layer, the lower boundary of SWC around zero at each site is not realistic considering that the flux sites are relatively wet. Comparison of uncalibrated and calibrated SWC shows that model parameter adjustment clearly improved the performance of the models (Fig. 3 right). The models mostly provided data within the expected SWC range, with no values beyond levels of SWC. The most prominent improvement was at GRI. At both LAQ1 and LAQ2, calibration introduced positive biases in some years (where uncalibrated biases were low).

Figs. K-O (Supplementary material) show the performance of the individual grassland models for both blind (nine models) and calibrated simulations (seven models). The results clearly show that systematic errors are present in all models. An interesting common error of the models is that the range of simulated SWC values is smaller than in reality (model 8 is exception). The scatterplots in Supplementary material also reveal that the above-mentioned wide range of model results (e.g. Figs. K1 and K2 for Oensingen) is caused by model 8 alone (in Fig. K2, the x- and y-axis ranges are smaller than in Fig. K1 because of the smaller overall range of SWC values.). The scatterplot indicate some improvement (remarkable with models 5 and 6) in the simulation of SWC in terms of R^2. However, model calibration was globally unable to address the systematic errors present in the blind tests.

Appendix 2 shows the performance indicators of the model results, for GRI and LAQ1, which are the driest and wettest site among the flux sites, respectively (for other sites, see Tables H-J of Supplementary material with weekly resolution). In general, high variability of changes was observed across sites for the models. Overall, none of the models under study revealed considerable improvement. SWC simulation was the most successful at GRI and OEN. At these sites, ME values up to 0.8 were obtained in some cases, with mostly negative values obtained in the other sites. It is evident that SWC representation is not satisfactory in spite of parameter adjustments. This means that all of the studied models have difficulties at the eddy covariance sites, which are all characterized by ample precipitation and lack of severe drought stress.

3.1.3. Evaluation of plant biomass estimates

Fig. 4a, b shows the comparison of measured and simulated biomass values for a dry and a wet site (SAS and KEM1; KEM2 is not shown) over the full measurement period (for the other sites, see Figs. P1-Q5 in the Supplementary material).

The shaded area represents the full range of model results (all nine models provided data for the blind tests, but only seven of them contributed to the calibrated tests), and the black line shows the multi-model median. The figures show that simulated biomass from the blind simulations varied in a wide range at all experimental sites. In general, measured biomass was within the range that was defined by the ensemble of the models. After calibration, the range of model results decreased for all sites except for MAT. As models 8 and 9 did not provide data for the calibrated simulations, it is not clear whether this decrease is the result of the calibration or it also incorporates the smaller number of models considered. For nine sites (SAS, KEM2, LEL, ROT1, ROT2, GRI, LAQ1, LAQ2, OEN), some of the measured data were outside the range that was defined by the seven models.

Fig. 4. Comparison of simulated and measured yield biomass (harvested aboveground biomass) at (a) SAS and (b) KEM1 sites (ID as in Table 1): without calibration (top) and with calibration (bottom). The shaded area represents the range of estimations provided by the individual models while solid line shows the multi-model median (MMM). Black circles show the measured yield biomass values.

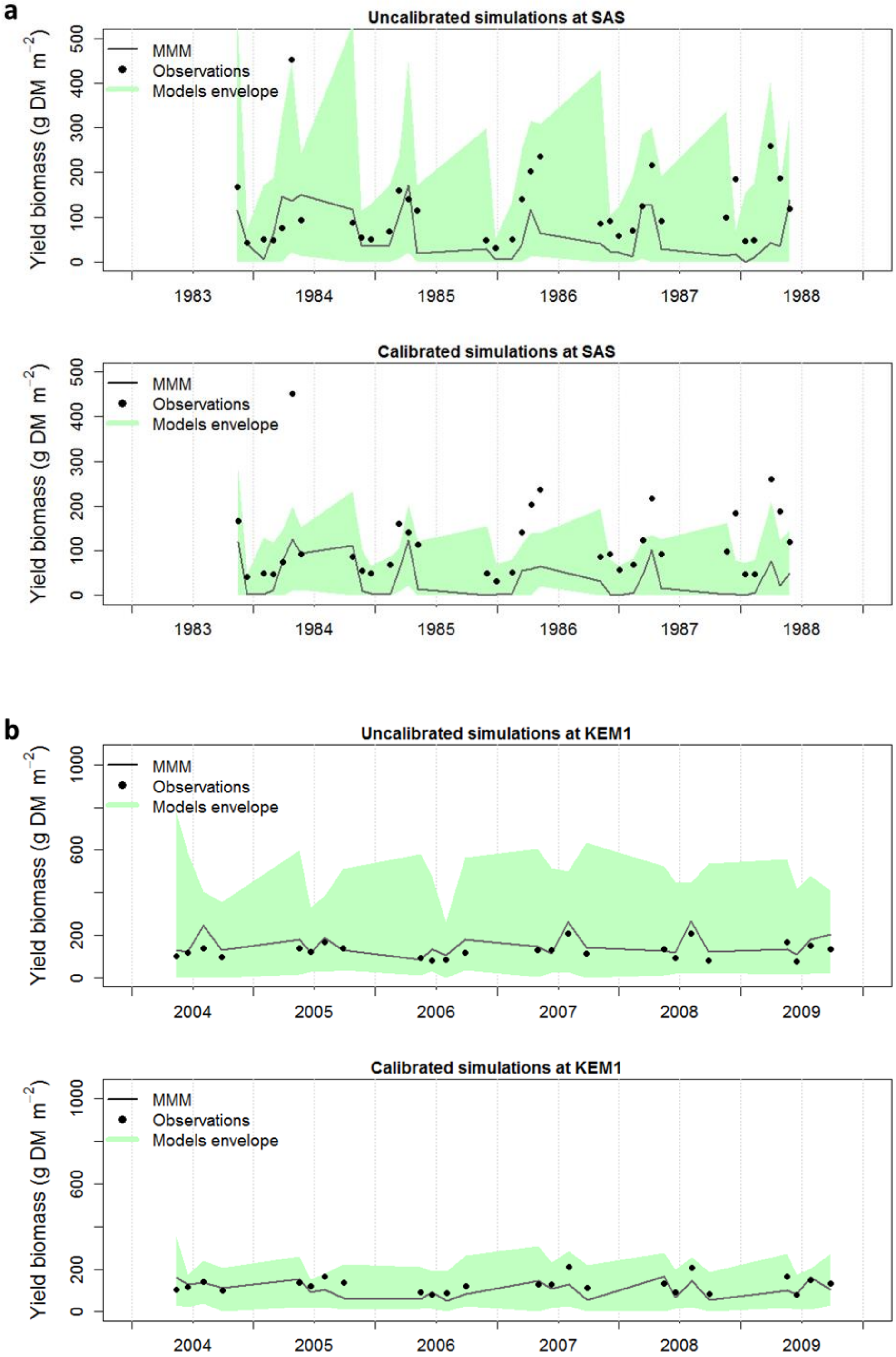


Fig. R (Supplementary material) shows the performance of the individual grassland models for the blind and the calibrated simulations, separately for the dry and wet site (SAS and KEM1, respectively; see also Figs. S1-S20 in the Supplementary material for the other sites), revealing that the performance of the grassland models is rather heterogeneous, and varies considerably between sites and models.

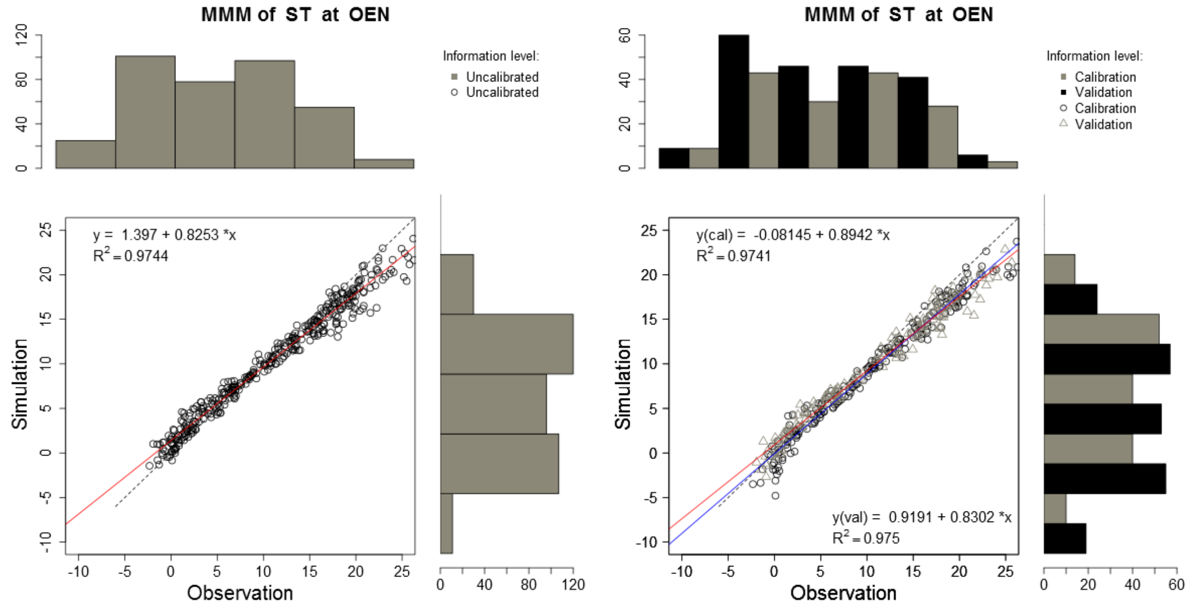
Overall, considering all sites and models (see also Supplementary material, Figs. Q1-S20), underestimation of biomass is more common than overestimation. Data points are distributed around the 1:1 line for ~1/3 of all model-site combinations that reported results. There is no clear systematic behaviour for the models in terms of over- or underestimation with a few exceptions. After calibration the overall picture changed to some extent: underestimation decreased, and tendency to approach the 1:1 line improved slightly. Percent of model-site combinations that provided data near the 1:1 line increased to some extent. Explained variance of the models (not considering MBO, due to the limited number of data points) varied in a wide range, spanning the interval of 0.00-0.78 for the blind runs, and 0.00-0.98 for the calibrated simulations.

For biomass, Appendix 3 shows the statistical evaluation of simulation performances at SAS and KEM1, for the uncalibrated and calibrated models separately (other sites in Tables K-T in Supplementary material). In this case, there is no distinction between *U1* and *U2*, and also *C* and *V* years, as yield data were not used for model calibration. Data from OEN were excluded from this analysis due to the low number of samples. High variability of changes in statistical indicators can be detected based on Table 4. Multi-site mean ME was negative for all models. There was no systematic fashion in the change of ME between the sites. In spite of the improvement of ME, the calibrated, multi-site mean ME was still negative for all models, which reflects poor model performance. The largest calibrated ME is characteristic to model 7 (multi-site mean ME is -2.57).

3.2. Analysis of the ensemble approach

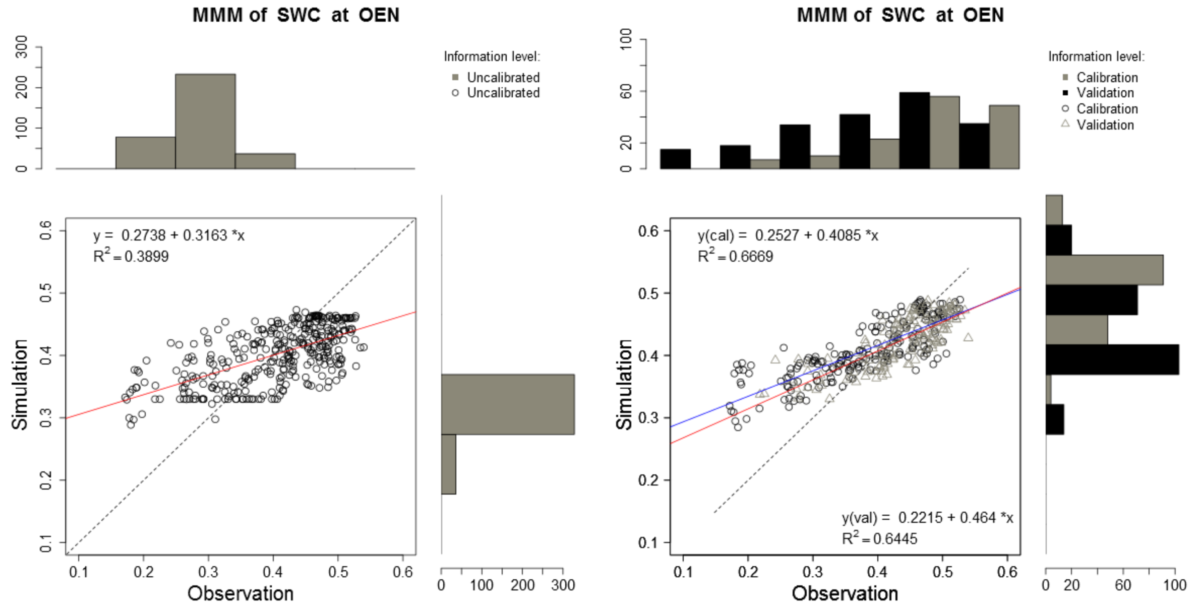
Fig. 5 shows the MMM (or in other words, ensemble), uncalibrated and calibrated-validated ST simulations compared with observed values on weekly resolution at OEN (see, for other sites, Figs. T1-T4 in Supplementary material).

Fig. 5. Multi-model median (MMM) of uncalibrated (left) and calibrated-validated (right) soil temperature (ST) simulations compared with observed values with weekly resolution at OEN site (ID as in Table 1): x-y scatterplots with associated x and y histograms with estimated densities.



The figures indicate that MMM ST from the blind simulations provided reliable estimates in terms of explained variance and slope of the linear regression. Explained variance varied between 91 and 97%, while the slope varied between 0.83 and 0.92 (which means small underestimation by the ensemble). Calibration did not change the overall quality of the MMM. Explained variance changed slightly with very small overall decrease, while the slope became closer to the 1:1 line in some cases. The performance indicators were calculated using the *U2* and *V* years only. Considering ME, the MMM ST taken from the blind runs was a better predictor than 62.5% of the models. After calibration, 71% of the models gave worse ME than the MMM. Considering the explained variance, blind MMM ST was better than any of the models, while after calibration 86% of the models provided worse performance than the ensemble median. Fig. 6 shows the comparison of the measured and the simulated MMM SWC results (separately for the uncalibrated and the calibrated-validated runs) at OEN, which is the best site in terms of MMM SWC performance (see, for other sites, Figs. U1-U4 in Supplementary material).

Fig. 6. Multi-model median (MMM) of uncalibrated (left) and calibrated-validated (right) soil water content (SWC) simulations compared with observed values with weekly resolution at OEN site (ID as in Table 1): x-y scatterplots with associated x and y histograms with estimated densities).



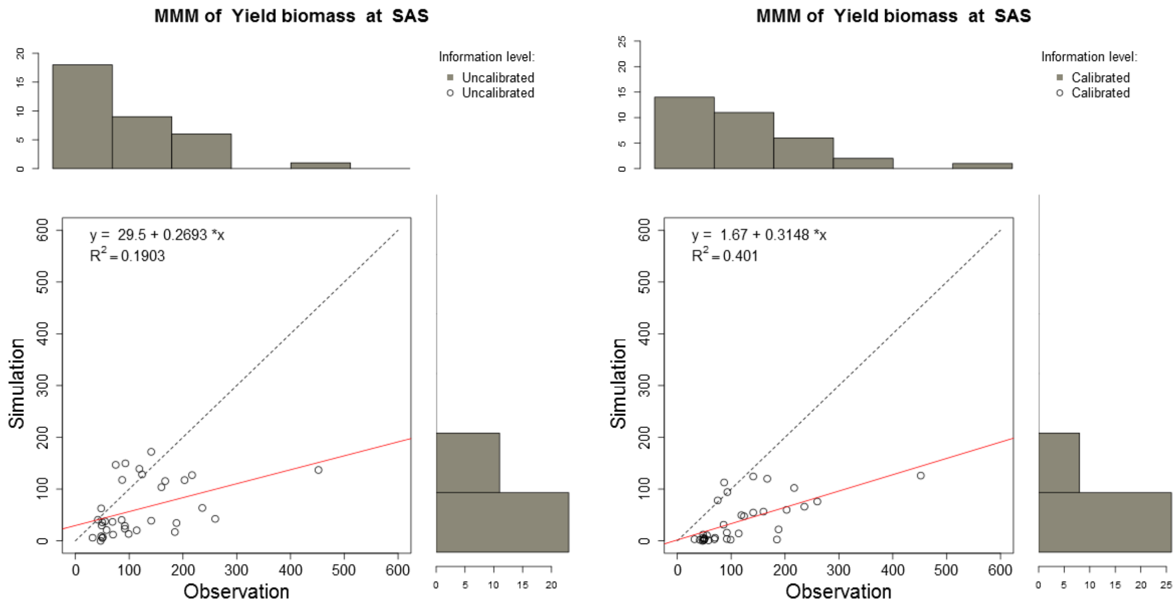
The results indicate that MMM SWC inherits the problems associated with the individual models. MMM SWC constructed from the blind simulation results shows poor performance at all sites. Low explained variance (maximum R^2 ~0.4 at OEN) and departure of the data from the 1:1 line are indicators of the low reliability of simulations. The range of simulated ensemble SWC values is smaller than in reality, similarly to the results obtained with the individual models. After calibration, the quality of the MMM SWC simulations was mainly improved, though the performance of the validated and calibrated years differed markedly in some cases. Explained variance increased for all five sites, and ranged between 11% (LAQ2, validated years) and 73% (OEN, calibrated years). The simulated MMM SWC remained confined within a relatively narrow range for all sites, which means that the intra-annual variability of SWC was not captured by the MMM. Similarly to ST, multi-site mean error statistics were calculated and compared with the multi-site mean statistical indicators of the MMM SWC (for the *U2* and *V* years). ME of the MMM SWC was better than 78% of the models and 57% of the models for the blind and calibrated simulations, respectively. Multi-site mean ME remained negative for all models in both time periods (*U2* and *V*), which means that the mean of the observations is more useful for SWC estimation than any of the models.

Fig. V (Supplementary material) shows that after calibration better estimations in yield were reached at the grassland sites other than the flux sites. In general, the MMM underestimated the expected yield at the production sites but overestimated it at the flux sites. Additionally, the observed yield was poorly represented at those sites characterized by extensive treatments (LAQ2, KEM2, ROT2).

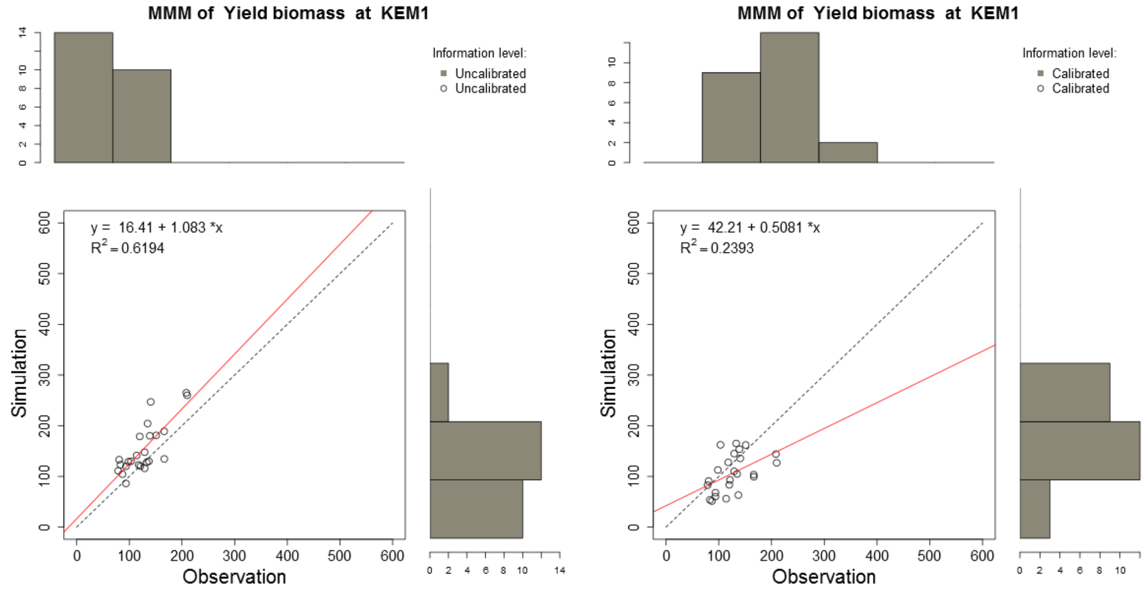
Fig. 7a, b shows the observed and the modelled ensemble (MMM) biomass data for SAS and KEM1 (Figs. W1-X5 in the Supplementary material present the results for the other situations, considering that MBO is not discussed due to the low number of data).

Fig. 7. Multi-model median (MMM) of uncalibrated (left) and calibrated (right) yield biomass simulations compared with observed values at the arid SAS site (a) and the humid KEM1 site (b) (ID as in Table 1): x-y scatterplots with associated x and y histograms with estimated densities.

a



b



The figures indicate that the performance of the MMM biomass estimation changed from site to site. Interestingly, the pattern on the scatterplots is similar for the blind and calibrated ensembles, which means that parameter adjustment did not cause radical change on the overall performance of the multi-model ensemble. With a few exceptions, systematic over- or underestimation is typical. Explained variance varies considerably among sites. With respect to ME, MMM outperformed the individual models in 100% of the cases. In terms of R^2, the MMM gave better explained variance than seven out of the nine models (78%) for the blind runs, while MMM outperformed five models (out of seven) for the calibrated simulations (71%).

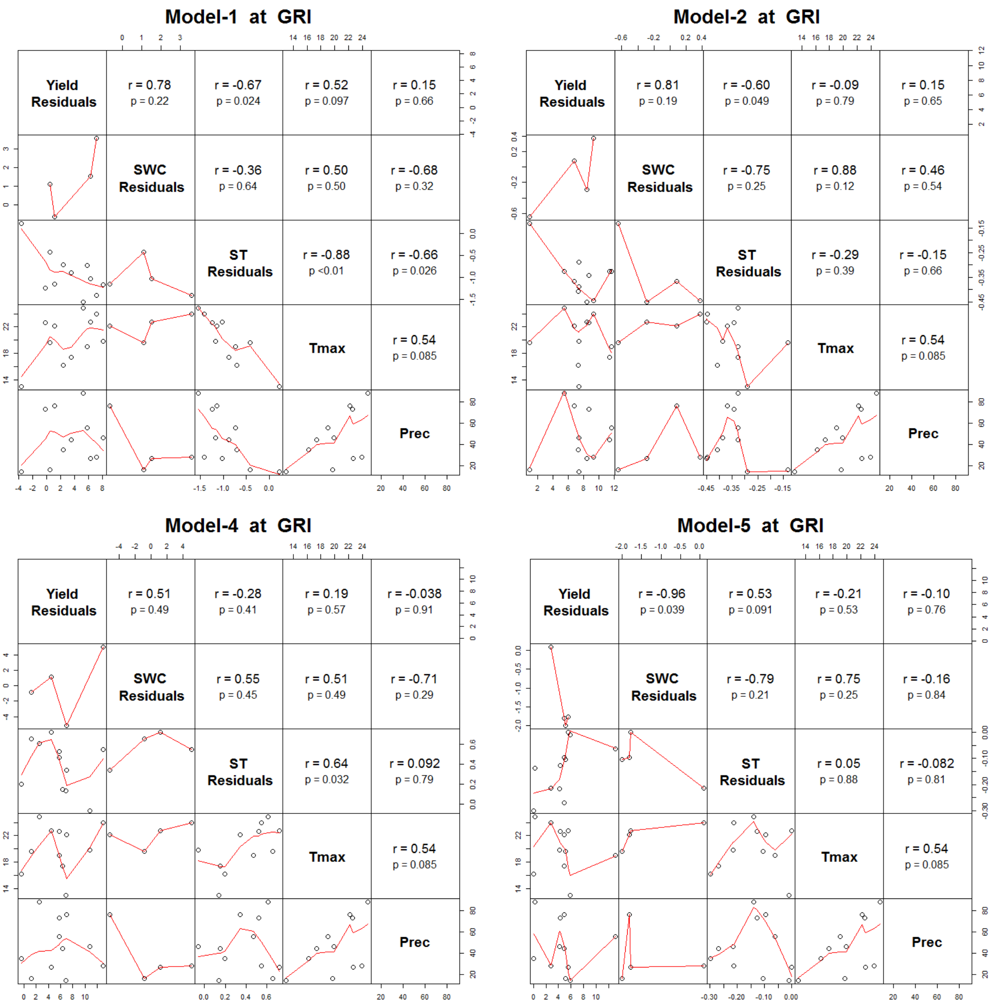
3.3. Relationship between model errors and uncertainty assessment

3.3.1. Relationship between residuals

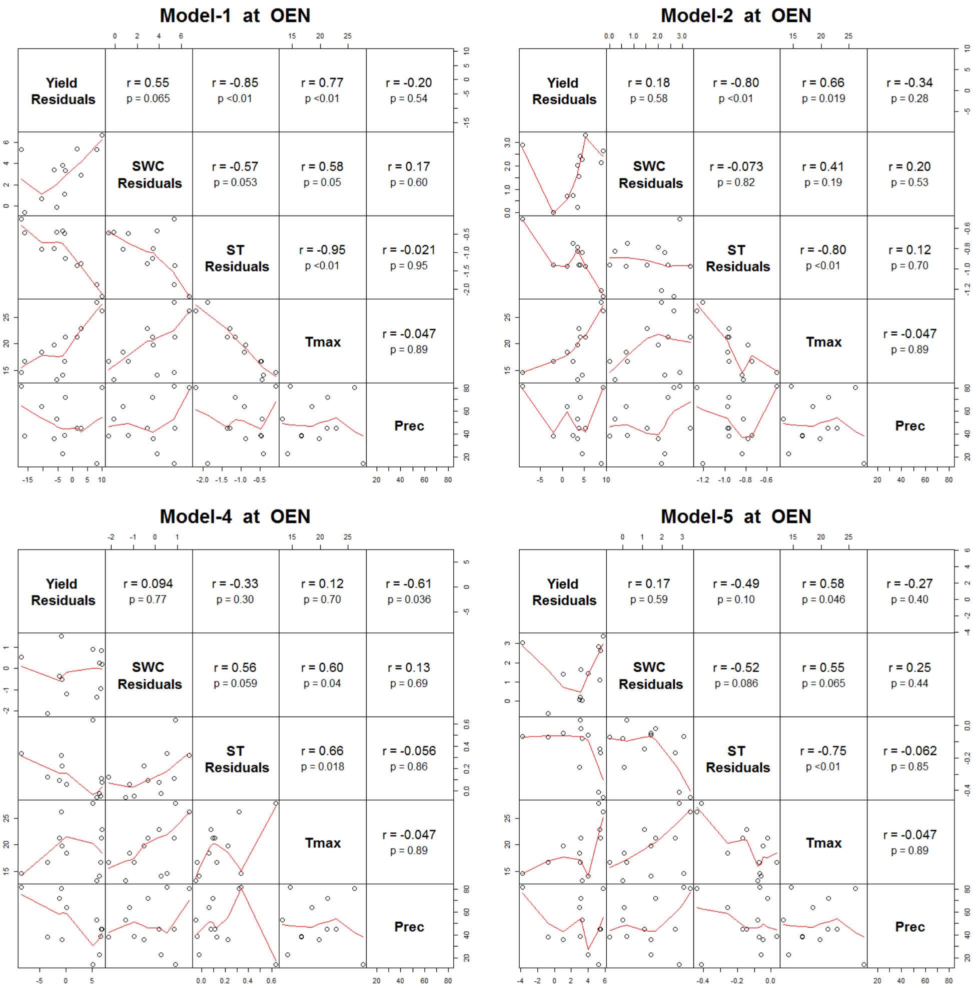
Due to data availability, the analysis of the relationship between standardized residuals was restricted to four eddy covariance sites (at MBO the number of biomass data was too low). Models 1, 2, 4, 5, 6 and 7 provided all data needed to analyse the residuals in this fashion (other models reported data to only a subset of the flux sites). Fig. 8 shows the relationship between the selected variables for OEN and GRI for models 1, 2, 4 and 5. Supplementary material contains results for other sites and models (Figs. Y1-Y5).

Fig. 8. Correlation between the standardized residuals of simulated yield biomass (cutting events) of models 1, 2, 4 and 5, soil water content (SWC), soil temperature (ST), maximum temperature (mean of the two weeks before cutting) and precipitation (total of the two weeks before cutting) at GRI (a) and OEN (b) sites (ID as in Table 1).

a



b



The figures visualize the relationship between the selected variables as squared matrix-like configurations. The lower triangular part of the squared matrices shows the scatterplots between the specific variables defined in the main diagonal of the matrix, with the overlying spline (without inferential character). For readability, the correlation between the variables and the significance of the relationship (p value) are shown in the upper triangular part of the matrix. The figures show that at some sites (mostly at GRI and OEN) a relatively strong relationship exists between some of the residuals, and also between the environmental factors and the residuals (relationship between maximum temperature and precipitation is not informative in the present context). The existing relationship is not uniform and, in some cases, the correlation is negative between some of the residuals (e.g. relationship between yield and SWC residuals at GRI for model 5). Considering that the number of available SWC residuals at GRI is low, the statistical comparison is not well justified here for SWC.

In the followings, we focus mainly on GRI and OEN sites. The individual models show considerably differences in terms of relationship between the yield, the SWC and the ST standardized residuals. High positive correlation was established between the yield and SWC residuals for models 1, 2 and 4, whilst models 5 and 6 had a strong negative correlation at Grillenburg, which is the northern flux site (Fig. 8 a and Fig. Y1 in Supplementary material). Similarly, positive correlation characterizes the relationship between yield and SWC residuals at OEN, but the relationship is weaker than at the GRI site (Fig. 8b and Fig. Y1 in Supplementary material). We found a general negative correlation between the yield and ST residuals, with the exception of models 5, 6 and 7 (Fig. Y1 in Supplementary material), as well as between the ST and SWC residuals (except for model 4) at all sites (the correlation was moderate at the grazed sites; see Figs. Y1 and Y2 in Supplementary material). Meteorological factors such as the mean maximum temperature and precipitation (2-weeks means and totals, respectively) also had a notable effect on the residuals. In some cases there was no clear pattern among the sites. The relationship between the selected variables can be alternatively characterized as well. We can select an arbitrary (but high enough) absolute minimum threshold and identify the number of cases when the covariance equals or exceeds this expected minimum in absolute terms. Selecting the 0.66 correlation threshold (which represents ~44% explained variance), and considering only OEN and GRI, the most common relationship is the ST residual - maximum temperature, which is typical for models 1, 2, 4, 5 and 6. The second most common feature is the SWC residual - yield residual relationship, which is present in the case of models 1, 2, 5 and 6. Strong precipitation - SWC residual, maximum temperature - SWC residual and ST residual - SWC residual relationships are present for three models. Maximum temperature - yield residual and ST residual - yield residual relationships were strong for two models. The correlation between the other possible variable combinations did not reach the 0.66 threshold for GRI and OEN. Though the multi-model medians of ST, SWC and yield are statistically-derived datasets, and not the result of a process-based model, it might be interesting to check their behaviour in terms of correlation between MMM residuals, and also the effect of environmental variables on the residuals. The MMM correlations were generally moderate probably owing to the decreased model uncertainty (Fig. Y5 in Supplementary material). We found a general negative correlation between the SWC and ST residuals, while the maximum temperatures were positively correlated with the SWC and negatively with the ST residuals at all sites (the highest correlation was characteristic to the GRI and OEN sites). These results are in accordance with our previous finding, namely that the MMM approach may give a better estimation than the individual models (here in terms of unexpected correlation between the residuals).

3.3.2. Uncertainty assessment related to multi-model ensemble

Appendix 4 shows, for both individual models and MMM, the ratios between the variability of the models envelope and standardized model residuals. Values greater than one indicate that the spread is larger than the model residual, i.e. the uncertainty associated with the ensemble of models is high. For ST, ratios >1 indicate that with both individual models (90%) and MMM (100%) model error was generally lower than the variability in the multi-model ensemble (with ratio equal to 1, M1 at LAQ1 is the only exception). With SWC, the pattern of responses is more complex, ranging from ratios <1 with M1 at all sites to ratios >1 with M6 and M7, and mixed situations with the other models and MMM (overall ratios >1 are 68% with individual models and 60% with MMM). This complexity is also reflected in the yield responses (ratios >1 are 54% with individual models and 58% with MMM), where only

M3 shows ratios <1 at all sites expect MBO (where only two values of measured biomass were available).

4. Discussion

4.1. Soil temperature (ST)

All the models simulated ST relatively well, and their performance for representing ST generally improved after calibration. However, modelling efficiency (ME, at times <0) indicated problems with the quality of the results. It means that the information content of the simulations is questionable in spite of the level of explained variance, which appears high. Therefore, developments are still needed in terms of ST representation of the models to improve the quality of the simulations. Error statistics show the utility of the ensemble ST simulations against individual models. Ensemble median ST based on the blind runs over-performed the majority of the models (except in terms of ME), while ensemble median ST derived from the calibrated runs was still more appropriate than ~2/3 of the models. The results indicate that satisfactory results can already be acquired based on the ensemble of uncalibrated runs.

4.2. Soil water content (SWC)

Even though bias can exist in the measurements of SWC (e.g. in the case of the widely used water content reflectometers; Weitz et al., 1997; Chow et al., 2009), performance indicators clearly indicated that the models used in this study are not sufficiently accurate to estimate SWC. This was mainly associated with the unrealistic small amplitude of the annual cycle of the SWC curve, as compared to the measurements. Due to the known role of SWC on evapotranspiration, stomatal conductance and other processes, this problem has obvious consequences at sites where water shortage is a typical feature. According to the De

Martonne-Gottmann aridity index (Supplementary material, Fig. A), water shortage affected the majority of the sites, at least in some years. Proper response of the models to the water-limited conditions is thus questionable, which means that the applicability of the models in semi-arid or arid ecosystems is not supported.

This finding may be to some extent related to the ability of roots to extract soil water, which differs between perennial species dominating continental Europe and annual self-seeding species dominating Mediterranean (semi-arid) sites (e.g. Volaire and Lelièvre, 2001; Mapfumo et al., 2002).

Quality of SWC simulation might seriously affect model parameter estimation as well. Calibration usually means a statistical method where the internal model parameters are adjusted, so that the agreement between model outputs and measurements is improved (e.g. Hidy et al., 2012). The pitfall of model calibration is the possible bias introduced to the optimized internal parameters when model structural errors are compensated with distorted parameters (e.g. Carvalhais et al., 2008; Martre et al., 2015). This is especially problematic if the model parameters are physical quantities (like C:N ratio, specific leaf are index, etc.) not merely coefficients of some empirical equation. Our results indicate that due to the deficient SWC estimation there is a high possibility that calibration will result in distorted parameter values. Further model developments are clearly and essentially needed in terms of soil hydrology to address structural errors within the models, and to avoid the systematic errors associated in some of the model parameters.

The utility of the MMM SWC estimation is not as straightforward as in the case of ST. Ensemble median of the blind results usually performs better than 2/3 of the models (with the exception of R^2), which means that some benefit can be expected by using an ensemble approach. Considering the calibrated models, the number of models that are outperformed by the median is decreased. These results indicate the usefulness of the ensemble approach though the performance of the MMM still indicates several areas of improvement. In summary, the results indicate that SWC estimation should be used with caution in regional or continental scale simulations, and model developments focusing on soil hydrology are essential.

4.3. Plant biomass

Biomass data are discontinuously measured and rather large uncertainties on biomass measurements (mainly owing to spatial heterogeneity) may hinder model evaluation (Vuichard et al., 2007). Simulated yield dynamics were essentially dissimilar across the models used in this intercomparison. The results indicate that there is no systematic fashion in the response of the models to the environmental factors. This highlights the complexity of interactions between meteorology, soil properties, grassland floristic composition and their related resilience to environmental stress, management and other biogeochemical factors. This also indicates that the models are not developed enough to capture systematic differences between the sites.

In our model intercomparison, calibration was performed using eddy covariance based on C flux and evapotranspiration data, together with SWC and ST (but some modelling groups only used a subset of measured data for calibration). Thus, biomass data were not used as a control variable for model optimization, which means that errors associated with the proper estimation of biomass can partly be explained by the lack of adjustments of some internal model parameters associated with biomass. Multi-objective model calibration should be extended to include biomass as a control variable with equal weight as the other, sometimes more data-rich data streams like GPP (Keenan et al., 2011). Besides uncertainty associated with the model parameters, structural problems might also affect the performance of models on yield. For example, constant ratios of the above- to below-ground biomass allocation may cause unsatisfactory model performance on biomass. Ensemble simulation of grassland production is an opportunity as shown in the present study. Uncalibrated ensemble median was the most successful in terms of error statistics, in spite of the fact that the quality of the performance based on the median was still problematic at almost all the sites. Due to calibration, the multi-model median was still useful.

4.4. Ensemble approach of grassland simulation

We used such a simple approach (median of all simulations) to construct ensemble results, but there are alternative ways (see Schwalm et al., 2015 for an overview) to calculate multi-model ensembles to take into account the skill of individual models with weighting according to errors. Schwalm et el. (2015) studied the effect of "naive" (i.e. simple multi-model ensemble like in our case) versus optimal techniques in terms of performance of terrestrial biosphere models. They found that sophisticated, skill-based methods are not superior in comparison with the naive approach in statistical sense. This means that our simple multi-model median approach might already capture the essentials considering the possible applicability of the ensemble technique. Further steps are needed, probably with the inclusion of additional grassland models and ensemble integration techniques to evaluate the usefulness of the ensemble technique. This would mean a major step towards robust and reliable estimation of production and greenhouse gas balance of grasslands.

4.5. Possible explanations for model errors (residual analysis)

We presented an approach that uses a covariance matrix (with graphical representation) to take into account all possible correlations between ST, SWC and yield residuals and, additionally, mean maximum air temperatures and precipitation totals. This residual analysis can help find relationships between some variables, and between variables and external drivers (and thus it can help find additional variables that may need to be included in the models as predictors; Medlyn et al., 2005). This analysis might indicate dependency of errors in one process that is related to another (which is a typical case of error propagation within the model), though the way of error propagation cannot be easily retrieved from the covariance matrix. For example, overestimation of biomass may cause overestimated shading of the soil surface that interferes with the ST simulation. In turn, bias in ST may interact with ecosystem respiration that affects plant growth and thus biomass amount. Underestimation of leaf biomass may interact with evapotranspiration (by decreasing it) which can cause errors in SWC due to slower water depletion. SWC effect on biomass is probably more straightforward. The results indicated that the SWC annual cycle is not well represented by model simulations and, hence, drought stress on plant growth and biomass could not be captured by models. This is particularly well illustrated at GRI.

Considering the specific models that provided calibrated outputs, the results can be used to make recommendations for model improvement (Supplementary material, section 4). The results indicate that the structural errors can be detected based on the analysis of model residuals. The lack of strong correlation between the residuals at the grazed site (LAQ1 and LAQ2) as well as extensive sites (ROT2, KEM2) indicates that the process representation of state-of-the-art grassland models is not satisfactory, and more research is needed to accurately simulate biogeochemical processes and grass yield at grazed and extensively managed sites. As we only used a few variables in the correlation matrix, additional variables might be added to the covariance matrix analysis of residuals.

4.6. Uncertainties in grassland modelling

Uncertainty of output data, defined as spread of results arising from unknown or imperfectly characterized processes, is an inherent property of mathematical modelling. In grassland modelling and, generally, in ecological modelling, uncertainty is caused by internal variability, errors in the initial and boundary conditions, parameterization, and model structure. In multi-model frameworks, uncertainty is also associated with the different model formulations (Schwalm et al., 2015).

Considering the nine grassland models, our study suggests that the spread of the ensemble members tends to be higher than the model error. This means that variability of simulation results can be explained by model formulation rather than structural uncertainties within the models. Work is needed to constrain the multi-model results and decrease uncertainty in simulating grassland functioning. Uncertainty is associated with the measurements which are used to train (i.e. calibrate) the individual grassland models. For example, eddy covariance measurements that were used in the present study inherently contain random and systematic errors that might interact with the parameter estimation (Richardson et al., 2006). Errors associated to the training dataset might cause bias in the optimized parameters for a given model structure. Initial conditions are typically estimated by self-initialization or equilibrium run (e.g. Lardy et al., 2011), which creates consistent initial conditions for the simulations in terms of different pools and nutrient availability. However, the equilibrium pools might deviate strongly from reality. Incorrect estimation of boundary conditions (i.e. meteorological drivers) might also cause uncertainty in the results.

Grassland models typically use many parameters (i.e. constants) that are variables in reality, which substantially alter the biophysical and biogeochemical processes. In many cases, these parameters are hard to define due to lack of measurement (e.g. for plant traits like leaf C:N ratio or specific leaf area), or due to the nature of the parameter (e.g. in empirical equations without physical meaning). Thus, model calibration is essential to optimize model results for a given ecosystem. However, parameters are highly variable in time and space (e.g. Zaehle et al., 2005), thus their general applicability as one defined plant functional type (PFT, Bonan et al., 2002) is problematic. Grassland models can simulate management in such a way that the user prescribes the management related data to the model (e.g. Hidy et al., 2012). However, due to the nature of management the settings are often affected by uncertainties. A typical example is grass cutting, or grazing. Within the present model intercomparison, yield simulation was rather unsuccessful at the grazed site (LAQ1 and LAQ2; Figs. R13 and R14 in the Supplementary material), which can be the consequence of management-related uncertainty. Individual grassland models are constructed using diverse representations of specific processes (Table B1 in Supplementary material). Though there are similarities in the applied methods (e.g. the Penman-Monteith method is used usually for evapotranspiration simulation), the heterogeneity of the process representations is obvious. Scientific level of understanding of plant processes is far from being perfect. Here we mention a few processes that are widely discussed in the literature.

Plant phenology is clearly problematic as timing of onset of vegetation growth and litter production in autumn strongly influence grassland functionality (e.g. Zhang et al., 2013). Photosynthesis routines coupled with stomatal conductance parameterization are subjected to uncertainties due to parameterization. Plant respiration formulation is quite heterogeneous among the models, which is a major source of model output uncertainty in grassland models and biogeochemical models in general. Soil water balance representation is another source of uncertainty for the models that was clearly demonstrated in the present study.

Although grassland models typically have some kind of representation of drought related senescence and changes of plant functioning due to water limitation and/or heat, this logic is still based on the above-mentioned PFT logic. Van der Molen et al. (2011) suggested that grassland ecosystems cannot be considered as a single PFT but should be treated as mixtures of plants with different plant strategic properties. For example, at the drought-prone Bugac-puszta site in Hungary (Nagy et al., 2007), observations revealed that C3 grasses dominate the spring/early summer intensive growth, then during the summer drought resistant C4 grass species start to interact with the overall C balance also due to their delayed phenological cycle at this extensively managed sandy grassland (Nagy Z., personal communication). None of the studied grassland models is at present prepared to represent this strategy for mixtures of grassland species.

Other processes not mentioned here might also be poorly represented within state-of-the-art grassland models. In any case, it is clear that our understanding is not satisfactory yet to provide reliable estimations for grassland functioning and biogeochemistry.

5. Conclusions and future directions

Quantitative representations of the uncertainty in models can be used to study strategies for decision-making. Estimating uncertainty derived from multi-model ensembles is a relatively recent topic in climate-related agronomic research, and it has gained a lot of momentum over the last few years (e.g. Asseng et al., 2013). The uncertainties that are embodied by a spectrum of modelling choices are thus represented and by the inherent imperfection of each and every one of them. In this study, we presented a framework for proper interpretation of model performances and uncertainties obtained with a set of biophysical models (individually and in an ensemble) simulating grasslands systems at a variety of sites.

There are multiple foci when designing multi-model studies of complex ecosystems (such as grasslands) depending on the questions to be answered. We have not identified the best model for grasslands and we have not assigned probability of success to prove the suitability of using one or another model. We are not even claiming that a set of parameter values of general validity was produced by calibrating grassland models. Rather, we have pursued questions to be answered about drivers of grassland processes and modelled responses (and their uncertainties).

The results indicated that some of the main drivers and results of the grassland processes are not represented well by state-of-the-art grassland models. Especially SWC and yield had severe problems that may prevent their applicability in reliable, larger scale experiments. Model errors were presented for the studied processes in a tabular form, which may provide comparability basis for further studies. Presentation of daily, weekly and monthly results might be useful for other researchers to compare model performance at the same sites. Calibration seemed to improve the model results to some extent, but there was no dramatic increase in model performance for any of the studied models, at any of the sites. Ensemble technique seems to be a feasible method for the simulation of grassland processes, but model development is inevitable to improve the multi-model approach. In our intercomparison, we highlighted the uncertainties that are associated with the models, and we created recommendations to some of the models. Uncertainty was characterized in a fashion, which allowed highlighting the scientific challenges faced in simulating soil processes (temperature and water content) and biomass on European and peri-European grasslands with a variety of state-of-the art models used individually or within an ensemble. What seems to be a message from our intercomparison is that grassland models should be further developed and tested at a large number of experimental sites. In order to provide validation and calibration data for the models, essential processes and outputs like GPP, RECO, SWC, ST, C allocation, emission of non-CO_2 GHGs, and also magnitude and timing of human intervention should be characterized in systematic and accurate fashion in multiple grassland sites covering large climatic gradients.

Though the exercise of the presented model intercomparison performed (the first on permanent grasslands) is large enough, we are aware that it does not completely cover most of the modelling approaches used to simulate grasslands. An example is the process-based, biogeochemical model ORCHIDEE-GM, which includes an enhanced representation of grassland management derived from PaSim (Chang et al., 2013, 2015). Another example is represented by a grassland-specific model derived from STICS (BioMA-Grassland, personal communication by G. De Sanctis, Joint Research Centre of the European Commission, Ispra, Italy), which is being developed for the platform BioMA (Biophysical Models Applications, http://bioma.jrc.ec.europa.eu). Grassland model intercomparisons with the inclusion of more models should therefore be continued to improve our ability to simulate grassland processes with acceptable quality. We also think that further analyses and better understanding of these ensembles are required to achieve fundamental progress in grassland modelling by investigating the sensitivity of models to climate and management drivers. This assessment goes beyond the scope of this paper, and a paper on this topic should be arranged later as a natural evolution of what has already been presented here.

Acknowledgements

The results of this research were obtained within an international research project named "FACCE MACSUR – Modelling European Agriculture with Climate Change for Food Security, a FACCE JPI knowledge hub", with the support of the Hungarian Scientific Research Fund (OTKA K104816) and the BioVeL project (Biodiversity Virtual e-Laboratory Project, FP7-INFRASTRUCTURES-2011-2, project number 283359), the German Ministry of Education and Research (031A103A), the Italian Ministry of Agricultural, Food and Forestry Policies, the Cabinet of the French Community of Belgium, and the metaprogramme Adaptation of Agriculture and Forests to Climate Change (AAFCC) of the French National Institute for Agricultural Research (INRA). We thank the individual site PIs (Katja Klumpp, French National Institute for Agricultural Research, Clermont-Ferrand, France; Christof Ammann, Agroscope, Zurich, Switzerland; Damiano Gianelle, Edmund Mach Foundation, San Michele all'Adige, Italy; Christian Bernhofer, Dresden University of Technology, Germany) and the technical staff for sharing their eddy covariance data. We also acknowledge technical support from the European Fluxes Database Cluster (http://www.europe-fluxdata.eu). We thank Luigi Ledda (University of Sassari, Italy) for providing data from Sassari grassland site and Katharina Braunmiller (Thünen Institute of Market Analysis, Braunschweig, Germany) for facilitating contacts with the Partner Institutions which provided other grassland data. Raphaël Martin and Haythem Ben Touhami helped in the running and calibration of PaSim at the French National Institute for Agricultural Research (Clermont-Ferrand, France). Biome-BGC version 4.1.1 (the predecessor of BBGC MuSo) was provided by the Numerical Terradynamic Simulation Group (NTSG) at the University of Montana, Missoula MT (USA), which assumes no responsibility for the proper use by others. We are grateful to the Laboratory of Parallel and Distributed Systems, Institute for Computer Science and Control (MTA SZTAKI), that provided consultation, technical expertise and access to the EDGeS@home volunteer desk top grid system in computation demanding analysis.

References

Ammann, C., Flechard, C.R., Leifeld, J., Neftel, A., Fuhrer, J., 2007. The carbon budget of newly established temperate grassland depends on management intensity. Agr. Ecosyst. Environ. 121, 5-20. doi:10.1016/j.agee.2006.12.002

Asseng, S., Ewert, F., Rosenzweig, C., Jones, J.W., Hatfield, J.L., Ruane, A., Boote, K.J., Thorburn, P., Rötter, R.P., Cammarano, D., Brisson, N., Basso, B., Martre, P., Aggarwal, P.K., Angulo, C., Bertuzzi, P., Biernath, C., Doltra, J., Gayler, S., Goldberg, R., Grant, R., Heng, L., Hooker, J.E., Hunt, L.A., Ingwersen, J., Izaurralde, R.C., Kersebaum, K.C., Müller, C., Naresh Kumar, S., Nendel, C., O'Leary, G., Olesen, J.E., Osborne, T.M., Palosuo, T., Priesack, E., Ripoche, D., Semenov, M.A., Shcherbak, I., Steduto, P., Stöckle, C.O., Stratonovitch, P., Streck, T., Supit, I., Travasso, M., Tao, F., Waha, K., Wallach, D., White, J.W., Wolf, J., 2013. Uncertainties in simulating wheat yields under climate change. Nat. Clim. Change 3, 827-832. doi:10.1038/nclimate1916

Aubinet, M., Vesala, T., Papale, D., 2012. Eddy covariance: A practical guide to measurement and data analysis. Springer, Dordrecht.

Bonan, G.B., Levis, S., Kergoat, L., Oleson, K.W., 2002. Landscapes as patches of plant functional types: An integrating concept for climate and ecosystem models. Global Biogeochem. Cy. 16, 5.1–5.23. doi:10.1029/2000GB001360

Cavallero, A., Talamucci, P., Grignani, C., Reyneri, A., Ziliotto, U., Scotton, M., Bianchi, A.A., Santilocchi, R., Basso, F., Postiglione, L., Carone, F., Corleto, A., Cazzato, E., Cassaniti, S., Cosentino, S., Litrico, P.G., Leonardi, S., Sarno, R., Stringi, L., Gristina, L., Amato, G., Bullitta, P., Caredda, S., Roggero, P.P., Caporali, F., D'Antuono, L.F., Pardini, A., Zagni, C., Piemontese, S., Pazzi, G., Costa, G., Pascal, G., Acutis, M., 1992. Caratterizzazione della dinamica produttiva di pascoli naturali italiani. Rivista di Agronomia 26, n. 3 suppl., 325-343. (in Italian)

Carvalhais, N., Reichstein, M., Seixas, J., Colltaz, G.J., Pereira, J.S., Berbigier, P., Carrara, A., Granier, A., Montagnani, L., Papale, D., Rambal, S., Sanz, M.J., Valentini, R., 2008. Implications of the carbon cycle steady state assumption for biogeochemical modeling performance and inverse parameter retrieval. Global Biogeochem. Cy. 22, GB2007. doi:10.1029/2007GB003033

Chang, J., Viovy, N., Vuichard, N., Ciais, P., Wang, T., Cozic, A., Lardy, R., Graux, A.-I., Klumpp, K., Martin, R., Soussana, J.-F., 2013. Incorporating grassland management in ORCHIDEE: model description and evaluation at 11 eddy-covariance sites in Europe. Geosci. Model Dev. 6, 2165-2181. doi:10.5194/gmd-6-2165-2013

Chang, J., Viovy, N., Vuichard, N., Ciais, P., Campioli, M., Klumpp, K., Martin, R., Leip, A., Soussana, J., 2015. Modelled changes in potential grassland productivity and in ruminant livestock density in Europe over 1961–2010. PLoS One 10, e0127554. doi:10.1371/journal.pone.0127554554

Chow, L., Xing, Z., Rees, H.W., Meng, F., Monteith, J., Stevens, L., 2009. Field performance of nine soil water content sensors on a sandy loam soil in New Brunswick, Maritime Region, Canada. Sensors 9, 9398–9413. doi:10.3390/s91109398

Collins, S.L., 1995. The measurement of stability in grasslands. Trends Ecol. Evol. 10, 95-96.

De Martonne, E., 1942. Nouvelle carte mondiale de l'indice d'aridité. Annales de Géographie 51, 242–250 (in French).

Fodor N., Rajkai K., 2011. Computer program (SOILarium 1.0) for estimating the physical and hydrophysical properties of soils from other soil characteristics. Agrochemistry and Soil Science 60, 27-40.

Golodets, C., Sternberg, M., Kigel, J., Boeken, B., Henkin, Z., Seligman, N.G., Ungar, D.E., 2013. From desert to Mediterranean rangelands: will increasing drought and inter-annual rainfall variability affect herbaceous annual primary productivity? Climatic Change 119, 785-798. doi:10.1007/s10584-013-0758-8

Graux A.-I., Bellocchi G., Lardy R., Soussana J.-F., 2013. Ensemble modelling of climate change risks and opportunities for managed grasslands in France. Agr. Forest Meteorol. 170, 114-131. doi:10.1016/j.agrformet.2012.06.010

Hidy, D., Barcza, Z., Haszpra, L., Churkina, G., Pintér, K., Nagy, Z., 2012. Development of the Biome-BGC model for simulation of managed herbaceous ecosystems. Ecol. Model. 226, 99-119. doi:10.1016/j.ecolmodel.2011.11.008

Huyghe, C., 2008. La multifonctionnalité des prairies I - Les fonctions de production. Cahiers Agricultures 17, 427-435. (in French)

Keenan, T.F., Carbone, M.S., Reichstein, M., Richardson, A.D., 2011. The model-data fusion pitfall: assuming certainty in an uncertain world. Oecologia 167, 587–597. doi:10.1007/s00442-011-2106-x

Klumpp, K., Tallec, T., Guix, N., Soussana, J.-F., 2011. Long-term impacts of agricultural practices and climatic variability on carbon storage in a permanent pasture. Global Change Biol. 17, 3534–3545. doi:10.1111/j.1365-2486.2011.02490.x

Lardy R., Bellocchi G., Soussana J.F., 2011. A new method to determine soil organic carbon equilibrium. Environ. Modell. Softw. 26, 1759-1763. doi:10.1016/j.envsoft.2011.05.016

Ma, S., Acutis, M., Barcza, Z., Ben Touhami, H., Doro, L., Hidy, D., Köchy, M., Minet, J., Lellei-Kovács, E., Perego, A., Rolinski, S., Ruget, F., Seddaiu, G., Wu, L., Bellocchi, G. 2014. The grassland model intercomparison of the MACSUR (Modelling European Agriculture with Climate Change for Food Security) European knowledge hub. In: Ames, D.P. Quinn, N. (Eds.) Proceedings of the 7th International Congress of the Environmental

Modelling and Software Society, 15-19 June, San Diego, CA. http://www.iemss.org/society/index.php/iemss-2014-proceedings (accessed 18.11.2015)

Ma, S., Lardy, R., Graux, A.-I., Ben Touhami, H., Klumpp, K., Martin, R., Bellocchi, G., 2015. Regional-scale analysis of carbon and water cycles on managed grassland systems. Environ. Modell. Softw. 72, 356-371, doi:10.1016/j.envsoft.2015.03.007.

Mapfumo, E., Naeth, M.A., Baron, V.S., Dick, A.C., Chanasyk, D.S., 2002. Grazing impacts on litter and roots: perennial versus annual grasses. Journal of Range Management 55, 16-22.

Marriott, C., Fothergill, M., Jeangros, B., Scotton, M., Louault, F., 2004. Long-term impacts of extensification of grassland management on biodiversity and productivity in upland area. A review. Agronomie 24, 447-461.

Martre, P., Wallach, D., Asseng, S., Ewert, F., Jones, J.W., Rotter, R.P., Boote, K.J., Ruane, A.C., Thorburn, P.J., Cammarano, D., Hatfield, J.L., Rosenzweig, C., Aggarwal, P.K., Angulo, C., Basso, B., Bertuzzi, P., Biernath, C., Brisson, N., Challinor, A.J., Doltra, J., Gayler, S., Goldberg, R., Grant, R.F., Heng, L., Hooker, J., Hunt, L.A., Ingwersen, J., Izaurralde, R.C., Kersebaum, K.C., Müller, C., Kumar, S.N., Nendel, C., O'leary, G., Olesen, J.E., Osborne, T.M., Palosuo, T., Priesack, E., Ripoche, D., Semenov, M.A., Shcherback, I., Steduto, P., Stöckle, C.O., Stratonovitch, P., Streck, T., Supit, I., Tao, F., Travasso, M., Waha, K., White, J.W., Wolf, J., 2015. Multimodel ensembles of wheat growth: many models are better than one. Global Change Biol. 21, 911-925.

Medlyn, B.E., Robinson, A.P., Clement, R., McMurtrie, R.E., 2005. On the validation of models of forest CO_2 exchange using eddy covariance data: some perils and pitfalls. Tree Physiol. 25, 839–857. doi:10.1093/treephys/25.7.839

Nagy Z., Pintér K., Czóbel Sz., Balogh J., Horváth L., Fóti Sz., Barcza Z., Weidinger T., Csintalan Zs., Dinh N.Q., Grosz B., Tuba Z., 2007. The carbon budget of a semi-arid grassland in a wet and a dry year in Hungary. Agr. Ecosyst. Environ. 121, 21-29. doi:10.1016/j.agee.2006.12.003

Peeters, A., 2012. Past and future of European grasslands. The challenge of the CAP towards 2020. In: Goliński, P., Warda, M., Stypiński, P. (Eds.), Proceedings of the 24th General Meeting of the European Grassland Federation, 3-7 June, Lublin, pp. 17-32. http://www.europeangrassland.org/fileadmin/media/EGF2012.pdf

Peyraud, J.-L., 2013. Réforme de la PAC et prairies permanentes. Dans : La PAC a 50 ans : le bel âge ? Colloque organisé par lnstitut National de la Recherche Agronomique dans le cadre du Salon International de l'Agriculture, February 26, Paris. (in French)

Prescher, A.-K., Grünwald, T., Bernhofer, C., 2010. Land use regulates carbon budgets in eastern Germany: from NEE to NBP. Agr. Forest Meteorol. 150, 1016–1025. doi:10.1016/j.agrformet.2010.03.008

Richardson, A. D., Hollinger, D. Y., Burba, G. G., Davis, K. J., Flanagan, L. B., Katul, G. G., Munger, J. W., Ricciuto, D. M., Stoy, P. C., Suyker, A. E., Verma, S. B., and Wofsy, S.C., 2006. A multi-site analysis of random error in tower-based measurements of carbon and energy fluxes. Agr. Forest Meteorol. 136, 1–18. doi: 10.1016/j.agrformet.2006.01.007

Rosenzweig, C., Jones, J.W., Hatfield, J.L., Ruane, A.C., Boote, K.J., Thorburn, P., Antle, J.M., Nelson, G.C., Porter, C., Janssen, S., Asseng, S., Basso, B., Ewert, F., Wallach, D., Baigorria, G., Winter, J.M., 2013. The Agricultural Model Intercomparison and Improvement Project (AgMIP): protocols and pilot studies. Agr. Forest Meteorol. 17, 166-182. doi:10.1016/j.agrformet.2012.09.011

Sándor, R., Ma, S., Acutis, M., Barcza, Z., Ben Touhami, H., Doro, L., Hidy, D., Köchy, M., Lellei-Kovács, E., Minet, J., Perego, A., Rolinski, S., Ruget, F., Seddaiu, G., Wu, L., Bellocchi, G., 2015. Uncertainty in simulating biomass yield and carbon-water fluxes from grasslands under climate change. Advances in Animal Biosciences 6, 49-51. doi:10.1017/S2040470014000545

Schader, C., Muller, A., Scialabba, N.E.-H., 2013. Sustainability and organic livestock modelling (SOL-m). Impacts of a global upscaling of low-input and organic livestock production. Preliminary results. Natural Resources Management and Environmental Department, FAO, Rome. http://www.fao.org/fileadmin/templates/nr/sustainability_pathways/docs/SOL-m_summary_Final.pdf (accessed 18.11.2015)

Schils, R., Snijders, P., 2004. The combined effect of fertiliser nitrogen and phosphorus on herbage yield and changes in soil nutrients of a grass/clover and grass-only sward. Nutr. Cycl. Agroecosys. 68, 165–179. doi:10.1023/B:FRES.0000019045.90791.a4

Schröpel, R., Diepolder, M., 2003. Auswirkungen der Grünlandextensivierung auf einer Weidelgras-Weißklee-Weide im Allgäuer Alpenvorland. Schuleund Beratung, Heft 11/2003, Seite III-13 bis III-15; Bayerisches Staatsministerium für Landwirtschaft und Forsten, Munich. (in German)

Schwalm, C.R., Huntinzger, D.N., Fisher, J.B., Michalak, A.M., Bowman, K., Cias, P., Cook, R., El-Masri, B., Hayes, D., Huang, M., Ito, A., Jain, A., King, A.W., Lei, H., Liu, J., Lu, C., Mao, J., Peng, S., Poulter, B., Ricciuto, D., Schaefer, K., Shi, X., Tao, B., Tian, H., Wang, W., Wei, Y., Yang, J., Zeng, N., 2015. Toward “optimal” integration of terrestrial biosphere models. Geophys. Res. Lett. 42, 4418–4428. doi: 10.1002/2015GL064002.

Silvertown, J., Poulton, P., Johnston, E., Edwards, G., Heard, M., Biss, P.M., 2006. The Park Grass Experiment 1856–2006: its contribution to ecology. J. Ecol. 94, 801–814. doi:10.1111/j.1365-2745.2006.01145.x

Soussana, J.F., Ehrhardt, F., Conant, R., Harrison, M., Lieffering, M., Bellocchi, G., Moore, A., Rolinski, S., Snow, V., Wu, L., Ruane, A., 2015. Projecting grassland sensitivity to climate change from an ensemble of models. In: Abstract Book of "Our Common Future Under Climate Change" International Scientific Conference, July 7-10, Paris, K-2223-02. http://pool7.kermeet.com/C/ewe/ewex/unesco/DOCS/CFCC_abstractBook.pdf (accessed 18.11.2015)

van der Molen, M.K., Dolman, A.J., Ciais, P., Eglin, T., Gobron, N., Law, B.E., Meir, P., Peters, W., Phillips, O.L., Reichstein, M., Chen, T., Dekker, S.C., Doubková, M., Friedl, M. A., Jung, M., van den Hurk, B.J.J.M., de Jeu, R.A.M., Kruijt, B., Ohta, T., Rebel, K.T., Plummer, S., Seneviratne, S.I., Sitch, S., Teuling, A.J., van der Werf, G.R., Wang, G., 2011. Drought and ecosystem carbon cycling. Agr. Forest Meteorol. 151, 765–773. doi:10.1016/j.agrformet.2011.01.018

Vital, J.-A., Gaurut, M., Lardy, R., Viovy, N., Soussana, J.-F., Bellocchi, G., Martin, R., 2013. High-performance computing for climate change impact studies with the Pasture Simulation model. Comput. Electron. Agr. 98, 131-135. doi:10.1016/j.compag.2013.08.004

Volaire, F., Lelièvre, F., 2001. Drought survival in *Dactylis glomerata* and *Festuca arundinacea* under similar rooting conditions. Plant Soil 229, 225-234. doi:10.1023/A:1004835116453

Vuichard, N., Soussana, J.-F., Ciais, P., Viovy, N., Ammann, C., Calanca, P., Clifton-Brown, J., Fuhrer, J., Jones, M., Martin, C., 2007. Estimating the greenhouse gas fluxes of European grasslands with a process-based model: 1. Model evaluation from in situ measurements. Global Biogeochem. Cy. 21, GB1004. doi:10.1029/2005GB002611

Weitz, A.M., Grauel, W.T., Keller, M., Veldkamp, E., 1997. Calibration of time domain reflectometry technique using undisturbed soil samples from humid tropical soils of volcanic origin. Water Resour. Res. 33, 1241-1249. doi:10.1029/96WR03956

Wohlfahrt, G., Anderson-Dunn, M., Bahn, M., Balzarolo, M., Berninger, F., Campbell, C., Carrara, A., Cescatti, A., Christensen, T., Dore, S., Eugster, W., Friborg, T., Furger, M., Gianelle, D., Gimeno, C., Hargreaves, K., Hari, P., Haslwanter, A., Johansson, T., Marcolla, B., Milford, C., Nagy, Z., Nemitz, E., Rogiers, N., Sanz, M.J., Siegwolf, R.T.W., Susiluoto, S., Sutton, M., Tuba, Z., Ugolini, F., Valentini, R., Zorer, R., Cernusca, A., 2008. Biotic, abiotic, and management controls on the net ecosystem CO_2 exchange of European mountain grassland ecosystems. Ecosystems 11, 1338-1351. doi:10.1007/s10021-008-9196-2

Wösten J.H.M., Lilly, A., Nemes, A., Le Bas, C., 1999. Development and use of a database of hydraulic properties of European soils. Geoderma 90,169-185. doi:10.1016/S0016-7061(98)00132-3

Zaehle, S., Sitch, S., Smith, B., Hatterman, F., 2005. Effects of parameter uncertainties on the modeling of terrestrial biosphere dynamics. Global Biogeochem. Cy. 19, GB3020. doi:10.1029/2004GB002395

Zhang, X., Tarpley, D., Sullivan, J.T., 2013. Diverse responses of vegetation phenology to a warming climate. Geophys. Res. Lett. 34, L19405. doi:10.1029/2007GL031447

APPENDICES

Appendix 1

Individual (M1-M8) and multi-model ensemble (MMM) performance at different information (SIM) levels - uncalibrated (*U1*, *U2*), calibrated (*C*) and validated (*V*) - at the most humid and the most arid flux sites (ID as in Table 1) based on different metrics calculated on weekly averaged soil temperature (ST). NA: not available ST simulations.

Model ID	**SIM**	**Mean of observations (°C)**		**Mean of simulations (°C)**		**BIAS (°C)**		**RRMSE (%)**		**ME**		**R^2**	
		GRI	**LAQ1**	**GRI**	**LAQ1**	**GRI**	**LAQ1**	**GRI**	**LAQ1**	**GRI**	**LAQ1**	**GRI**	**LAQ1**
M1	*U1*	9.74	8.95	8.71	7.69	-1.02	-1.26	32.25	28.77	-0.06	-0.15	0.77	0.83
	U2	10.17	8.54	9.69	7.63	-0.47	-0.90	14.63	25.80	0.07	0.36	0.95	0.93
	C	9.74	8.95	8.60	7.71	-1.14	-1.24	34.81	37.02	0.63	0.59	0.89	0.90
	V	10.17	8.54	9.39	7.49	-0.78	-1.05	30.60	41.71	0.70	0.72	0.96	0.93
M2	*U1*	9.74	8.95	5.01	7.36	-4.73	-1.59	54.21	28.66	-1.37	-0.65	0.90	0.89
	U2	10.17	8.54	6.91	6.92	-3.26	-1.62	39.92	27.81	-0.79	-0.19	0.92	0.94
	C	9.74	8.95	4.85	7.17	-4.89	-1.78	55.09	28.16	-1.36	-0.59	0.90	0.90
	V	10.17	8.54	6.81	6.58	-3.36	-1.96	40.69	29.96	-0.80	-0.09	0.92	0.94
M3	*U1*	9.74	8.95	10.38	10.26	0.64	1.31	50.53	50.83	0.88	0.79	0.70	0.78
	U2	10.17	8.54	10.44	10.26	0.27	1.72	44.54	56.98	0.88	0.81	0.73	0.78
	C	9.74	8.95	7.80	7.65	-1.94	-1.30	29.93	25.49	-0.17	-0.07	0.86	0.88
	V	10.17	8.54	9.15	7.31	-1.02	-1.23	18.44	31.00	0.13	0.31	0.94	0.88
M4	*U1*	9.74	8.95	10.04	8.70	0.31	-0.25	36.77	28.93	-1.10	-0.88	0.91	0.90
	U2	10.17	8.54	11.94	8.37	1.77	-0.16	35.82	23.48	-0.98	-0.37	0.91	0.94
	C	9.74	8.95	10.01	8.36	0.27	-0.59	35.59	26.59	-1.05	-0.81	0.91	0.91
	V	10.17	8.54	11.70	8.01	1.54	-0.53	32.55	20.71	-0.88	-0.27	0.93	0.95
M5	*U1*	9.74	8.95	7.80	NA	-1.94	NA	27.08	NA	-0.02	NA	0.89	NA
	U2	10.17	8.54	9.14	NA	-1.03	NA	17.06	NA	0.25	NA	0.97	NA
	C	9.74	8.95	7.84	NA	-1.89	NA	27.38	NA	-0.29	NA	0.90	NA
	V	10.17	8.54	9.31	NA	-0.86	NA	16.08	NA	0.02	NA	0.95	NA
M6	*U1*	9.74	8.95	6.95	7.21	-2.79	-1.74	31.92	24.50	-0.15	0.04	0.91	0.93
	U2	10.17	8.54	8.81	6.80	-1.36	-1.74	18.99	31.93	0.24	0.34	0.97	0.93
	C	9.74	8.95	11.45	7.20	1.72	-1.75	40.15	33.64	0.44	0.38	0.73	0.88
	V	10.17	8.54	10.50	5.96	0.33	-2.58	26.89	42.21	0.05	0.39	0.81	0.93
M7	*U1*	9.74	8.95	8.23	NA	-1.51	NA	25.37	NA	-0.33	NA	0.90	NA
	U2	10.17	8.54	9.72	NA	-0.45	NA	12.99	NA	-0.02	NA	0.96	NA
	C	9.74	8.95	7.86	NA	-1.88	NA	27.29	NA	-0.36	NA	0.90	NA
	V	10.17	8.54	9.36	NA	-0.81	NA	14.56	NA	-0.03	NA	0.96	NA
M8	*U1*	9.74	8.95	28.06	28.04	18.32	19.09	198.41	223.42	-7.42	-10.57	0.80	0.86
	U2	10.17	8.54	28.21	27.99	18.05	19.45	186.53	238.72	-7.48	-8.24	0.95	0.89
	C	9.74	8.95	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	V	10.17	8.54	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
MMM	*U1*	9.74	8.95	8.14	8.39	-1.60	-0.56	24.63	20.03	-0.12	0.00	0.90	0.92
	U2	10.17	8.54	9.66	8.03	-0.51	-0.50	12.12	18.58	0.17	0.31	0.97	0.97
	C	9.74	8.95	7.90	7.44	-1.83	-1.51	26.59	22.54	-0.26	0.02	0.90	0.93
	V	10.17	8.54	9.31	6.91	-0.86	-1.63	14.34	28.75	0.07	0.31	0.96	0.95

Appendix 2

Individual (M1-M9) and multi-model ensemble (MMM) model performance at different information (SIM) levels - uncalibrated (*U1*, *U2*), calibrated (*C*) and validated (*V*) - at the most humid and the most arid flux sites (ID as in Table 1) based on different metrics calculated on weekly averaged soil water content (SWC). NA: not available SWC simulations.

Model ID	SIM	Mean of observations (m^3 m^{-3})		Mean of simulations (m^3 m^{-3})		BIAS (m^3 m^{-3})		RRMSE (%)		ME		R^2	
		GRI	**LAQ1**	**GRI**	**LAQ1**	**GRI**	**LAQ1**	**GRI**	**LAQ1**	**GRI**	**LAQ1**	**GRI**	**LAQ1**
M1	*U1**	0.45	0.36	0.37	0.36	-0.08	0.01	14.17	11.20	-714.6	0.30	0.10	0.50
	U2	0.41	0.33	0.36	0.36	-0.06	0.04	18.01	15.98	0.32	-1.91	0.83	0.25
	*C**	0.45	0.36	0.39	0.39	-0.06	0.03	13.61	15.43	-329	0.34	0.08	0.46
	V	0.41	0.33	0.38	0.39	-0.03	0.06	17.84	21.38	0.82	-3.55	0.87	0.37
M2	*U1**	0.45	0.36	0.39	0.38	-0.06	0.02	16.35	14.38	-406.8	0.00	0.32	0.41
	U2	0.41	0.33	0.37	0.39	-0.04	0.06	14.94	21.67	0.42	-3.65	0.82	0.20
	*C**	0.45	0.36	0.39	0.37	-0.06	0.02	16.51	14.21	-409.9	-0.05	0.45	0.40
	V	0.41	0.33	0.38	0.39	-0.04	0.06	15.37	21.22	0.49	-3.53	0.76	0.20
M3	*U1**	0.45	0.36	0.24	0.26	-0.21	-0.10	44.31	31.65	-4291	-3.68	0.34	0.18
	U2	0.41	0.33	0.22	0.26	-0.19	-0.06	47.51	24.08	-3.87	-4.92	0.70	0.07
	*C**	0.45	0.36	0.30	0.35	-0.15	-0.01	33.46	19.79	-2219	-0.64	0.55	0.12
	V	0.41	0.33	0.27	0.43	-0.14	0.11	37.11	50.52	-1.80	-23.88	0.60	0.00
M4	*U1**	0.45	0.36	0.23	0.38	-0.22	0.02	50.95	14.41	-4336	0.21	0.09	0.31
	U2	0.41	0.33	0.23	0.38	-0.19	0.05	48.60	19.64	-3.44	-3.06	0.56	0.23
	*C**	0.45	0.36	0.34	0.36	-0.11	0.00	25.13	11.73	-1011	0.56	0.20	0.44
	V	0.41	0.33	0.34	0.36	-0.08	0.04	26.52	14.14	0.29	-0.86	0.66	0.29
M5	*U1**	0.45	0.36	0.31	NA	-0.14	NA	37.84	NA	-1934	1.00	0.00	NA
	U2	0.41	0.33	0.31	NA	-0.11	NA	33.85	NA	-0.52	1.00	0.02	NA
	*C**	0.45	0.36	0.29	NA	-0.16	NA	37.95	NA	-2368	1.00	0.38	NA
	V	0.41	0.33	0.29	NA	-0.13	NA	34.63	NA	-1.29	1.00	0.55	NA
M6	*U1**	0.45	0.36	0.31	0.29	-0.14	-0.06	42.54	24.92	-2066	-1.94	0.00	0.10
	U2	0.41	0.33	0.31	0.30	-0.10	-0.03	30.46	19.01	-0.75	-3.42	0.38	0.18
	*C**	0.45	0.36	0.45	0.33	0.00	-0.03	3.43	12.90	0.74	-0.40	0.26	0.48
	V	0.41	0.33	0.46	0.30	0.05	-0.03	18.81	12.51	0.29	-0.02	0.53	0.18
M7	*U1**	0.45	0.36	0.70	NA	0.25	NA	64.14	NA	-5817	1.00	0.45	NA
	U2	0.41	0.33	0.69	NA	0.28	NA	68.30	NA	-8.97	1.00	0.49	NA
	*C**	0.45	0.36	0.35	NA	-0.10	NA	17.90	NA	-1038	1.00	0.44	NA
	V	0.41	0.33	0.34	NA	-0.07	NA	24.44	NA	0.11	1.00	0.44	NA
M8	*U1**	0.45	0.36	0.19	0.22	-0.26	-0.13	66.39	62.01	-8509	-24.12	0.70	0.18
	U2	0.41	0.33	0.14	0.19	-0.27	-0.13	72.21	58.86	-10.05	-36.17	0.14	0.01
	*C**	0.45	0.36	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	V	0.41	0.33	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
M9	*U1**	0.45	0.36	0.29	0.31	-0.16	-0.05	46.04	25.15	-2627	-2.09	0.51	0.03
	U2	0.41	0.33	0.29	0.32	-0.12	-0.01	40.18	21.67	-1.60	-4.25	0.01	0.06
	*C**	0.45	0.36	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	V	0.41	0.33	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
MMM	*U1**	0.45	0.36	0,31	0,33	-0,14	-0,02	40,18	16,11	-1945,6	-0,44	0,01	0,22
	U2	0.41	0.33	0,30	0,33	-0,11	0,01	32,71	13,39	-0,72	-1,04	0,23	0,20
	*C**	0.45	0.36	0,35	0,36	-0,10	0,01	17,90	11,28	-975,49	0,43	0,44	0,55
	V	0.41	0.33	0,35	0,37	-0,07	0,05	22,91	17,39	0,30	-1,79	0,74	0,20

* Six available observed SWC data during *U1* and *C* simulations at Grillenburg.

Appendix 3

Individual (M1-M9) and multi-model ensemble (MMM) model performance at different information (SIM) levels - uncalibrated (*U*) and calibrated (*C*) - for SAS and KEM1 sites (ID as in Table 1) based on different metrics calculated on cutting events of yield biomass (harvested aboveground biomass). NA: not available yield simulations.

Model ID	SIM	Mean of observations (g DM m^{-2})		Mean of simulations (g DM m^{-2})		BIAS (g DM m^{-2})		RRMSE (%)		ME		R^2	
		SAS	**KEM1**	**SAS**	**KEM1**	**SAS**	**KEM1**	**SAS**	**KEM1**	**SAS**	**KEM1**	**SAS**	**KEM1**
M1	*U*	117.6	126.6	64.5	240.0	-53.1	113.4	89.4	132.3	-0.26	-22.99	0.15	0.09
	C			26.9	113.1	-90.7	-13.4	102.5	56.6	-0.46	-2.63	0.14	0.02
M2	*U*	117.6	126.6	11.1	93.2	-106.6	-33.4	111.4	46.8	-0.67	-1.18	0.22	0.02
	C			5.2	57.5	-112.5	-69.0	118.0	65.0	-0.81	-3.78	0.08	0.02
M3	*U*	117.6	126.6	62.6	36.1	-55.0	-90.4	129.8	80.3	-0.93	-6.20	0.02	0.01
	C			10.7	23.2	-107.0	-103.3	113.5	86.1	-0.62	-7.71	0.32	0.04
M4	*U*	117.6	126.6	34.8	124.9	-82.8	-1.7	97.8	25.7	0.02	0.84	0.21	0.14
	C			NA	184.0	NA	57.5	NA	54.5	NA	-2.39	NA	0.10
M5	*U*	117.6	126.6	85.6	38.4	-32.0	-88.1	72.5	79.4	0.00	-6.32	0.28	0.00
	C			85.6	101.8	-32.0	-24.8	72.5	67.7	0.00	-3.46	0.28	0.02
M6	*U*	117.6	126.6	190.3	335.8	72.6	209.3	139.8	181.5	-3.98	-42.07	0.28	0.05
	C			110.7	183.3	-6.9	56.7	73.9	62.0	0.68	-3.77	0.05	0.07
M7	*U*	117.6	126.6	99.7	166.5	-17.9	39.9	92.9	60.9	-0.87	-5.05	0.19	0.26
	C			65.9	155.6	-51.7	29.1	76.0	52.5	0.07	-4.13	0.29	0.37
M8	*U*	117.6	126.6	97.2	466.3	-20.4	339.7	88.4	294.5	0.44	-111.08	0.00	0.00
	C			NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
M9	*U*	117.6	126.6	107.0	179.9	-10.6	53.4	91.3	107.5	0.08	-13.92	0.03	0.02
	C			NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
MMM	*U*	117.6	126.6	61.2	153.5	-56.5	26.9	81.5	31.8	0.17	-1.48	0.19	0.62
	C			38.7	106.5	-78.9	-20.0	87.6	32.7	-0.12	-0.40	0.40	0.24

Appendix 4

Average ratio of the ensemble spread to model error: average absolute standardized spread (maximum-minimum) of model results / average absolute standardized model residual. Responses are from calibrated simulations of soil temperature (ST), soil water content (SWC) and yield biomass, as obtained at each site (ID as in Table 1) with both individual models (M1-M7) and the ensemble median (MMM). NA: not available simulations.

Output	**Site**	**M1**	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**	**MMM**
ST	OEN	1.10	1.92	3.90	6.19	5.03	1.95	5.58	4.95
	MBO	1.07	2.72	2.60	3.80	3.03	1.44	3.39	2.97
	GRI	1.54	2.42	3.91	4.15	4.78	2.25	5.16	4.95
	LAQ1	1.00	2.79	2.37	4.17	NA	1.39	NA	2.53
	LAQ2	1.53	3.04	3.54	4.51	NA	1.91	NA	4.19
SWC	OEN	0.64	1.23	1.09	1.33	1.13	4.42	1.07	2.04
	MBO	0.38	0.57	0.39	2.15	0.66	3.04	1.40	0.62
	GRI	0.83	2.03	1.01	0.29	0.91	2.66	1.05	0.82
	LAQ1	0.83	1.56	2.58	1.61	NA	2.48	NA	1.60
	LAQ2	0.74	1.62	3.09	1.46	NA	1.33	NA	2.27
Yield biomass	KEM1	0.96	0.95	0.14	1.10	2.49	1.18	2.27	1.89
	KEM2	0.75	0.76	0.15	0.72	1.65	1.48	2.30	0.76
	ROT1	1.92	2.51	0.14	4.96	1.47	1.78	3.76	2.07
	ROT2	1.82	2.44	0.15	6.05	1.63	1.66	4.30	2.29
	LEL	0.28	0.73	0.13	2.62	1.97	0.52	1.10	0.44
	MAT	0.20	1.52	0.11	0.09	0.94	2.18	1.04	1.07
	SAS	0.71	0.15	0.10	NA	2.09	4.57	1.12	0.75
	OEN	0.09	0.61	0.99	1.05	0.48	0.47	1.09	0.50
	MBO	0.52	0.55	4.67	0.39	0.85	3.27	2.56	0.79
	GRI	0.63	1.02	0.96	0.99	2.08	0.93	1.84	1.10
	LAQ1	1.28	1.42	0.55	0.83	NA	2.56	NA	2.17
	LAQ2	1.67	1.32	0.19	1.09	NA	1.15	NA	1.86