



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

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

7

8 ^a Grassland Ecosystem Research Unit, French National Institute for Agricultural Research, Clermont-Ferrand, France

9 ^b MTA Centre for Agricultural Research, Institute for Soil Sciences and Agricultural Chemistry, Budapest, Hungary

10 ^c Eötvös Loránd University, Department of Meteorology, Budapest, Hungary

11 ^d University of Milan, Department of Agricultural and Environmental Sciences - Production, Landscape, Agroenergy, Milan,
12 Italy

13 ^e University of Sassari, Desertification Research Centre, Sassari, Italy

14 ^f Szent István University, MTA-SZIE Plant Ecology Research Group, Gödöllő, Hungary

15 ^g Thünen Institute of Market Analysis, Braunschweig, Germany

16 ^h University of Liège, Arlon Environment Campus, Arlon, Belgium

17 ⁱ MTA Centre for Ecological Research, Institute of Ecology and Botany, Vácrátót, Hungary

18 ^j Potsdam Institute for Climate Impact Research, Potsdam, Germany

19 ^k French National Institute for Agricultural Research, Modelling Agricultural and Hydrological Systems in the
20 Mediterranean Environment, Avignon, France

21 ^l Rothamsted Research, North Wyke, Okehampton, United Kingdom

22

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

24 ^{††} Currently at: Catholic University of the Sacred Heart, Department of Sustainable Food Production, Piacenza, Italy

25

26 * Corresponding author. 5 chemin de Beaulieu, 63039 Clermont-Ferrand (France);

27 gianni.bellocchi@clermont.inra.fr

28

29 **Abstract**

30 This study presents results from a major grassland model intercomparison exercise, and
31 highlights the main challenges faced in the implementation of a multi-model ensemble
32 prediction system in grasslands. Nine, independently developed simulation models linking
33 climate, soil, vegetation and management to grassland biogeochemical cycles and production
34 were compared in a simulation of soil water content (SWC) and soil temperature (ST) in the
35 topsoil, and of biomass production. The results were assessed against SWC and ST data from
36 five observational grassland sites representing a range of conditions - Grillenburg in
37 Germany, Laqueuille in France with both extensive and intensive management, Monte
38 Bondone in Italy and Oensingen in Switzerland - and against yield measurements from the
39 same sites and other experimental grassland sites in Europe and Israel. We present a
40 comparison of model estimates from individual models to the multi-model ensemble
41 (represented by multi-model median: MMM). With calibration (seven out of nine models), the
42 performances were acceptable for weekly-aggregated ST ($R^2 > 0.7$ with individual models and
43 $> 0.8-0.9$ with MMM), but less satisfactory with SWC ($R^2 < 0.6$ with individual models and
44 $< \sim 0.5$ with MMM) and biomass ($R^2 < \sim 0.3$ with both individual models and MMM). With
45 individual models, maximum biases of about $-5\text{ }^\circ\text{C}$ for ST, $-0.3\text{ m}^3\text{ m}^{-3}$ for SWC and 360 g
46 DM m^{-2} for yield, as well as negative modelling efficiencies and some high relative root mean
47 square errors indicate low model performance, especially for biomass. We also found
48 substantial discrepancies across different models, indicating considerable uncertainties
49 regarding the simulation of grassland processes. The multi-model approach allowed for
50 improved performance, but further progress is strongly needed in the way models represent
51 processes in managed grassland systems.

52

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

54

55 **1. Introduction**

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

64 Considering the role played by grasslands in maintaining food production, grassland
65 biomass yield is an important agro-technical indicator to evaluate the economic viability of
66 grassland-based milk and meat production systems as compared to concentrate feeding (e.g.
67 Schader et al., 2013). In a climate-change context, for instance, adaptation of grasslands to
68 climate change necessarily includes minimizing fluctuations in biomass produced (Collins,
69 1995). Considering the viability of grassland-based systems depending on their ability to
70 produce meat from forage harvested on-farm, it is critical to examine the dynamics of
71 grassland biomass production, where management plays a role by influencing the temporal
72 forage availability and the interactions between herd and grassland.

73 Grassland ecosystem models have become important tools for extrapolating local
74 observations and testing hypotheses on grassland ecosystem functioning (Chang et al., 2013;
75 Graux et al., 2013; Vital et al., 2013; Ma et al., 2015). Under the auspices of the FACCE
76 MACSUR knowledge hub (<http://macsur.eu>), a model intercomparison was conducted using
77 datasets from an observational and experimental network of nine multi-year flux and
78 production sites spread across Europe (France, Italy, Germany, Switzerland, The Netherlands,
79 and United Kingdom) and Israel, and engaging a modelling community using a suite of

80 different models to understand grassland functioning. In particular, the collected datasets of
81 meteorological data, C, energy and water fluxes were used to drive and evaluate the
82 performance of nine grassland models.

83 The identified models are an inventory of modelling approaches made available through
84 the MACSUR consortium and applied worldwide. Grassland-specific approaches were used
85 together with other approaches, mainly conceived to simulate crops and plant functional
86 types. The primary goal of this study is to synthesize and compare the participating grassland
87 models to assess current understanding of soil processes (soil temperature and soil water
88 content, which are fundamental drivers of ecosystem-scale processes) and
89 aboveground/harvested biomass (which is the output of major significance in agricultural
90 production) in Europe and Israel. To achieve this goal, model evaluation against actual
91 measurements was performed before and after model calibration. To the best of authors'
92 knowledge, this is the first model intercomparison performed specifically on permanent
93 grasslands. The present study, focused on grassland sites across Europe and a neighbour
94 country (Israel), extends preliminary analyses (Ma et al., 2014; Sándor et al., 2015), and
95 parallels other initiatives on the comparison of grassland models worldwide, such as the
96 Agricultural Model Intercomparison and Improvement Project (AgMIP, Rosenzweig et al.,
97 2013) and other international projects (Soussana et al., 2015).

98 The present grassland model intercomparison tries to answer five fundamental questions in
99 a multi-site, multi-model framework: (1) are the main drivers of grassland processes
100 represented well by state-of-the-art grassland models?, (2) what is the skill of the studied
101 models considering the different processes?, (3) can calibration improve the models in terms
102 of quality of simulation of different processes?, (4) can the ensemble of model results be used
103 to estimate soil properties and grassland biomass in the study sites?, and (5) what
104 uncertainties are associated with the different models, and how can uncertainty be quantified

105 in a multi-model framework? In addition, areas are identified where structural changes in
106 models may be needed to improve performances and decrease uncertainty of process
107 representation.

108

109 2. Material and methods

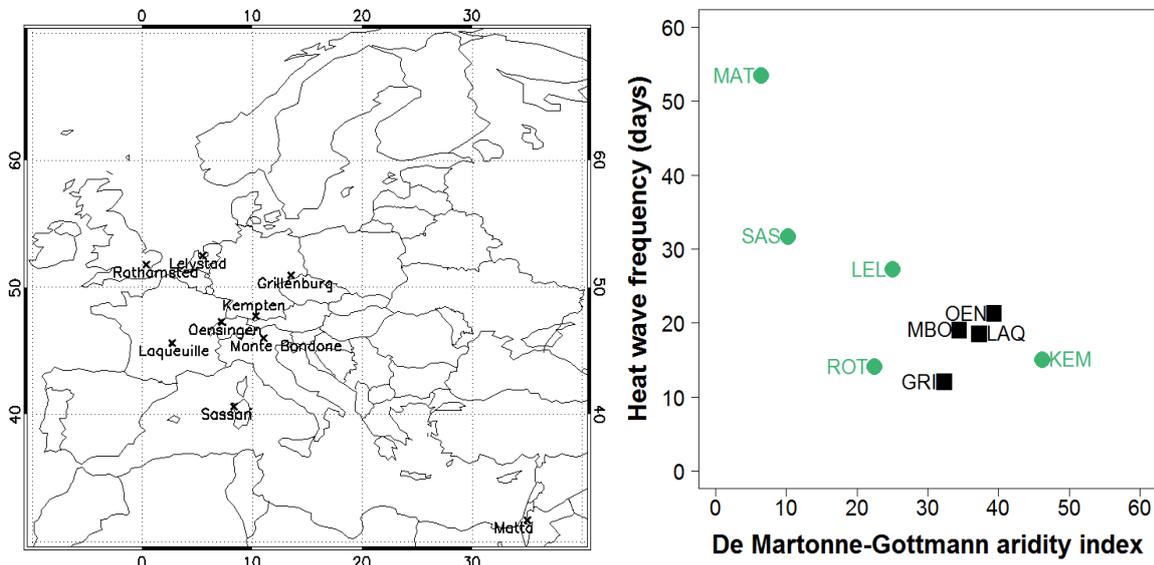
110 2.1. Study sites

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

115

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

120



121 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)
Grillenbug (GRI), Germany	50° 57' N	13° 30' E	380	2004-2008	Flux-tower mowed, extensively managed site.	Prescher et al. (2010)
Kempton (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)

122

123 Four of the study sites (Laqueuille, Monte Bondone, Grillenburg, Oensingen) are
124 equipped with an eddy covariance system to determine the net ecosystem exchange (NEE) of
125 CO₂ and automated weather stations for hourly weather reports. They are essentially old semi-
126 natural grasslands including vegetation types representative of the zone (with the exception of
127 OEN, which was established in 2001). The flux-tower sites are the most data-rich grasslands
128 in Europe, covering a variety of components of grassland ecosystem, including gross primary
129 production (GPP), that is an estimate of the plant production of organic compounds from
130 atmospheric CO₂, and ecosystem respiration (RECO), the latter playing an important role to

131 estimate global C balances of terrestrial ecosystems (by definition $NEE = RECO - GPP$, with
132 positive values indicating the system is a source of C, and negative values indicating that the
133 system takes up C from the atmosphere). The flux-tower sites also record actual
134 evapotranspiration, soil temperature (top 0.1 m) and soil water content (top 0.1 m). The eddy
135 covariance system consists of a fast response 3D sonic anemometer coupled with fast CO₂-
136 H₂O analysers measuring fluxes of CO₂, latent and sensible heat, and momentum fluxes at a
137 30-min time step. The basic data used in this study are at daily resolution to fit the temporal
138 resolution of models. They are the result of a filtering process, quality check and gap filling
139 according to European flux database guidelines (Aubinet et al., 2012). Data are also available
140 on the standing aboveground biomass at given dates. Biomass was measured destructively at
141 given dates in all the study sites (at ground level at Laqueuille, at site-specific canopy heights
142 as part of regular mowing in the other sites).

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

150

151 *2.2. Models description*

152 The first phase of the study was to identify a wide selection of grassland models to be able
153 to represent processes controlling energy, water and C cycle dynamics. The selection phase
154 allowed identifying nine models in which processes are represented with different levels of
155 detail. Whereas some models are empirically based with relatively simple relationships

156 between driver variables and fluxes, others are more complex, simulating the coupled C,
157 nutrient, and water cycles (process-based models). Models also differ in their representation
158 of soil properties, vegetation type, farming practices, and environmental forcing, as well as
159 the initialization of C pools.

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

170

171 *2.3. Simulation study design*

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

177 For the uncalibrated (blind) simulations, the models were run at each site using the
178 available data of weather, soil and management, with no parameter adjustment. After the
179 blind simulations were completed, additional plant and soil information from a sub-set of
180 flux-tower site data was supplied to each modelling group, i.e. the first half of the whole

181 series of available data or the first half plus one in the case of an uneven number of years
182 (Table 1). The information provided were daily time series of GPP, RECO, soil water content,
183 soil temperature, and actual evapotranspiration (some groups only used a subset of
184 observations for calibration). For the same output variables, calibrated simulation results were
185 evaluated against observations from the validation sub-set of years. Biomass data were not
186 used for calibration and held back for validation purpose.

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

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

200 Performance metrics were calculated for four time series: uncalibrated ($U1$, $U2$), calibrated
201 (C) and validated (V) years. $U1$ and C refer to the first half of the whole series of available
202 data (or the first half plus one in the case of an uneven number of years) which was used for
203 calibration, while $U2$ and V refer to the years which were excluded from calibration. Possible
204 improvement of model performance due to calibration was evaluated using the metrics from
205 the $U2$ and V years. This logic was used because validation implies that model performance is

206 assessed with calibration-independent data. Thus, possible improvement of model
207 performance can be most clearly judged by comparing error measures from *U2* and *V*. Multi-
208 site mean (i.e. average data from all sites) error statistics were analysed to quantify the overall
209 effect of model calibration on the simulated processes.

210

211 *2.4. Uncertainty assessment*

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

231 quantified the relationship between the residuals and mean maximum temperature and
232 precipitation sum values of the preceding two weeks relative to the biomass sampling.

233

234 **3. Results**

235 *3.1. Analysis of individual model performance*

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

246

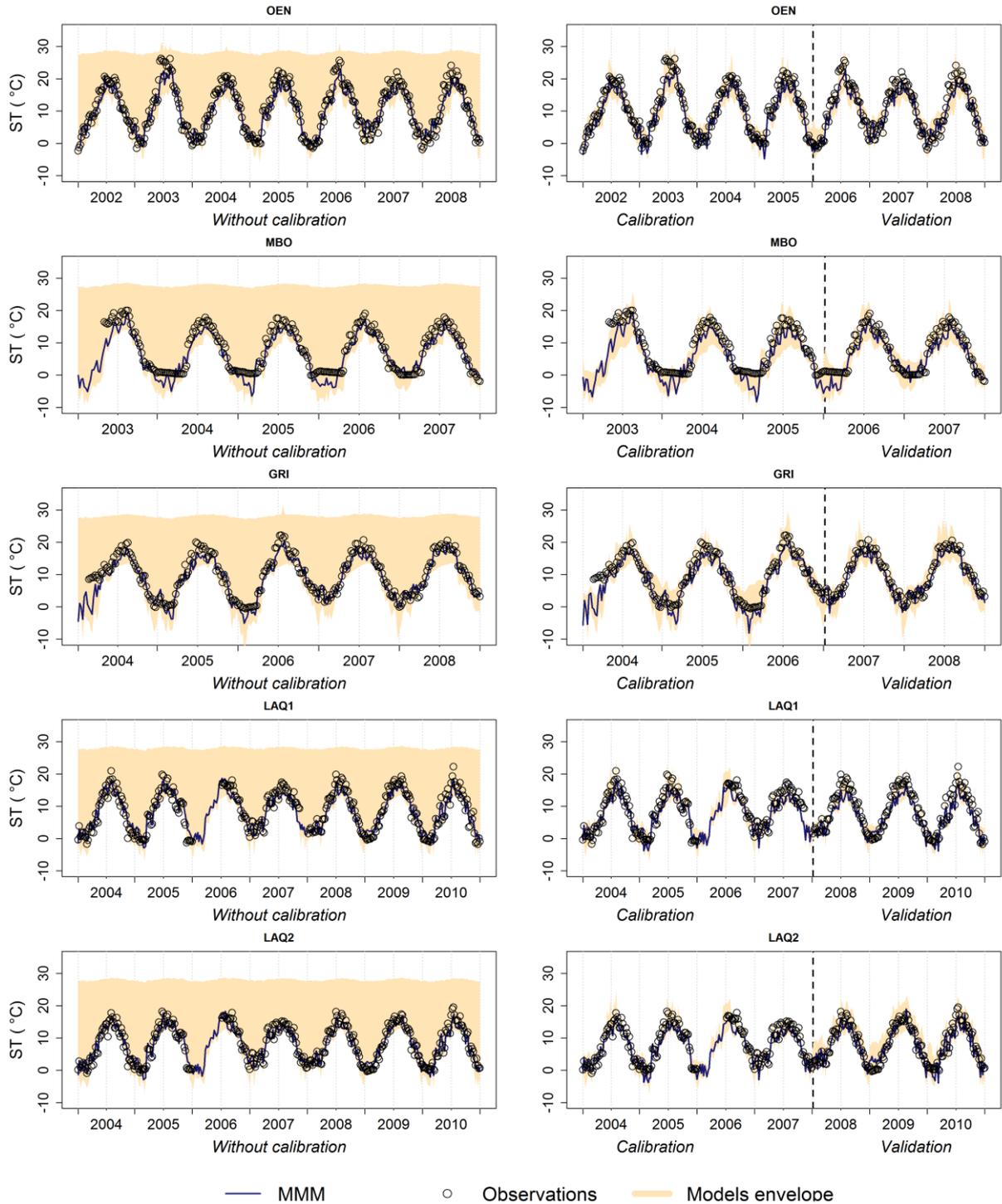
247 *3.1.1. Evaluation of soil temperature (ST) estimates (flux sites)*

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

252

253

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



259
 260 The figure suggests that the range of model results decreased drastically after calibration.
 261 However, it is worth noting that the upper bound in Fig. 2 (left) (almost constant ST around

262 28 °C) is caused by model 8 only, which did not provide results for the calibrated simulations.

263 The rest of the models provided ST values in a more realistic fashion (not shown here).

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

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

273

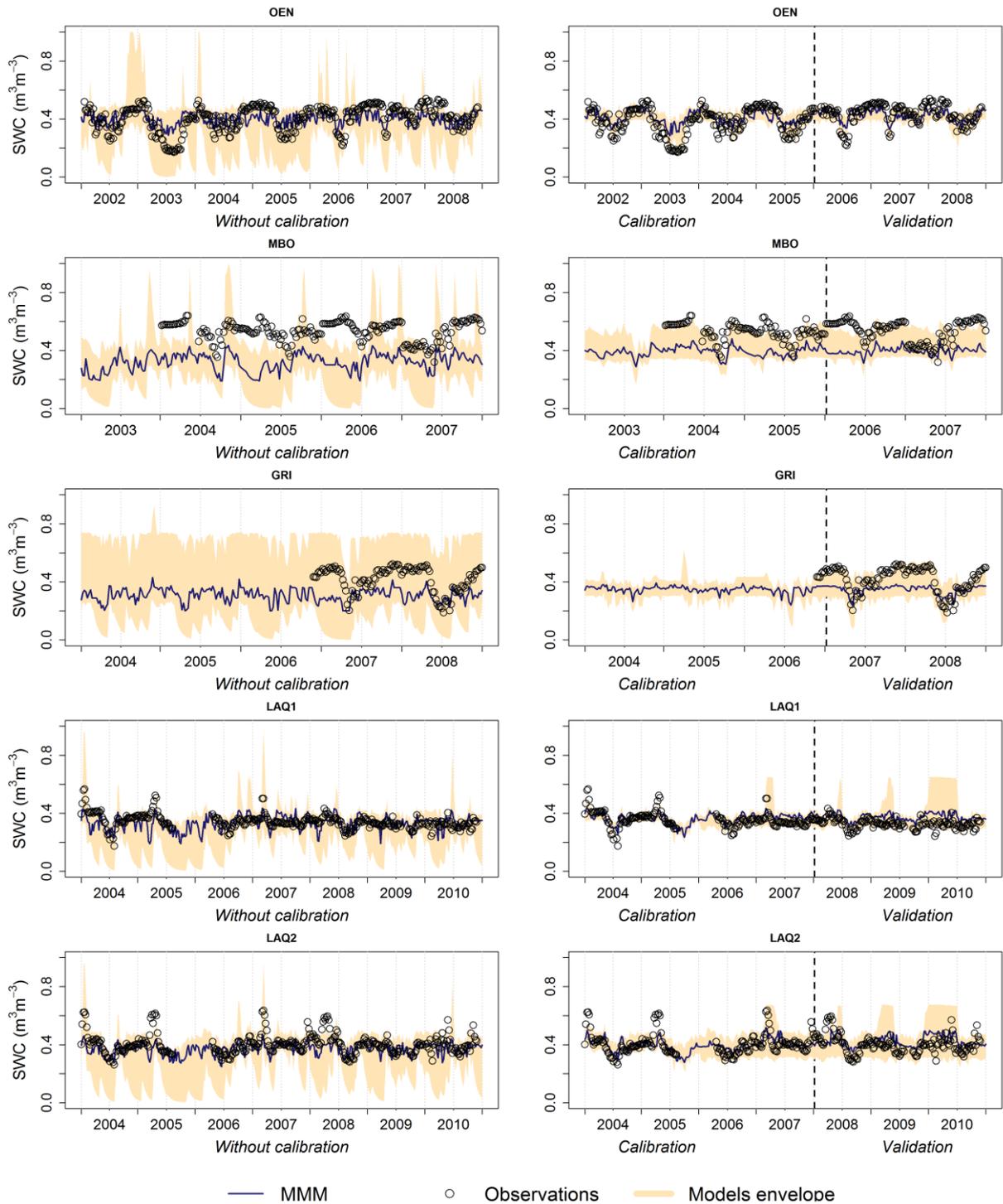
274 *3.1.2. Evaluation of soil water content (SWC) estimates (flux sites)*

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

280

281

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



287

288

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

303 Figs. K-O (Supplementary material) show the performance of the individual grassland
304 models for both blind (nine models) and calibrated simulations (seven models). The results
305 clearly show that systematic errors are present in all models. An interesting common error of
306 the models is that the range of simulated SWC values is smaller than in reality (model 8 is
307 exception). The scatterplots in Supplementary material also reveal that the above-mentioned
308 wide range of model results (e.g. Figs. K1 and K2 for Oensingen) is caused by model 8 alone
309 (in Fig. K2, the x- and y-axis ranges are smaller than in Fig. K1 because of the smaller overall
310 range of SWC values.). The scatterplot indicate some improvement (remarkable with models
311 5 and 6) in the simulation of SWC in terms of R^2 . However, model calibration was globally
312 unable to address the systematic errors present in the blind tests.

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

323

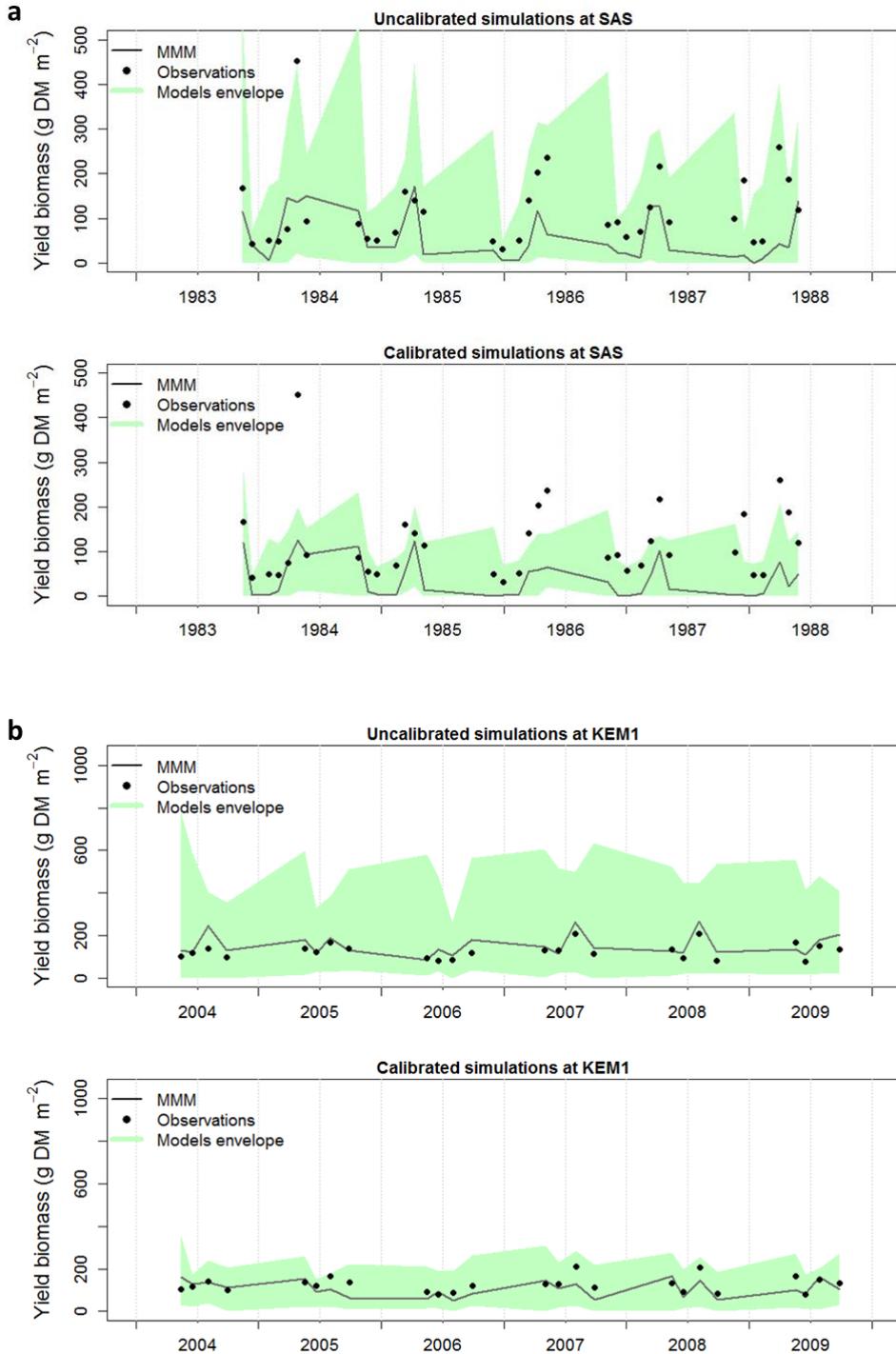
324 *3.1.3. Evaluation of plant biomass estimates*

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

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

338

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



344
 345

346

347

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

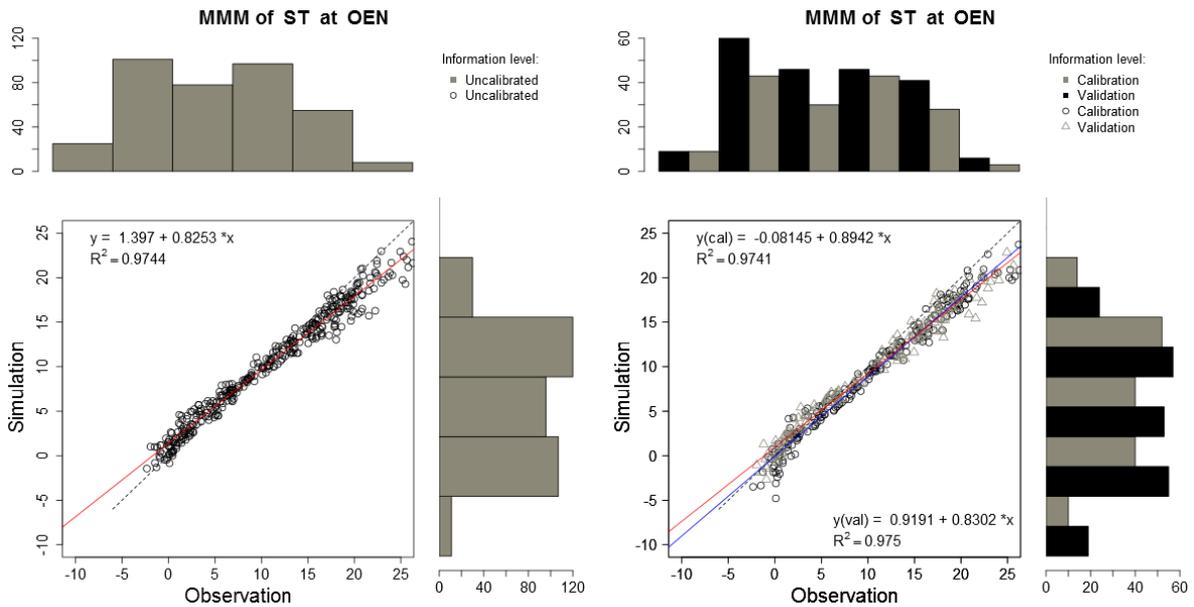
353 Overall, considering all sites and models (see also Supplementary material, Figs. Q1-S20),
354 underestimation of biomass is more common than overestimation. Data points are distributed
355 around the 1:1 line for ~1/3 of all model-site combinations that reported results. There is no
356 clear systematic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

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

373 3.2. Analysis of the ensemble approach

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

377
 378 Fig. 5. Multi-model median (MMM) of uncalibrated (left) and calibrated-validated (right) soil
 379 temperature (ST) simulations compared with observed values with weekly resolution at OEN
 380 site (ID as in Table 1): x-y scatterplots with associated x and y histograms with estimated
 381 densities.



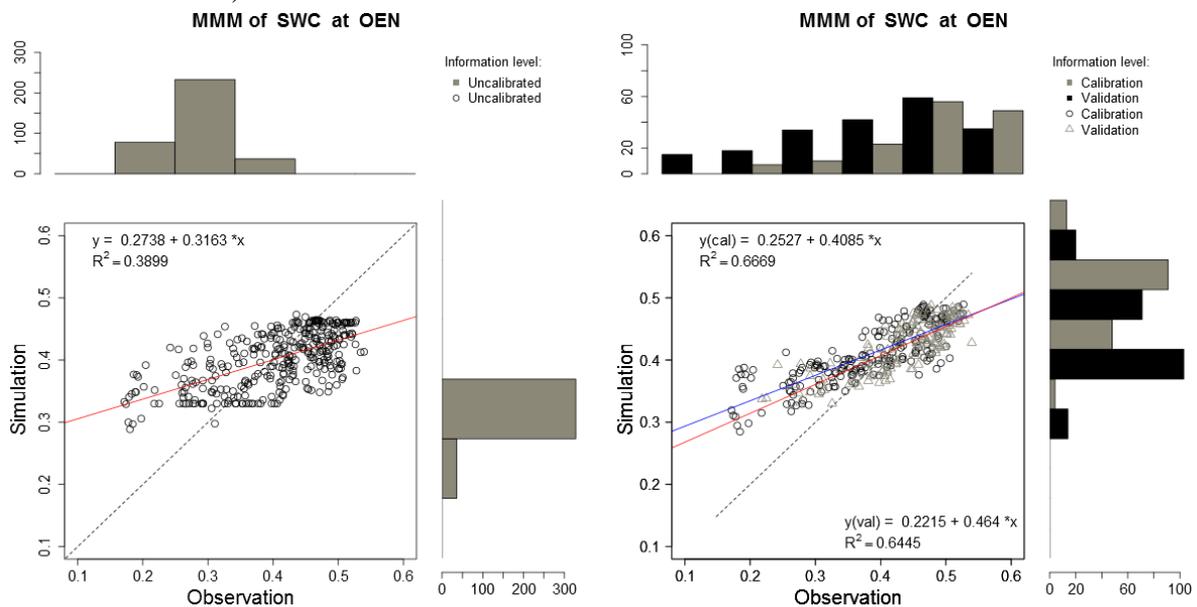
382

383

384 The figures indicate that MMM ST from the blind simulations provided reliable estimates
 385 in terms of explained variance and slope of the linear regression. Explained variance varied
 386 between 91 and 97%, while the slope varied between 0.83 and 0.92 (which means small
 387 underestimation by the ensemble). Calibration did not change the overall quality of the
 388 MMM. Explained variance changed slightly with very small overall decrease, while the slope
 389 became closer to the 1:1 line in some cases. The performance indicators were calculated using
 390 the U_2 and V years only. Considering ME, the MMM ST taken from the blind runs was a
 391 better predictor than 62.5% of the models. After calibration, 71% of the models gave worse
 392 ME than the MMM. Considering the explained variance, blind MMM ST was better than any

393 of the models, while after calibration 86% of the models provided worse performance than the
 394 ensemble median. Fig. 6 shows the comparison of the measured and the simulated MMM
 395 SWC results (separately for the uncalibrated and the calibrated-validated runs) at OEN, which
 396 is the best site in terms of MMM SWC performance (see, for other sites, Figs. U1-U4 in
 397 Supplementary material).

398
 399 Fig. 6. Multi-model median (MMM) of uncalibrated (left) and calibrated-validated (right) soil
 400 water content (SWC) simulations compared with observed values with weekly resolution at
 401 OEN site (ID as in Table 1): x-y scatterplots with associated x and y histograms with
 402 estimated densities).



403
 404
 405 The results indicate that MMM SWC inherits the problems associated with the individual
 406 models. MMM SWC constructed from the blind simulation results shows poor performance at
 407 all sites. Low explained variance (maximum $R^2 \sim 0.4$ at OEN) and departure of the data from
 408 the 1:1 line are indicators of the low reliability of simulations. The range of simulated
 409 ensemble SWC values is smaller than in reality, similarly to the results obtained with the
 410 individual models. After calibration, the quality of the MMM SWC simulations was mainly
 411 improved, though the performance of the validated and calibrated years differed markedly in
 412 some cases. Explained variance increased for all five sites, and ranged between 11% (LAQ2,

413 validated years) and 73% (OEN, calibrated years). The simulated MMM SWC remained
414 confined within a relatively narrow range for all sites, which means that the intra-annual
415 variability of SWC was not captured by the MMM. Similarly to ST, multi-site mean error
416 statistics were calculated and compared with the multi-site mean statistical indicators of the
417 MMM SWC (for the *U2* and *V* years). ME of the MMM SWC was better than 78% of the
418 models and 57% of the models for the blind and calibrated simulations, respectively. Multi-
419 site mean ME remained negative for all models in both time periods (*U2* and *V*), which means
420 that the mean of the observations is more useful for SWC estimation than any of the models.

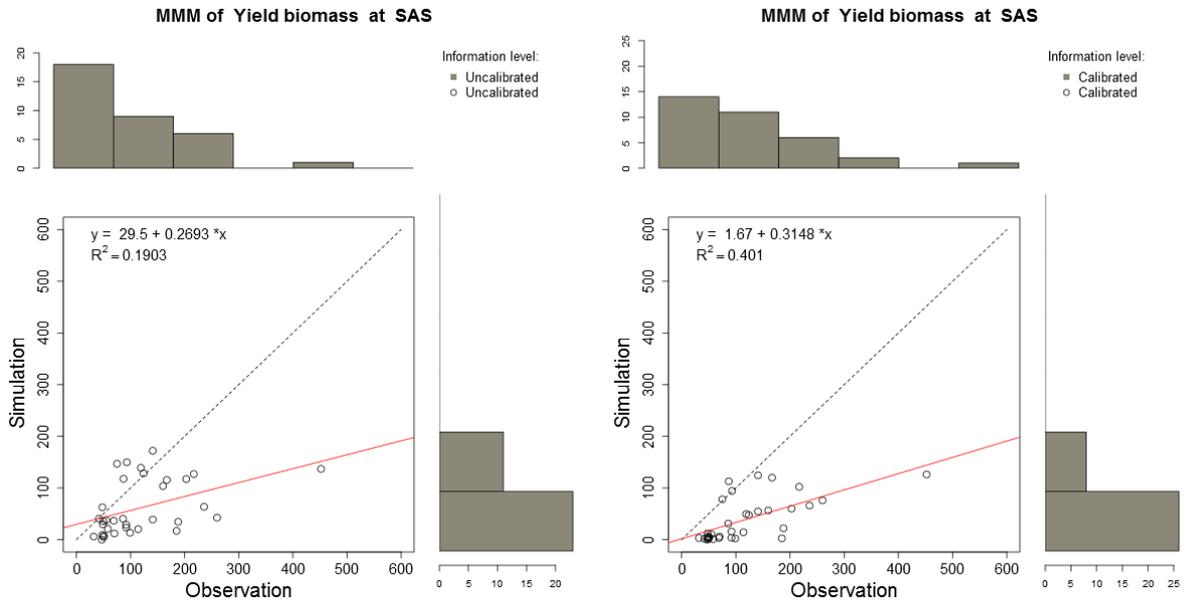
421 Fig. V (Supplementary material) shows that after calibration better estimations in yield
422 were reached at the grassland sites other than the flux sites. In general, the MMM
423 underestimated the expected yield at the production sites but overestimated it at the flux sites.
424 Additionally, the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

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

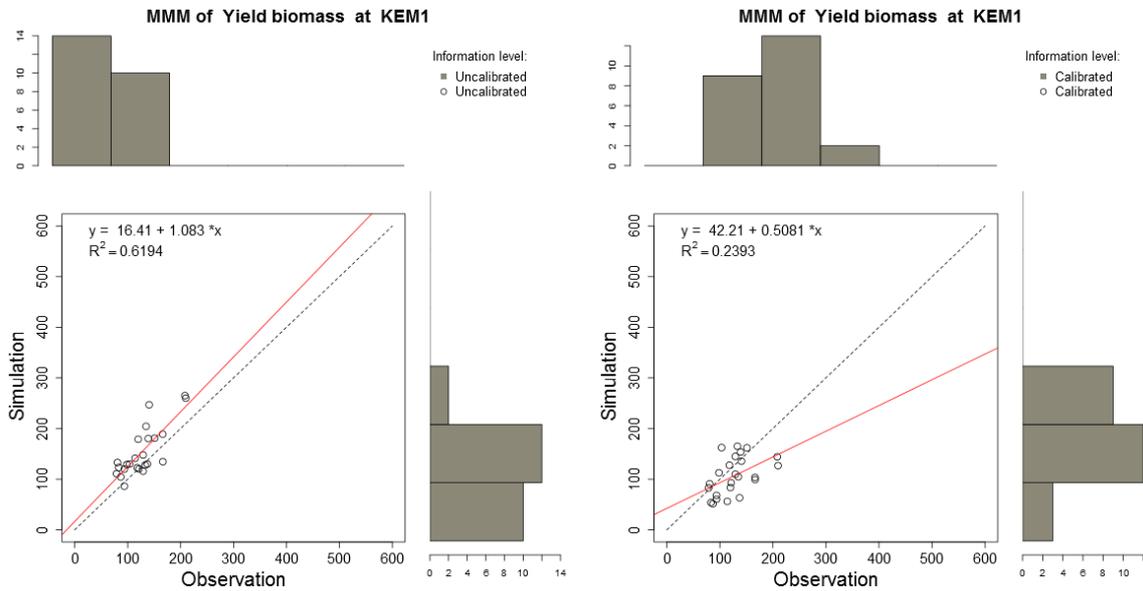
429

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

a



b



439 The figures indicate that the performance of the MMM biomass estimation changed from
440 site to site. Interestingly, the pattern on the scatterplots is similar for the blind and calibrated
441 ensembles, which means that parameter adjustment did not cause radical change on the
442 overall performance of the multi-model ensemble. With a few exceptions, systematic over- or
443 underestimation is typical. Explained variance varies considerably among sites. With respect
444 to ME, MMM outperformed the individual models in 100% of the cases. In terms of R^2 , the
445 MMM gave better explained variance than seven out of the nine models (78%) for the blind
446 runs, while MMM outperformed five models (out of seven) for the calibrated simulations
447 (71%).

448

449 *3.3. Relationship between model errors and uncertainty assessment*

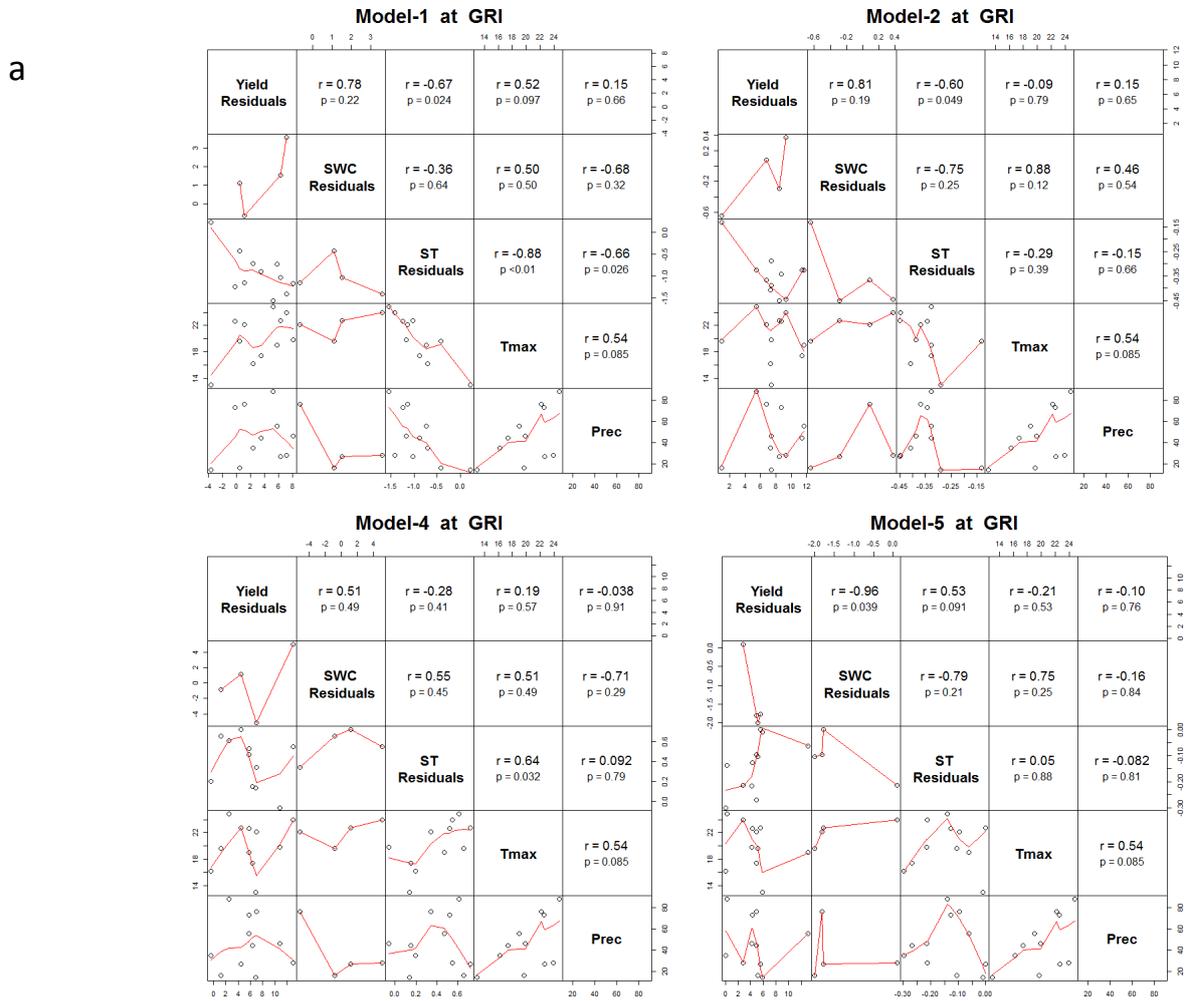
450 *3.3.1. Relationship between residuals*

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

457

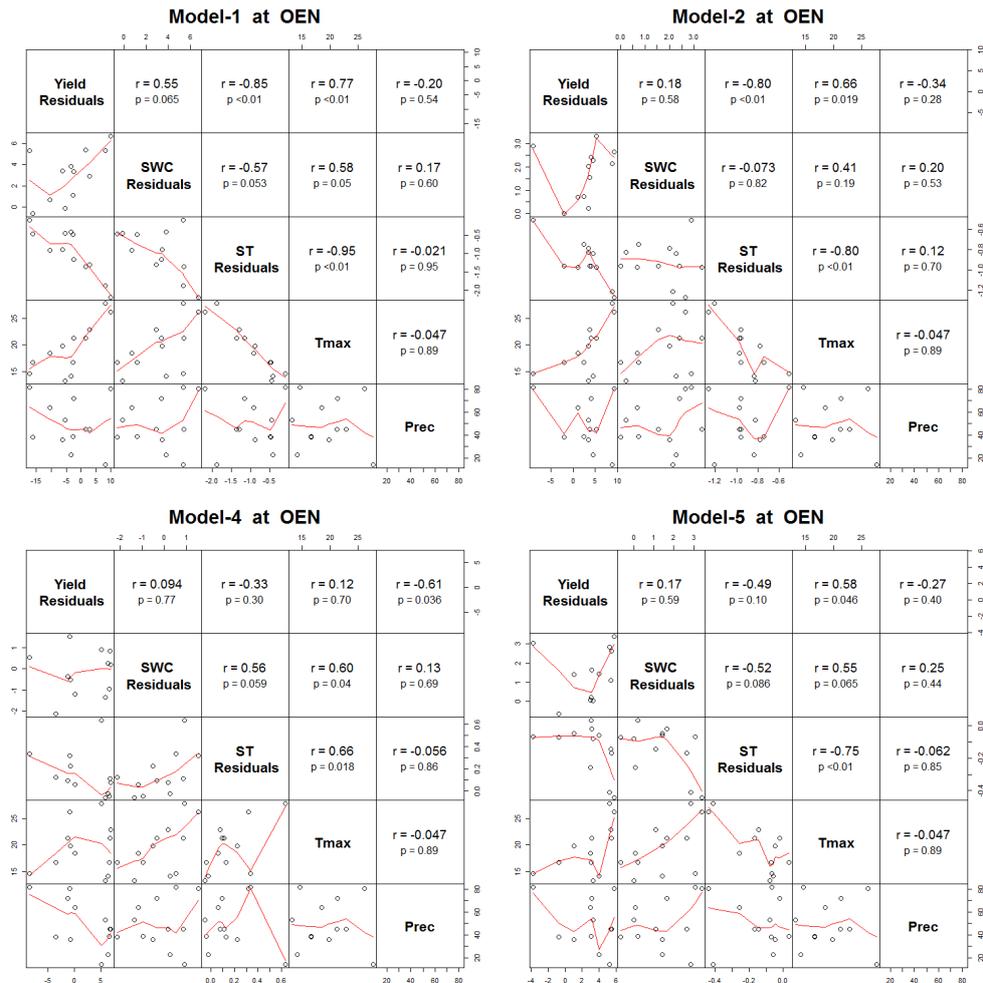
458

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



463

b



464

465

466 The figures visualize the relationship between the selected variables as squared matrix-like
 467 configurations. The lower triangular part of the squared matrices shows the scatterplots
 468 between the specific variables defined in the main diagonal of the matrix, with the overlying
 469 spline (without inferential character). For readability, the correlation between the variables
 470 and the significance of the relationship (p value) are shown in the upper triangular part of the
 471 matrix. The figures show that at some sites (mostly at GRI and OEN) a relatively strong
 472 relationship exists between some of the residuals, and also between the environmental factors
 473 and the residuals (relationship between maximum temperature and precipitation is not
 474 informative in the present context). The existing relationship is not uniform and, in some
 475 cases, the correlation is negative between some of the residuals (e.g. relationship between

476 yield and SWC residuals at GRI for model 5). Considering that the number of available SWC
477 residuals at GRI is low, the statistical comparison is not well justified here for SWC.

478 In the followings, we focus mainly on GRI and OEN sites. The individual models show
479 considerably differences in terms of relationship between the yield, the SWC and the ST
480 standardized residuals. High positive correlation was established between the yield and SWC
481 residuals for models 1, 2 and 4, whilst models 5 and 6 had a strong negative correlation at
482 Grillenburg, which is the northern flux site (Fig. 8 a and Fig. Y1 in Supplementary material).
483 Similarly, positive correlation characterizes the relationship between yield and SWC residuals
484 at OEN, but the relationship is weaker than at the GRI site (Fig. 8b and Fig. Y1 in
485 Supplementary material). We found a general negative correlation between the yield and ST
486 residuals, with the exception of models 5, 6 and 7 (Fig. Y1 in Supplementary material), as
487 well as between the ST and SWC residuals (except for model 4) at all sites (the correlation
488 was moderate at the grazed sites; see Figs. Y1 and Y2 in Supplementary material).
489 Meteorological factors such as the mean maximum temperature and precipitation (2-weeks
490 means and totals, respectively) also had a notable effect on the residuals. In some cases there
491 was no clear pattern among the sites. The relationship between the selected variables can be
492 alternatively characterized as well. We can select an arbitrary (but high enough) absolute
493 minimum threshold and identify the number of cases when the covariance equals or exceeds
494 this expected minimum in absolute terms. Selecting the 0.66 correlation threshold (which
495 represents ~44% explained variance), and considering only OEN and GRI, the most common
496 relationship is the ST residual - maximum temperature, which is typical for models 1, 2, 4, 5
497 and 6. The second most common feature is the SWC residual - yield residual relationship,
498 which is present in the case of models 1, 2, 5 and 6. Strong precipitation - SWC residual,
499 maximum temperature - SWC residual and ST residual - SWC residual relationships are
500 present for three models. Maximum temperature - yield residual and ST residual - yield

501 residual relationships were strong for two models. The correlation between the other possible
502 variable combinations did not reach the 0.66 threshold for GRI and OEN. Though the multi-
503 model medians of ST, SWC and yield are statistically-derived datasets, and not the result of a
504 process-based model, it might be interesting to check their behaviour in terms of correlation
505 between MMM residuals, and also the effect of environmental variables on the residuals. The
506 MMM correlations were generally moderate probably owing to the decreased model
507 uncertainty (Fig. Y5 in Supplementary material). We found a general negative correlation
508 between the SWC and ST residuals, while the maximum temperatures were positively
509 correlated with the SWC and negatively with the ST residuals at all sites (the highest
510 correlation was characteristic to the GRI and OEN sites). These results are in accordance with
511 our previous finding, namely that the MMM approach may give a better estimation than the
512 individual models (here in terms of unexpected correlation between the residuals).

513

514 *3.3.2. Uncertainty assessment related to multi-model ensemble*

515 Appendix 4 shows, for both individual models and MMM, the ratios between the
516 variability of the models envelope and standardized model residuals. Values greater than one
517 indicate that the spread is larger than the model residual, i.e. the uncertainty associated with
518 the ensemble of models is high. For ST, ratios >1 indicate that with both individual models
519 (90%) and MMM (100%) model error was generally lower than the variability in the multi-
520 model ensemble (with ratio equal to 1, M1 at LAQ1 is the only exception). With SWC, the
521 pattern of responses is more complex, ranging from ratios <1 with M1 at all sites to ratios >1
522 with M6 and M7, and mixed situations with the other models and MMM (overall ratios >1 are
523 68% with individual models and 60% with MMM). This complexity is also reflected in the
524 yield responses (ratios >1 are 54% with individual models and 58% with MMM), where only

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

527

528 **4. Discussion**

529 *4.1. Soil temperature (ST)*

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

541

542 *4.2. Soil water content (SWC)*

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

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

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

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

570 The utility of the MMM SWC estimation is not as straightforward as in the case of ST.
571 Ensemble median of the blind results usually performs better than 2/3 of the models (with the
572 exception of R^2), which means that some benefit can be expected by using an ensemble
573 approach. Considering the calibrated models, the number of models that are outperformed by
574 the median is decreased. These results indicate the usefulness of the ensemble approach

575 though the performance of the MMM still indicates several areas of improvement. In
576 summary, the results indicate that SWC estimation should be used with caution in regional or
577 continental scale simulations, and model developments focusing on soil hydrology are
578 essential.

579

580 *4.3. Plant biomass*

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

590 In our model intercomparison, calibration was performed using eddy covariance based on
591 C flux and evapotranspiration data, together with SWC and ST (but some modelling groups
592 only used a subset of measured data for calibration). Thus, biomass data were not used as a
593 control variable for model optimization, which means that errors associated with the proper
594 estimation of biomass can partly be explained by the lack of adjustments of some internal
595 model parameters associated with biomass. Multi-objective model calibration should be
596 extended to include biomass as a control variable with equal weight as the other, sometimes
597 more data-rich data streams like GPP (Keenan et al., 2011). Besides uncertainty associated
598 with the model parameters, structural problems might also affect the performance of models
599 on yield. For example, constant ratios of the above- to below-ground biomass allocation may

600 cause unsatisfactory model performance on biomass. Ensemble simulation of grassland
601 production is an opportunity as shown in the present study. Uncalibrated ensemble median
602 was the most successful in terms of error statistics, in spite of the fact that the quality of the
603 performance based on the median was still problematic at almost all the sites. Due to
604 calibration, the multi-model median was still useful.

605

606 *4.4. Ensemble approach of grassland simulation*

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

619

620 *4.5. Possible explanations for model errors (residual analysis)*

621 We presented an approach that uses a covariance matrix (with graphical representation) to
622 take into account all possible correlations between ST, SWC and yield residuals and,
623 additionally, mean maximum air temperatures and precipitation totals. This residual analysis
624 can help find relationships between some variables, and between variables and external

625 drivers (and thus it can help find additional variables that may need to be included in the
626 models as predictors; Medlyn et al., 2005). This analysis might indicate dependency of errors
627 in one process that is related to another (which is a typical case of error propagation within the
628 model), though the way of error propagation cannot be easily retrieved from the covariance
629 matrix. For example, overestimation of biomass may cause overestimated shading of the soil
630 surface that interferes with the ST simulation. In turn, bias in ST may interact with ecosystem
631 respiration that affects plant growth and thus biomass amount. Underestimation of leaf
632 biomass may interact with evapotranspiration (by decreasing it) which can cause errors in
633 SWC due to slower water depletion. SWC effect on biomass is probably more
634 straightforward. The results indicated that the SWC annual cycle is not well represented by
635 model simulations and, hence, drought stress on plant growth and biomass could not be
636 captured by models. This is particularly well illustrated at GRI.

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

646

647 *4.6. Uncertainties in grassland modelling*

648 Uncertainty of output data, defined as spread of results arising from unknown or
649 imperfectly characterized processes, is an inherent property of mathematical modelling. In

650 grassland modelling and, generally, in ecological modelling, uncertainty is caused by internal
651 variability, errors in the initial and boundary conditions, parameterization, and model
652 structure. In multi-model frameworks, uncertainty is also associated with the different model
653 formulations (Schwalm et al., 2015).

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

668 Grassland models typically use many parameters (i.e. constants) that are variables in
669 reality, which substantially alter the biophysical and biogeochemical processes. In many
670 cases, these parameters are hard to define due to lack of measurement (e.g. for plant traits like
671 leaf C:N ratio or specific leaf area), or due to the nature of the parameter (e.g. in empirical
672 equations without physical meaning). Thus, model calibration is essential to optimize model
673 results for a given ecosystem. However, parameters are highly variable in time and space (e.g.
674 Zaehle et al., 2005), thus their general applicability as one defined plant functional type (PFT,

675 Bonan et al., 2002) is problematic. Grassland models can simulate management in such a way
676 that the user prescribes the management related data to the model (e.g. Hidy et al., 2012).
677 However, due to the nature of management the settings are often affected by uncertainties. A
678 typical example is grass cutting, or grazing. Within the present model intercomparison, yield
679 simulation was rather unsuccessful at the grazed site (LAQ1 and LAQ2; Figs. R13 and R14 in
680 the Supplementary material), which can be the consequence of management-related
681 uncertainty. Individual grassland models are constructed using diverse representations of
682 specific processes (Table B1 in Supplementary material). Though there are similarities in the
683 applied methods (e.g. the Penman-Monteith method is used usually for evapotranspiration
684 simulation), the heterogeneity of the process representations is obvious. Scientific level of
685 understanding of plant processes is far from being perfect. Here we mention a few processes
686 that are widely discussed in the literature.

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

694 Although grassland models typically have some kind of representation of drought related
695 senescence and changes of plant functioning due to water limitation and/or heat, this logic is
696 still based on the above-mentioned PFT logic. Van der Molen et al. (2011) suggested that
697 grassland ecosystems cannot be considered as a single PFT but should be treated as mixtures
698 of plants with different plant strategic properties. For example, at the drought-prone Bugac-
699 puszta site in Hungary (Nagy et al., 2007), observations revealed that C3 grasses dominate the

700 spring/early summer intensive growth, then during the summer drought resistant C4 grass
701 species start to interact with the overall C balance also due to their delayed phenological cycle
702 at this extensively managed sandy grassland (Nagy Z., personal communication). None of the
703 studied grassland models is at present prepared to represent this strategy for mixtures of
704 grassland species.

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

708

709 **5. Conclusions and future directions**

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

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

725 The results indicated that some of the main drivers and results of the grassland processes
726 are not represented well by state-of-the-art grassland models. Especially SWC and yield had
727 severe problems that may prevent their applicability in reliable, larger scale experiments.
728 Model errors were presented for the studied processes in a tabular form, which may provide
729 comparability basis for further studies. Presentation of daily, weekly and monthly results
730 might be useful for other researchers to compare model performance at the same sites.
731 Calibration seemed to improve the model results to some extent, but there was no dramatic
732 increase in model performance for any of the studied models, at any of the sites. Ensemble
733 technique seems to be a feasible method for the simulation of grassland processes, but model
734 development is inevitable to improve the multi-model approach. In our intercomparison, we
735 highlighted the uncertainties that are associated with the models, and we created
736 recommendations to some of the models. Uncertainty was characterized in a fashion, which
737 allowed highlighting the scientific challenges faced in simulating soil processes (temperature
738 and water content) and biomass on European and peri-European grasslands with a variety of
739 state-of-the art models used individually or within an ensemble. What seems to be a message
740 from our intercomparison is that grassland models should be further developed and tested at a
741 large number of experimental sites. In order to provide validation and calibration data for the
742 models, essential processes and outputs like GPP, RECO, SWC, ST, C allocation, emission of
743 non-CO₂ GHGs, and also magnitude and timing of human intervention should be
744 characterized in systematic and accurate fashion in multiple grassland sites covering large
745 climatic gradients.

746 Though the exercise of the presented model intercomparison performed (the first on
747 permanent grasslands) is large enough, we are aware that it does not completely cover most of
748 the modelling approaches used to simulate grasslands. An example is the process-based,
749 biogeochemical model ORCHIDEE-GM, which includes an enhanced representation of

750 grassland management derived from PaSim (Chang et al., 2013, 2015). Another example is
751 represented by a grassland-specific model derived from STICS (BioMA-Grassland, personal
752 communication by G. De Sanctis, Joint Research Centre of the European Commission, Ispra,
753 Italy), which is being developed for the platform BioMA (Biophysical Models Applications,
754 <http://bioma.jrc.ec.europa.eu>). Grassland model intercomparisons with the inclusion of more
755 models should therefore be continued to improve our ability to simulate grassland processes
756 with acceptable quality. We also think that further analyses and better understanding of these
757 ensembles are required to achieve fundamental progress in grassland modelling by
758 investigating the sensitivity of models to climate and management drivers. This assessment
759 goes beyond the scope of this paper, and a paper on this topic should be arranged later as a
760 natural evolution of what has already been presented here.

761

762 **Acknowledgements**

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

786

787 **References**

- 788 Ammann, C., Flechard, C.R., Leifeld, J., Neftel, A., Fuhrer, J., 2007. The carbon budget of
789 newly established temperate grassland depends on management intensity. *Agr. Ecosyst.*
790 *Environ.* 121, 5-20. doi:10.1016/j.agee.2006.12.002
- 791 Asseng, S., Ewert, F., Rosenzweig, C., Jones, J.W., Hatfield, J.L., Ruane, A., Boote, K.J.,
792 Thorburn, P., Rötter, R.P., Cammarano, D., Brisson, N., Basso, B., Martre, P., Aggarwal,
793 P.K., Angulo, C., Bertuzzi, P., Biernath, C., Doltra, J., Gayler, S., Goldberg, R., Grant, R.,
794 Heng, L., Hooker, J.E., Hunt, L.A., Ingwersen, J., Izaurrealde, R.C., Kersebaum, K.C.,
795 Müller, C., Naresh Kumar, S., Nendel, C., O’Leary, G., Olesen, J.E., Osborne, T.M.,
796 Palosuo, T., Priesack, E., Ripoche, D., Semenov, M.A., Shcherbak, I., Steduto, P., Stöckle,
797 C.O., Stratonovitch, P., Streck, T., Supit, I., Travasso, M., Tao, F., Waha, K., Wallach, D.,
798 White, J.W., Wolf, J., 2013. Uncertainties in simulating wheat yields under climate
799 change. *Nat. Clim. Change* 3, 827-832. doi:10.1038/nclimate1916
- 800 Aubinet, M., Vesala, T., Papale, D., 2012. *Eddy covariance: A practical guide to measurement*
801 *and data analysis.* Springer, Dordrecht.
- 802 Bonan, G.B., Levis, S., Kergoat, L., Oleson, K.W., 2002. Landscapes as patches of plant
803 functional types: An integrating concept for climate and ecosystem models. *Global*
804 *Biogeochem. Cy.* 16, 5.1–5.23. doi:10.1029/2000GB001360
- 805 Cavallero, A., Talamucci, P., Grignani, C., Reyneri, A., Ziliotto, U., Scotton, M., Bianchi,
806 A.A., Santilocchi, R., Basso, F., Postiglione, L., Carone, F., Corleto, A., Cazzato, E.,
807 Cassaniti, S., Cosentino, S., Litrico, P.G., Leonardi, S., Sarno, R., Stringi, L., Gristina, L.,
808 Amato, G., Bullitta, P., Caredda, S., Roggero, P.P., Caporali, F., D’Antuono, L.F., Pardini,
809 A., Zagni, C., Piemontese, S., Pazzi, G., Costa, G., Pascal, G., Acutis, M., 1992.
810 Caratterizzazione della dinamica produttiva di pascoli naturali italiani. *Rivista di*
811 *Agronomia* 26, n. 3 suppl., 325-343. (in Italian)

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

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

821 Chang, J., Viovy, N., Vuichard, N., Ciais, P., Campioli, M., Klumpp, K.,
822 Martin, R., Leip, A., Soussana, J., 2015. Modelled changes in potential grassland
823 productivity and in ruminant livestock density in Europe over 1961–2010. *PLoS One*
824 10, e0127554. doi:10.1371/journal.pone.0127554554

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

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

829 De Martonne, E., 1942. Nouvelle carte mondiale de l'indice d'aridité. *Annales de Géographie*
830 51, 242–250 (in French).

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

834 Golodets, C., Sternberg, M., Kigel, J., Boeken, B., Henkin, Z., Seligman, N.G., Ungar, D.E.,
835 2013. From desert to Mediterranean rangelands: will increasing drought and inter-annual

836 rainfall variability affect herbaceous annual primary productivity? *Climatic Change* 119,
837 785-798. doi:10.1007/s10584-013-0758-8

838 Graux A.-I., Bellocchi G., Lardy R., Soussana J.-F., 2013. Ensemble modelling of climate
839 change risks and opportunities for managed grasslands in France. *Agr. Forest Meteorol.*
840 170, 114-131. doi:10.1016/j.agrformet.2012.06.010

841 Hidy, D., Barcza, Z., Haszpra, L., Churkina, G., Pintér, K., Nagy, Z., 2012. Development of
842 the Biome-BGC model for simulation of managed herbaceous ecosystems. *Ecol. Model.*
843 226, 99-119. doi:10.1016/j.ecolmodel.2011.11.008

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

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

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

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

855 Ma, S., Acutis, M., Barcza, Z., Ben Touhami, H., Doro, L., Hidy, D., Köchy, M., Minet, J.,
856 Lellei-Kovács, E., Perego, A., Rolinski, S., Ruget, F., Seddaiu, G., Wu, L., Bellocchi, G.
857 2014. The grassland model intercomparison of the MACSUR (Modelling European
858 Agriculture with Climate Change for Food Security) European knowledge hub. In: Ames,
859 D.P. Quinn, N. (Eds.) *Proceedings of the 7th International Congress of the Environmental*

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

862 Ma, S., Lardy, R., Graux, A.-I., Ben Touhami, H., Klumpp, K., Martin, R., Bellocchi, G.,
863 2015. Regional-scale analysis of carbon and water cycles on managed grassland systems.
864 *Environ. Modell. Softw.* 72, 356-371, doi:10.1016/j.envsoft.2015.03.007.

865 Mapfumo, E., Naeth, M.A., Baron, V.S., Dick, A.C., Chanasyk, D.S., 2002. Grazing impacts
866 on litter and roots: perennial versus annual grasses. *Journal of Range Management* 55, 16-
867 22.

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

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

880 Medlyn, B.E., Robinson, A.P., Clement, R., McMurtrie, R.E., 2005. On the validation of
881 models of forest CO₂ exchange using eddy covariance data: some perils and pitfalls. *Tree*
882 *Physiol.* 25, 839-857. doi:10.1093/treephys/25.7.839

883 Nagy Z., Pintér K., Czóbel Sz., Balogh J., Horváth L., Fóti Sz., Barcza Z., Weidinger T.,
884 Csintalan Zs., Dinh N.Q., Grosz B., Tuba Z., 2007. The carbon budget of a semi-arid

885 grassland in a wet and a dry year in Hungary. *Agr. Ecosyst. Environ.* 121, 21-29.
886 doi:10.1016/j.agee.2006.12.003

887 Peeters, A., 2012. Past and future of European grasslands. The challenge of the CAP towards
888 2020. In: Goliński, P., Warda, M., Stypiński, P. (Eds.), Proceedings of the 24th General
889 Meeting of the European Grassland Federation, 3-7 June, Lublin, pp. 17-32.
890 <http://www.europeangrassland.org/fileadmin/media/EGF2012.pdf>

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

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

897 Richardson, A. D., Hollinger, D. Y., Burba, G. G., Davis, K. J., Flanagan, L. B., Katul, G. G.,
898 Munger, J. W., Ricciuto, D. M., Stoy, P. C., Suyker, A. E., Verma, S. B., and Wofsy, S.C.,
899 2006. A multi-site analysis of random error in tower-based measurements of carbon and
900 energy fluxes. *Agr. Forest Meteorol.* 136, 1–18. doi: 10.1016/j.agrformet.2006.01.007

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

906 Sándor, R., Ma, S., Acutis, M., Barcza, Z., Ben Touhami, H., Doro, L., Hidy, D., Köchy, M.,
907 Lellei-Kovács, E., Minet, J., Perego, A., Rolinski, S., Ruget, F., Seddaiu, G., Wu, L.,
908 Bellocchi, G., 2015. Uncertainty in simulating biomass yield and carbon-water fluxes from

909 grasslands under climate change. *Advances in Animal Biosciences* 6, 49-51.
910 doi:10.1017/S2040470014000545

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

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

920 Schröpel, R., Diepolder, M., 2003. Auswirkungen der Grünlandextensivierung auf einer
921 Weidelgras-Weißklee-Weide im Allgäuer Alpenvorland. Schule und Beratung, Heft
922 11/2003, Seite III-13 bis III-15; Bayerisches Staatsministerium für Landwirtschaft und
923 Forsten, Munich. (in German)

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

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

932 Soussana, J.F., Ehrhardt, F., Conant, R., Harrison, M., Lieffering, M., Bellocchi, G., Moore,
933 A., Rolinski, S., Snow, V., Wu, L., Ruane, A., 2015. Projecting grassland sensitivity to

934 climate change from an ensemble of models. In: Abstract Book of “Our Common Future
935 Under Climate Change” International Scientific Conference, July 7-10, Paris, K-2223-02.
936 http://pool7.kermeet.com/C/ewe/ewex/unesco/DOCS/CFCC_abstractBook.pdf (accessed
937 18.11.2015)

938 van der Molen, M.K., Dolman, A.J., Ciais, P., Eglin, T., Gobron, N., Law, B.E., Meir, P.,
939 Peters, W., Phillips, O.L., Reichstein, M., Chen, T., Dekker, S.C., Doubková, M., Friedl,
940 M. A., Jung, M., van den Hurk, B.J.J.M., de Jeu, R.A.M., Kruijt, B., Ohta, T., Rebel, K.T.,
941 Plummer, S., Seneviratne, S.I., Sitch, S., Teuling, A.J., van der Werf, G.R., Wang, G.,
942 2011. Drought and ecosystem carbon cycling. *Agr. Forest Meteorol.* 151, 765–773.
943 doi:10.1016/j.agrformet.2011.01.018

944 Vital, J.-A., Gaurut, M., Lardy, R., Viovy, N., Soussana, J.-F., Bellocchi, G., Martin, R.,
945 2013. High-performance computing for climate change impact studies with the Pasture
946 Simulation model. *Comput. Electron. Agr.* 98, 131-135.
947 doi:10.1016/j.compag.2013.08.004

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

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

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

958 Wohlfahrt, G., Anderson-Dunn, M., Bahn, M., Balzarolo, M., Berninger, F., Campbell, C.,
959 Carrara, A., Cescatti, A., Christensen, T., Dore, S., Eugster, W., Friborg, T., Furger, M.,
960 Gianelle, D., Gimeno, C., Hargreaves, K., Hari, P., Haslwanter, A., Johansson, T.,
961 Marcolla, B., Milford, C., Nagy, Z., Nemitz, E., Rogiers, N., Sanz, M.J., Siegwolf, R.T.W.,
962 Susiluoto, S., Sutton, M., Tuba, Z., Ugolini, F., Valentini, R., Zorer, R., Cernusca, A.,
963 2008. Biotic, abiotic, and management controls on the net ecosystem CO₂ exchange of
964 European mountain grassland ecosystems. *Ecosystems* 11, 1338-1351.
965 doi:10.1007/s10021-008-9196-2

966 Wösten J.H.M., Lilly, A., Nemes, A., Le Bas, C., 1999. Development and use of a database of
967 hydraulic properties of European soils. *Geoderma* 90,169-185. doi:10.1016/S0016-
968 7061(98)00132-3

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

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

974

975 **APPENDICES**

976

977 **Appendix 1**

978

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

983

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	<i>U1</i>	9.74	8.95	8.71	7.69	-1.02	-1.26	32.25	28.77	-0.06	-0.15	0.77	0.83
	<i>U2</i>	10.17	8.54	9.69	7.63	-0.47	-0.90	14.63	25.80	0.07	0.36	0.95	0.93
	<i>C</i>	9.74	8.95	8.60	7.71	-1.14	-1.24	34.81	37.02	0.63	0.59	0.89	0.90
	<i>V</i>	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	<i>U1</i>	9.74	8.95	5.01	7.36	-4.73	-1.59	54.21	28.66	-1.37	-0.65	0.90	0.89
	<i>U2</i>	10.17	8.54	6.91	6.92	-3.26	-1.62	39.92	27.81	-0.79	-0.19	0.92	0.94
	<i>C</i>	9.74	8.95	4.85	7.17	-4.89	-1.78	55.09	28.16	-1.36	-0.59	0.90	0.90
	<i>V</i>	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	<i>U1</i>	9.74	8.95	10.38	10.26	0.64	1.31	50.53	50.83	0.88	0.79	0.70	0.78
	<i>U2</i>	10.17	8.54	10.44	10.26	0.27	1.72	44.54	56.98	0.88	0.81	0.73	0.78
	<i>C</i>	9.74	8.95	7.80	7.65	-1.94	-1.30	29.93	25.49	-0.17	-0.07	0.86	0.88
	<i>V</i>	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	<i>U1</i>	9.74	8.95	10.04	8.70	0.31	-0.25	36.77	28.93	-1.10	-0.88	0.91	0.90
	<i>U2</i>	10.17	8.54	11.94	8.37	1.77	-0.16	35.82	23.48	-0.98	-0.37	0.91	0.94
	<i>C</i>	9.74	8.95	10.01	8.36	0.27	-0.59	35.59	26.59	-1.05	-0.81	0.91	0.91
	<i>V</i>	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	<i>U1</i>	9.74	8.95	7.80	NA	-1.94	NA	27.08	NA	-0.02	NA	0.89	NA
	<i>U2</i>	10.17	8.54	9.14	NA	-1.03	NA	17.06	NA	0.25	NA	0.97	NA
	<i>C</i>	9.74	8.95	7.84	NA	-1.89	NA	27.38	NA	-0.29	NA	0.90	NA
	<i>V</i>	10.17	8.54	9.31	NA	-0.86	NA	16.08	NA	0.02	NA	0.95	NA
M6	<i>U1</i>	9.74	8.95	6.95	7.21	-2.79	-1.74	31.92	24.50	-0.15	0.04	0.91	0.93
	<i>U2</i>	10.17	8.54	8.81	6.80	-1.36	-1.74	18.99	31.93	0.24	0.34	0.97	0.93
	<i>C</i>	9.74	8.95	11.45	7.20	1.72	-1.75	40.15	33.64	0.44	0.38	0.73	0.88
	<i>V</i>	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	<i>U1</i>	9.74	8.95	8.23	NA	-1.51	NA	25.37	NA	-0.33	NA	0.90	NA
	<i>U2</i>	10.17	8.54	9.72	NA	-0.45	NA	12.99	NA	-0.02	NA	0.96	NA
	<i>C</i>	9.74	8.95	7.86	NA	-1.88	NA	27.29	NA	-0.36	NA	0.90	NA
	<i>V</i>	10.17	8.54	9.36	NA	-0.81	NA	14.56	NA	-0.03	NA	0.96	NA
M8	<i>U1</i>	9.74	8.95	28.06	28.04	18.32	19.09	198.41	223.42	-7.42	-10.57	0.80	0.86
	<i>U2</i>	10.17	8.54	28.21	27.99	18.05	19.45	186.53	238.72	-7.48	-8.24	0.95	0.89
	<i>C</i>	9.74	8.95	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	<i>V</i>	10.17	8.54	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
MMM	<i>U1</i>	9.74	8.95	8.14	8.39	-1.60	-0.56	24.63	20.03	-0.12	0.00	0.90	0.92
	<i>U2</i>	10.17	8.54	9.66	8.03	-0.51	-0.50	12.12	18.58	0.17	0.31	0.97	0.97
	<i>C</i>	9.74	8.95	7.90	7.44	-1.83	-1.51	26.59	22.54	-0.26	0.02	0.90	0.93
	<i>V</i>	10.17	8.54	9.31	6.91	-0.86	-1.63	14.34	28.75	0.07	0.31	0.96	0.95

984

985

986 **Appendix 2**

987

988 Individual (M1-M9) and multi-model ensemble (MMM) model performance at different
 989 information (SIM) levels - uncalibrated (*U1*, *U2*), calibrated (*C*) and validated (*V*) - at the
 990 most humid and the most arid flux sites (ID as in Table 1) based on different metrics
 991 calculated on weekly averaged soil water content (SWC). NA: not available SWC
 992 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	<i>U1</i> *	0.45	0.36	0.37	0.36	-0.08	0.01	14.17	11.20	-714.6	0.30	0.10	0.50
	<i>U2</i>	0.41	0.33	0.36	0.36	-0.06	0.04	18.01	15.98	0.32	-1.91	0.83	0.25
	<i>C</i> *	0.45	0.36	0.39	0.39	-0.06	0.03	13.61	15.43	-329	0.34	0.08	0.46
	<i>V</i>	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	<i>U1</i> *	0.45	0.36	0.39	0.38	-0.06	0.02	16.35	14.38	-406.8	0.00	0.32	0.41
	<i>U2</i>	0.41	0.33	0.37	0.39	-0.04	0.06	14.94	21.67	0.42	-3.65	0.82	0.20
	<i>C</i> *	0.45	0.36	0.39	0.37	-0.06	0.02	16.51	14.21	-409.9	-0.05	0.45	0.40
	<i>V</i>	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	<i>U1</i> *	0.45	0.36	0.24	0.26	-0.21	-0.10	44.31	31.65	-4291	-3.68	0.34	0.18
	<i>U2</i>	0.41	0.33	0.22	0.26	-0.19	-0.06	47.51	24.08	-3.87	-4.92	0.70	0.07
	<i>C</i> *	0.45	0.36	0.30	0.35	-0.15	-0.01	33.46	19.79	-2219	-0.64	0.55	0.12
	<i>V</i>	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	<i>U1</i> *	0.45	0.36	0.23	0.38	-0.22	0.02	50.95	14.41	-4336	0.21	0.09	0.31
	<i>U2</i>	0.41	0.33	0.23	0.38	-0.19	0.05	48.60	19.64	-3.44	-3.06	0.56	0.23
	<i>C</i> *	0.45	0.36	0.34	0.36	-0.11	0.00	25.13	11.73	-1011	0.56	0.20	0.44
	<i>V</i>	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	<i>U1</i> *	0.45	0.36	0.31	NA	-0.14	NA	37.84	NA	-1934	1.00	0.00	NA
	<i>U2</i>	0.41	0.33	0.31	NA	-0.11	NA	33.85	NA	-0.52	1.00	0.02	NA
	<i>C</i> *	0.45	0.36	0.29	NA	-0.16	NA	37.95	NA	-2368	1.00	0.38	NA
	<i>V</i>	0.41	0.33	0.29	NA	-0.13	NA	34.63	NA	-1.29	1.00	0.55	NA
M6	<i>U1</i> *	0.45	0.36	0.31	0.29	-0.14	-0.06	42.54	24.92	-2066	-1.94	0.00	0.10
	<i>U2</i>	0.41	0.33	0.31	0.30	-0.10	-0.03	30.46	19.01	-0.75	-3.42	0.38	0.18
	<i>C</i> *	0.45	0.36	0.45	0.33	0.00	-0.03	3.43	12.90	0.74	-0.40	0.26	0.48
	<i>V</i>	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	<i>U1</i> *	0.45	0.36	0.70	NA	0.25	NA	64.14	NA	-5817	1.00	0.45	NA
	<i>U2</i>	0.41	0.33	0.69	NA	0.28	NA	68.30	NA	-8.97	1.00	0.49	NA
	<i>C</i> *	0.45	0.36	0.35	NA	-0.10	NA	17.90	NA	-1038	1.00	0.44	NA
	<i>V</i>	0.41	0.33	0.34	NA	-0.07	NA	24.44	NA	0.11	1.00	0.44	NA
M8	<i>U1</i> *	0.45	0.36	0.19	0.22	-0.26	-0.13	66.39	62.01	-8509	-24.12	0.70	0.18
	<i>U2</i>	0.41	0.33	0.14	0.19	-0.27	-0.13	72.21	58.86	-10.05	-36.17	0.14	0.01
	<i>C</i> *	0.45	0.36	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	<i>V</i>	0.41	0.33	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
M9	<i>U1</i> *	0.45	0.36	0.29	0.31	-0.16	-0.05	46.04	25.15	-2627	-2.09	0.51	0.03
	<i>U2</i>	0.41	0.33	0.29	0.32	-0.12	-0.01	40.18	21.67	-1.60	-4.25	0.01	0.06
	<i>C</i> *	0.45	0.36	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	<i>V</i>	0.41	0.33	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
MMM	<i>U1</i> *	0.45	0.36	0.31	0.33	-0.14	-0.02	40.18	16.11	-1945.6	-0.44	0.01	0.22
	<i>U2</i>	0.41	0.33	0.30	0.33	-0.11	0.01	32.71	13.39	-0.72	-1.04	0.23	0.20
	<i>C</i> *	0.45	0.36	0.35	0.36	-0.10	0.01	17.90	11.28	-975.49	0.43	0.44	0.55
	<i>V</i>	0.41	0.33	0.35	0.37	-0.07	0.05	22.91	17.39	0.30	-1.79	0.74	0.20

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

994

995

996 **Appendix 3**

997

998 Individual (M1-M9) and multi-model ensemble (MMM) model performance at different
 999 information (SIM) levels - uncalibrated (*U*) and calibrated (*C*) - for SAS and KEM1 sites (ID
 1000 as in Table 1) based on different metrics calculated on cutting events of yield biomass
 1001 (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	<i>U</i>	117.6	126.6	64.5	240.0	-53.1	113.4	89.4	132.3	-0.26	-22.99	0.15	0.09
	<i>C</i>			26.9	113.1	-90.7	-13.4	102.5	56.6	-0.46	-2.63	0.14	0.02
M2	<i>U</i>	117.6	126.6	11.1	93.2	-106.6	-33.4	111.4	46.8	-0.67	-1.18	0.22	0.02
	<i>C</i>			5.2	57.5	-112.5	-69.0	118.0	65.0	-0.81	-3.78	0.08	0.02
M3	<i>U</i>	117.6	126.6	62.6	36.1	-55.0	-90.4	129.8	80.3	-0.93	-6.20	0.02	0.01
	<i>C</i>			10.7	23.2	-107.0	-103.3	113.5	86.1	-0.62	-7.71	0.32	0.04
M4	<i>U</i>	117.6	126.6	34.8	124.9	-82.8	-1.7	97.8	25.7	0.02	0.84	0.21	0.14
	<i>C</i>			NA	184.0	NA	57.5	NA	54.5	NA	-2.39	NA	0.10
M5	<i>U</i>	117.6	126.6	85.6	38.4	-32.0	-88.1	72.5	79.4	0.00	-6.32	0.28	0.00
	<i>C</i>			85.6	101.8	-32.0	-24.8	72.5	67.7	0.00	-3.46	0.28	0.02
M6	<i>U</i>	117.6	126.6	190.3	335.8	72.6	209.3	139.8	181.5	-3.98	-42.07	0.28	0.05
	<i>C</i>			110.7	183.3	-6.9	56.7	73.9	62.0	0.68	-3.77	0.05	0.07
M7	<i>U</i>	117.6	126.6	99.7	166.5	-17.9	39.9	92.9	60.9	-0.87	-5.05	0.19	0.26
	<i>C</i>			65.9	155.6	-51.7	29.1	76.0	52.5	0.07	-4.13	0.29	0.37
M8	<i>U</i>	117.6	126.6	97.2	466.3	-20.4	339.7	88.4	294.5	0.44	-111.08	0.00	0.00
	<i>C</i>			NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
M9	<i>U</i>	117.6	126.6	107.0	179.9	-10.6	53.4	91.3	107.5	0.08	-13.92	0.03	0.02
	<i>C</i>			NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
MMM	<i>U</i>	117.6	126.6	61.2	153.5	-56.5	26.9	81.5	31.8	0.17	-1.48	0.19	0.62
	<i>C</i>			38.7	106.5	-78.9	-20.0	87.6	32.7	-0.12	-0.40	0.40	0.24

1002

1003 **Appendix 4**

1004 Average ratio of the ensemble spread to model error: average absolute standardized spread
 1005 (maximum-minimum) of model results / average absolute standardized model residual.
 1006 Responses are from calibrated simulations of soil temperature (ST), soil water content (SWC)
 1007 and yield biomass, as obtained at each site (ID as in Table 1) with both individual models
 1008 (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	

1009