Contents lists available at SciVerse ScienceDirect

Field Crops Research

journal homepage: www.elsevier.com/locate/fcr

Simulation of spring barley yield in different climatic zones of Northern and Central Europe: A comparison of nine crop models

Reimund P. Rötter[a,*], Taru Palosuo[a], Kurt Christian Kersebaum[b], Carlos Angulo[c], Marco Bindi[d], Frank Ewert[c], Roberto Ferrise[d], Petr Hlavinka[e], Marco Moriondo[f], Claas Nendel[b], Jørgen E. Olesen[g], Ravi H. Patil[g,h], Françoise Ruget[i], Jozef Takáč[j], Miroslav Trnka[e,k]

[a] *MTT Agrifood Research Finland, Lönnrotinkatu 5, 50100 Mikkeli, Finland*
[b] *ZALF, Leibniz-Centre for Agricultural Landscape Research, Eberswalder Str. 84, D-15374 Müncheberg, Germany*
[c] *University of Bonn, Institute of Crop Science and Resource Conservation, Katzenburgweg 5, D-53115 Bonn, Germany*
[d] *University of Florence, DIPSA, Department of Plant, Soil and Environmental Science, Piazzale delle Cascine 18, 50144 Florence, Italy*
[e] *Institute of Agrosystems and Bioclimatology, Mendel University in Brno, Zemedelska 1, Brno 613 00, Czech Republic*
[f] *National Research Council of Italy, IBIMET-CNR, Institute of Biometeorology, via Caproni 8, 50145 Florence, Italy*
[g] *Department of Agroecology and Environment, Aarhus University, DK-8830 Tjele, Denmark*
[h] *Agricultural and Biological Engineering Department, University of Florida, Gainesville, 32611-0570 FL, USA*
[i] *INRA, UMR 1114 EMMAH Environement et Agronomie, F-84000 Avignon, France*
[j] *Soil Science and Conservation Research Institute, Gagarinova 10, 827 13 Bratislava, Slovak Republic*
[k] *Global Change Research Centre AS CR, v.v.i., Bělidla 986/4a, 603 00 Brno, Czech Republic*

ARTICLE INFO

Article history:
Received 21 December 2011
Received in revised form 30 March 2012
Accepted 30 March 2012

Keywords:
Climate
Crop growth simulation
Model comparison
Spring barley
Yield variability
Uncertainty

ABSTRACT

In this study, the performance of nine widely used and accessible crop growth simulation models (APES-ACE, CROPSYST, DAISY, DSSAT-CERES, FASSET, HERMES, MONICA, STICS and WOFOST) was compared during 44 growing seasons of spring barley (*Hordeum vulgare* L.) at seven sites in Northern and Central Europe. The aims of this model comparison were to examine how different process-based crop models perform at multiple sites across Europe when applied with minimal information for model calibration of spring barley at field scale, whether individual models perform better than the multi-model mean, and what the uncertainty ranges are in simulated grain yields. The reasons for differences among the models and how results for barley compare to winter wheat are discussed.

Regarding yield estimation, best performing based on the root mean square error (RMSE) were models HERMES, MONICA and WOFOST with lowest values of 1124, 1282 and 1325 (kg ha^{-1}), respectively. Applying the index of agreement (IA), models WOFOST, DAISY and HERMES scored best having highest values (0.632, 0.631 and 0.585, respectively). Most models systematically underestimated yields, whereby CROPSYST showed the highest deviation as indicated by the mean bias error (MBE) (−1159 kg ha^{-1}). While the wide range of simulated yields across all sites and years shows the high uncertainties in model estimates with only restricted calibration, mean predictions from the nine models agreed well with observations. Results of this paper also show that models that were more accurate in predicting phenology were not necessarily the ones better estimating grain yields. Total above-ground biomass estimates often did not follow the patterns of grain yield estimates and, thus, harvest indices were also different. Estimates of soil moisture dynamics varied greatly.

In comparison, even though the growing cycle for winter wheat is several months longer than for spring barley, using RMSE and IA as indicators, models performed slightly, but not significantly, better in predicting wheat yields. Errors in reproducing crop phenology were similar, which in conjunction with the shorter growth cycle of barley has higher effects on accuracy in yield prediction.

1. Introduction

Various model-based tools are used to support the decision making and planning in agriculture (Brouwer and van Ittersum, 2010; Ewert et al., 2011). Crop growth simulation models (hereafter referred to as crop models) are increasingly being applied, particularly in climate change-related agricultural

* Corresponding author. Tel.: +358 40 353 4506.
E-mail address: reimund.rotter@mtt.fi (R.P. Rötter).

0378-4290/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.fcr.2012.03.016

impact assessments (Rosenzweig and Wilbanks, 2010; White et al., 2011).

Recently, there has been renewed interest and discussion about the need for improved understanding and reporting of the uncertainties related to crop growth and yield predictions (Rötter et al., 2011a; Ferrise et al., 2011; Børgesen and Olesen, 2011). Comparison of different modelling approaches and models can reveal the uncertainties involved. Variation of model results in model comparisons involves also the uncertainty related to model structure, which is probably the source of uncertainty most difficult to quantify. Model comparisons, when combined with experimental data of the compared variables, may also be used to test the performance of different models. However, comprehensive data sets that would allow such thorough comparisons (see, e.g. Groot and Verberne, 1991 or Kleemola et al., 1995), are scarce and in most cases have already been utilized or published for model calibration or validation. This situation calls for a concerted effort to exploit existing (unused) and develop new high quality data sets for different locations (agro-climatic conditions) and crops (Rötter et al., 2011a). Since the 1980s, there have been many studies on comparing different process-based crop models on their performance in predicting yield variability in response to climate and other factors (see, e.g. Kersebaum et al., 2007; Palosuo et al., 2011), including a very active period during the 1990s (Porter et al., 1993; Diekkrüger et al., 1995; Ewert et al., 2002; Goudriaan et al., 1994; Jamieson et al., 1998; Kabat et al., 1995; Wolf et al., 1996). Most of these comparisons have been made for wheat while other crops such as barley, received much less attention (Tubiello and Ewert, 2002; see, e.g. Eitzinger et al., 2004 for an exception).

Since proper understanding and modelling of crop responses to heat and drought stress becomes increasingly important in climate impact assessments (Semenov and Shewry, 2011; Lobell et al., 2012), we also looked into this issue. In a couple of studies in different parts of the world specific responses of barley to heat and drought stress have been investigated (e.g. Jamieson et al., 1995; Passarella et al., 2005). For the critical growth stages during and immediately after flowering (Savin and Nicolas, 1999), it has been found that significant yield reduction is experienced if threshold temperatures of 28–30 °C are exceeded. Yield-reducing effects depend, however, on the timing and intensity of events (Passarella et al., 2005). Moreover, there is considerable response diversity among barley cultivars (see, e.g. Hakala et al., 2012). For drought stress, Jamieson et al. (1995) found no clear thresholds, but rather the importance of timing of drought for reduction in final biomass of barley, whereby final biomass was especially sensitive to soil moisture deficit for the early drought treatments.

To analyze sources of crop model uncertainties in climate impact assessments for Europe, four crop model intercomparisons were set-up during 2009–2010 in the framework of COST action 734, seeking coverage of the most widely used and accessible crop simulation models: one comparison for winter wheat (Palosuo et al., 2011) and another one for spring barley (this study) across multiple sites in Europe with restricted calibration, one on the sensitivity of crop models to extreme weather conditions for maize and winter wheat (Eitzinger et al., in press), and one with a detailed calibration using comprehensive barley datasets from one Finnish location (Salo et al., companion paper, in preparation).

This paper presents the results of the spring barley (*Hordeum vulgare* L.) comparison across multiple sites in Europe. Barley is currently the third most important cereal in Northern and Central Europe after wheat and grain maize (EUROSTAT, 2011). Since spring barley has been much less considered in crop modelling than winter wheat, and assuming that accordingly wheat models were developed with more experimental data than those for barley, we hypothesized that the uncertainties in simulation results for barley are higher.

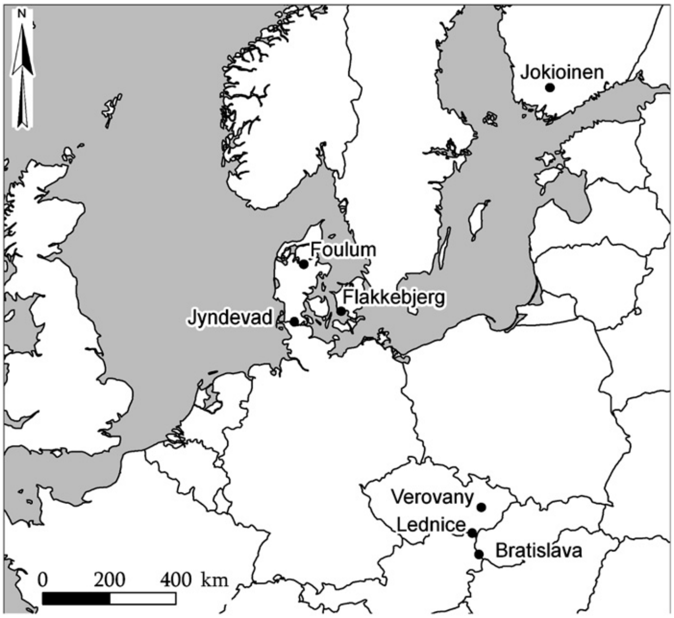

Fig. 1. Locations of the study sites.

The specific objectives of this model intercomparison study were to examine (1) how different process-based crop models perform at multiple sites across Europe in estimating grain yield when applied with minimal information for model calibration of spring barley at field scale, (2) whether individual models perform better than the multi-model mean, and (3) what the uncertainty ranges are in simulated grain yields. Furthermore, an initial effort is made to discuss the reasons for differences among the models and investigate how results for barley compare to winter wheat (Palosuo et al., 2011).

We applied nine crop models altogether for 44 growing seasons of spring barley at seven different study sites in Europe: in the Czech Republic, Denmark, Finland and Slovakia.

2. Material and methods

2.1. Models

Nine crop simulation models, APES-ACE, CROPSYST, DAISY, DSSAT-CERES, FASSET, HERMES, MONICA, STICS and WOFOST were applied at seven different study sites in Northern and Central Europe (Fig. 1). Details about these models can be obtained from the main references gathered by Palosuo et al. (2011), except for model MONICA, which has been described by Nendel et al. (2011). Table 1 gives an overview of the model version applied, model calibrations and their major applications for barley in Europe, while Table 2 provides an overview with characterization of basic process descriptions and how the models deal with heat and drought stress.

All models are summary models and work on a daily time step. The models differ considerably in the way they treat growth-defining, -limiting and -reducing factors (van Ittersum et al., 2003) and, correspondingly, in their structure, associated input data requirements and model parameters. One can group the models according to different criteria, such as the approach used for describing daily biomass or dry matter accumulation under non-limiting conditions (e.g. Confalioneri et al., 2009 distinguished three groups: SUCROS/WOFOST type; CERES type and a simpler one; in Table 2 we distinguished two types for light utilization or biomass growth – see, also Adam et al., 2011); but our models also

Table 1
Model version applied in this study, references to papers with data for model parameterization, applications, and model web address.

Model Version	Reference to relevant earlier model parameterizations (other evaluations/applications for barley in Europe)	Documentation/accessibility (weblink)
APES-ACE V. 1.0	Ewert et al. (2011)	Request from frank.ewert@uni-bonn.de
CROPSYST V. 3.04.08	Unpublished calibration for Poland (1995–2005) Donatelli et al. (1997)	http://www.bsyse.wsu.edu/CS_Suite/CropSyst/index.html
DAISY V. 4.01	Hansen et al. (1990) Svendsen et al. (1995), Smith et al. (1997), Refsgaard et al. (1998)	http://code.google.com/p/daisy-model/
DSSAT-CERES V. 4.0.1.0	Hlavinka et al. (2010) Eitzinger et al. (2004) and Trnka et al. (2004)	http://www.icasa.net/dssat/
FASSET V. 2.0	Olesen et al. (2000) Berntsen et al. (2004) and Doltra et al. (2011) and Sapkota et al. (2012)	http://www.fasset.dk
HERMES V. 4.26	Franko et al. (2007) Kersebaum et al. (2007)	http://www.zalf.de/en/forschung/institute/lsa/forschung/oekomod/hermes
MONICA V. 1.0	Nendel et al. (2011)	http://monica.agrosystem-models.com
STICS V. 6.9	Corre-Hellou et al. (2009) and Launay et al. (2009)	http://www.avignon.inra.fr/agroclim_stics_eng/
WOFOST V. 7.1	Boons-Prins et al. (1993) and Rötter et al. (2011b) Eitzinger et al. (2004)	http://www.wofost.wur.nl

operate under growth-limiting conditions (water and/or nutrient limitation), and, hence, can be classified according to other criteria, e.g. how they treat the soil moisture balance (see, Table 2 or van Ittersum et al., 2003), which makes it more difficult to rank models according to their complexity. Based on the characteristics provided in Table 2 and in related literature (Bouman et al., 1996; Brisson et al., 2003; Jones et al., 2003; Stöckle et al., 2003; van Ittersum et al., 2003), we can, however, roughly classify them: DAISY as the most complex, is followed by a group containing MONICA, HERMES and STICS. Then come the less complex WOFOST, FASSET, DSSAT-CERES and, finally, APES-ACE and CROPSYST. Much more complex models than DAISY are usually not applied in regional climate impact assessments due to their much higher data requirements.

2.2. Study sites

The model comparison was carried out using data from seven research sites in North and central Europe, Denmark, Czech Republic, Finland and Slovakia (Fig. 1). The principal characteristics of these sites are summarized in Table 3. Data contained altogether 44 growing seasons of spring barley. The longest time series, 14 and 13 years, were available for the two Czech sites at Verovany and Lednice, respectively. For the rest of the sites data from three to four years were available. Soils varied widely in their soil moisture retention characteristics, ranging from less favorable sandy soils (Jyndevad, Denmark) to favorable silt loams (at the Czech and Slovakian sites) (Table 3). Irrigation was applied at the Danish sites of Jyndevad (2006: 153 mm, 2007: 68 mm, 2008: 178 mm) and Foulum (2006: 103 mm, 2007: 53 mm, 2008: 94 mm).

In all experiments, the plots were kept weed free and plant protection was applied as necessary to avoid the presence of pests and diseases. Years during which the yields were reported to be affected by pests or diseases in spite of these plant protection activities were excluded from the study.

For Bratislava site, we selected three years (1996, 1999 and 2002) to illustrate how high temperature stress events between flowering and maturity (with $T_{max} > 30\,^{\circ}C$) combined with three distinctly different seasonal soil moisture patterns influenced simulated and observed total above-ground biomass production (TAGB) and simulated actual evapotranspiration. The soil moisture patterns were characterized as: (i) moderate early drought with

Table 2
Modelling approaches applied in this study regarding the major processes determining crop growth and development.

Model	LA development and LI[a]	Light utilization[b]	Yield formation[c]	Root distribution over depth[d]	Heat stress around flowering[e]	Drought stress[f]	Water dynamics[g]	Evapo-transpiration[h]
APES	D	RUE	Y(Prt)	Exp	No	ET_a/ET_p	C	P
CROPSYST	S	RUE	Y(HI,B)	Lin	No	T_a/T_p	C	PT
DAISY	D	P-R	Y(PRT)	Exp	No	cl-SM	R	PM
DSSAT	S	RUE	Y(HI(Gn),B)	Exp	No[i]	T_a/T_p	C	PT
FASSET	D	RUE	Y(HI,B)	Exp	No	ET_a/ET_p	C	MA
HERMES	D	P-R	Y(Prt)	Exp	No	T_a/T_p	C	PM
MONICA	D	P-R	Y(Prt)	Exp	Yes	ET_a/ET_p	C	PM
STICS	D	RUE	Y(HI(Gn),B)	Sig	No	cl SM	C	SW
WOFOST	D	P-R	Y(Prt,B)	Lin	No	cl-SM	C	P

[a] Leaf area development and light interception; simple (=S) or detailed (=D) approach.

[b] Light utilization or biomass growth: RUE = simple (descriptive) Radiation use efficiency approach, P-R = detailed (explanatory) Gross photosynthesis–respiration; (for more details, see e.g. Adam et al. (2011)).

[c] *Y*(*x*) yield formation depending on: HI = fixed harvest index, B = total (above-ground) biomass, Gn = number of grains, Prt = partitioning during reproductive stages.

[d] Root distribution over depth: linear (Lin), exponential (Exp), sigmoidal (Sig).

[e] Heat stress around flowering described in the model: Yes/No.

[f] Drought stress: T_a/T_p or ET_a/ET_p, or crop/crop group specific cl-SM = critical limits for plant available soil moisture in root zone.

[g] Water dynamics approach: C = capacity approach, R = Richards approach.

[h] Method to calculate evapo-transpiration: P = Penman; PM = Penman–Monteith, PT = Priestley–Taylor, TW = Turc–Wendling, MA = Makkink, HAR = Hargreaves, SW = Shuttleworth and Wallace (resistive model).

[i] No heat stress was reported by the model at any test site.

Table 3
Characteristics of the study sites.

Location *Environmental zone*[a]	Position Latitude/longitude/altitude a.s.l.	Precipitation[b] [mm yr^{-1}]	Temperature[c] [°C]	Period	Crop variety	Sand[d] [%]	Silt [%]	Clay [%]	C_{org} [%]	Root depth [cm]	fc–wp root zone[e] [mm]	Soil name
Lednice (CZ)	48°48′/16°48′/176 m	539	10.0	1984–1991	Orbit	T: 17	61	22	1.41	150	258	Chernozem
CON 2				1993–1996		S: 18	62	20				
				1998								
Verovany (CZ)	49°28′/17°17′/214 m	576	9.0	1984–1996	Orbit	T: 17	66	17	1.17	150	265	Chernozem
CON 10				1998		S: 15	64	21				
Bratislava (SK)	48°10′/17°/131 m	523	10.0	1994, 1996	Jubilant	T: 15	63	22	1.49	120	211	Chernozem
PAN 2				1999, 2002	Akcent	S: 16	64	20				
Foulum (DK)	56°30′/9°35′/54 m	694	8.8	2006–2008	Mixture[f]	T: 78	13	9	2.15	130	239	Mollic Luvisol
ATN 3						S: 75	12	13				
Flakkebjerg (DK)	55°11′/11°14′/32 m	607	9.6	2006–2008	Mixture[f]	T: 73	12	15	0.98	160	243	Glossic Phaeozem
CON 5						S: 69	12	19				
Jyndevad (DK)	54°54′/9°08′/14 m	864	9.6	2006–2008	Mixture[f]	T: 92	4	4	1.13	60	59	Humic Podzol
ATN 3						S: 93	3	4	0.6			
Jokioinen (FI)	60°42′/23°30′/104 m	506	4.3	2005–2008	Scarlett	T: 39	15	46	3.4	70	171	Dystric Cambisol
BOR 5						S: 24	24	52	3.2			

[a] Environmental zone according to Metzger et al. (2005), as quoted by Trnka et al. (2011).
[b] Average annual precipitation for period of simulations.
[c] Average annual temperature for period of simulations.
[d] Texture is given as average for T = topsoil (ploughing zone) and S = subsoil (until given root depth).
[e] mm water at field capacity (fc) and wilting point (wp) in specific root zone.
[f] The mixture of cultivars Simba, Smilla and Cicero.

Table 4
Input data provided to models.

Category	Variable	Type
Meteorological data	Minimum temperature	Daily minimum [°C]
	Maximum temperature	Daily maximum [°C]
	Relative air humidity	Daily average [%]
	Global radiation	Daily sum [MJ m^{-2}]
	Wind speed	Daily average [m s^{-1}]
	Precipitation	Daily sum [mm]
Soil data (0 cm to maximum rooting depth)	Texture	Per layer clay, silt, sand [mass%]
	C_{org}	Per layer [mass%]
	C:N ratio	Per layer [unitless]
	Bulk density	Per layer [cm^3 cm^{-3}]
	pH	Per layer [unitless]
	Field capacity	Per layer [cm^3 cm^{-3}]
	Wilting point	Per layer [cm^3 cm^{-3}]
	Total pore space	Per layer [cm^3 cm^{-3}]
	Max. rooting depth	[cm]
Crop data	Cultivar	
	Crop density	Crops per m^2
	Flowering (or heading)	doy (=day of year)
	Yellow ripeness	doy
Initial status	Water content[a]	Per layer [cm^3 cm^{-3}]
	Soil mineral N[a]	Per layer [kg ha^{-1}]
Management	Sowing date	doy
	Harvest date	doy
	N fertilization	doy, fert. type, amount [kg N ha^{-1}]
	Irrigation	doy, amount [mm]
	Tillage	doy, type, depth [cm]
	Previous crop sowing/harvest	doy
	Previous crop yield/residues	Res. export (y/n), amount [t ha^{-1}]

[a]Estimated for all but Bratislava site.

pronounced depletion around flowering (year 1996); (ii) favorable conditions with only short and moderate dry spell (year 1999), and (iii) steady depletion with terminal drought (year 2002).

2.3. Setup of model intercomparison

2.3.1. Information available for model users

The current study was implemented as a "blind test", i.e. the model users were not provided with the information on the variables they were asked to deliver as model results before they submitted the results. For the simulations, the input data provided are listed in Table 4. These also included information on key phenological dates during the growing period for each of the various spring barley cultivars used in different sites and seasons.

2.3.2. Calibration of the models

Phenological data were the only ones used in calibrating the models for the various barley cultivars grown at the different sites. It was further agreed that only one crop phenology parameter set is derived per cultivar to best match the phenological dates observed in the experiments. That set was then applied to all seasons in which the specific barley cultivar was grown. However, we did not exactly specify a procedure how model users should interpret and convert this information into parameter values. We further only recommended that all other crop parameters needed for the models were taken from earlier applications of the models thought to be eco-physiologically relevant (see, references in Table 1). But here again, the definition of "relevance" was left to the individual modeller. These parameters were then kept unchanged for all years and locations. While daily weather variables, basic soil physical characteristics and estimates of soil moisture and mineral nitrogen at start of the growing seasons were provided (Table 4), it was not prescribed exactly how that information should be used to generate initial soil moisture and soil nitrogen conditions.

2.4. Methods used for evaluating model performance and assessing uncertainties

The methods of how to assess and compare the performance of models have been discussed widely (see e.g. Bellocchi et al., 2009; Kobayashi and Salam, 2000; Wallach et al., 2006; Willmott, 1981). The combined use of various statistical indicators is seen important to achieve a balanced picture. Grain yields and growth duration (from emergence to flowering and maturity) simulated by the various models were compared with observed values.

For assessing and comparing model performance we calculated a set of statistical parameters in line with those reported by Palosuo et al. (2011): the root mean square error (RMSE) was taken as a measure of the relative average difference between the model estimates and measurements. CV(RMSE) is defined as RMSE normalized to the mean of the observed values:

$$\mathrm{CV(RMSE)} = \frac{\sqrt{N^{-1}\sum_{i=1}^{N}(P_i - O_i)^2}}{\bar{O}}, \quad (1)$$

where N is the number of estimate-observation-pair, P_i is the model prediction, O_i is the observed value and $\bar{O}$ is the mean of observations.

Mean bias error (MBE) was taken as an indicator telling whether the models under- or overestimate the yields, i.e. the direction and magnitude of bias:

$$\mathrm{MBE} = N^{-1}\sum_{i=1}^{N}(P_i - O_i) \quad (2)$$

The variance of the distribution of differences (s_d^2) was used to quantify the error variability:

$$s_d^2 = (N-1)^{-1}\sum_{i=1}^{N}(P_i - O_i - \mathrm{MBE})^2 \quad (3)$$

Overall systematic error relative to total mean squared error ($\mathrm{MSE_S}$/MSE) was used to identify how much or what proportion of RMSE is systematic in nature. It is calculated as a share between the systematic error and mean square error.

$$\mathrm{MSE_S} = N^{-1}\sum_{i=1}^{N}(\hat{P}_i - O_i)^2 \quad (4)$$

where $\hat{P}$ is derived from $\hat{P}_i = a + bO_i$.

Index of agreement (IA) developed by Willmott (1981) was used as a more general indicator of modeling efficiency.

$$\mathrm{IA} = 1 - \frac{N \cdot \mathrm{MSE}}{\mathrm{PE}} \quad (5)$$

where $\mathrm{PE} = \sum_{i=1}^{N}(|\dot{P}| + |\dot{O}|)^2$ and where $\dot{P} = P_i - \overline{P}$ and $\dot{O} = O_i - \overline{P}$, IA can have values within the range [0,1], and the values closer to 1 indicate the better simulation quality.

Above this, for comparison, the traditional r^2 regression statistic (least-squares coefficient of determination) was calculated even though it does not take into account model bias, which is central when assessing the performance of simulation models.

We provide an indication of uncertainties in model simulations attributable to using a variety of crop models (representing different complexity) and model user groups (representing different application skills) by showing outcomes from the nine individual models, and comparing these to observed mean yields. Uncertainty is represented by a distribution of simulated model results, whereas error is the difference between observed and predicted values, applying to cases where we have the true value (Wallach et al., 2006). Bias means an average (over sites or years, etc.) over- or under-estimates by the models (illustrated by the MBE). Here, we need to stress that the observed data presented in this study cannot unambiguously be considered as true values, but for the evaluation we use them as such.

3. Results

3.1. Assessment of model performance

3.1.1. Crop phenology

Calibration results for spring barley phenology show considerable discrepancies with observations, amounting to ±11 days for the start of flowering (Zadoks 61) and up to +12 days for physiological maturity (Zadoks 90). The most accurate estimates of phenology were provided by models STICS and WOFOST (Fig. 2a and b). The grain filling period was longest for FASSET and notably short for CROPSYST and HERMES (Fig. 2a).

3.1.2. Grain yield

A detailed comparison of the grain yield estimates with observed values showed that none of the models perfectly reproduced observations at all sites and in all years (Figs. 3–5a). However, some models (e.g. HERMES) clearly performed better than others. Two models (CROPSYST, DAISY) systematically underestimated yields, while one model (WOFOST) mostly overestimated yields (Figs. 3 and 4b). The statistical analysis of the grain yield results show that the best performance regarding yield estimation was found for HERMES, MONICA, WOFOST and DAISY, for which the RMSE values were lowest and the IA values highest (Fig. 4a and e). IA was lowest for FASSET and CROPSYST, and highest for WOFOST and DAISY (Fig. 4e).

The overall or average systematic error (MSEs/MSE) was lowest for HERMES (Fig. 4d), whereas CROPSYST had by far the highest systematic error (Fig. 4d). DSSAT-CERES and FASSET showed the highest variance of model residuals (Fig. 4c) indicating some high individual discrepancies between simulated and observed yields. Clearly, DAISY model showed the highest coefficient of determination (r^2 = 0.48) (Fig. 4f).

The ability of the models to capture the variability of grain yield at field-level was studied using the Verovany study site with the longest time series (14 years), with results similar to the Palosuo et al. (2011) study. Observed mean yields and its variability were best captured by model DSSAT-CERES, followed by MONICA and HERMES (not shown). This result was confirmed by statistical indicators RMSE (573, 719 and 897 kg ha^{-1}, respectively) and by IA (0.78, 0.59 and 0.51, respectively).

When comparing the performance of multi-model mean (hereafter referred to as MMM) and individual models in estimating grain yields, we found that the MMM is a better predictor over all sites and seasons (Fig. 3). When taking RMSE and IA as indicators, we find that several models like HERMES, MONICA, WOFOST and DAISY do not perform much worse than the MMM (Fig. 4a and e). How this comparison looks for individual sites and seasons is presented in Section 3.2.

For a short duration crop like barley inaccuracies of a few days in simulated phenology can have quite an effect on simulated dry matter increase and final grain yield. Hence, we also examined, whether accuracy in estimating phenology is correlated to the accuracy of models in estimating grain yield. Results of associated regression analyses (not shown), however, suggested that the correlation between the accuracies in phenology and yield estimation is weak.

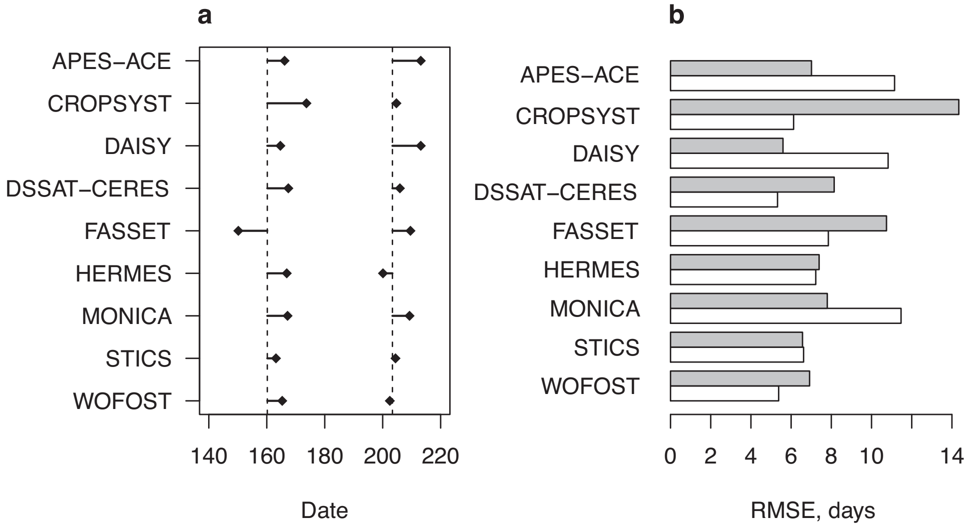

Fig. 2. Calibration results: model performance for phenology (a) mean model estimates for date of start of anthesis (Zadoks 61) (black tetragons around left column) and (b) RMSE of the model-calculated anthesis (grey bars) and maturity (Zadoks 90) (white).

3.1.3. *Root biomass, above-ground biomass and harvest index*

Maximum root biomass estimates were available for all models except CROPSYST. According to Fig. 5b, there are two groups of models performing quite differently in estimating root biomass. APES-ACE, DAISY, DSSAT-CERES and HERMES estimated average root biomass around 1000 kg ha^{-1} or less while FASSET, MONICA, STICS and WOFOST estimates were at 1750 kg ha^{-1} or higher. Observations of root biomass only were available from the Foulum site in 2008, amounting to 1730 kg ha^{-1} (Chirinda et al., in press).

In terms of simulated total above-ground biomass (TAGB) models followed a slightly different order than for simulating grain yields. DSSAT-CERES clearly showed the highest TAGB estimates, followed by DAISY, APES-ACE and WOFOST. FASSET and CROPSYST were the models with lowest estimates of TAGB (Fig. 5c).

Harvest indices (HI) varied more widely among models than grain yield or TAGB (Fig. 5d). HI estimates ranged from 0.4 to 0.6, which is plausible for spring cereals according to the literature (Peltonen-Sainio et al., 2008). DSSAT-CERES, DAISY and STICS were found at the low end, while HERMES and WOFOST were at the high end. DSSAT, MONICA and WOFOST showed highest variation in HI estimates across sites and seasons, whereas HI applied in CROPSYST were almost constant (at 0.48) (Fig. 5d).

3.1.4. *Dynamics of above-ground biomass and soil moisture*

Observed TAGB in years 1996 and 2002 at Bratislava site were almost the same (approx. 6200 kg ha^{-1}), but almost twice as high in 1999 (with approx. 11,900 kg ha^{-1}) (Fig. 6a–c). That was mainly due to a relatively good water availability in 1999 (Fig. 6g–i) compared to the other two years. This is also expressed by a generally higher cumulative actual evapotranspiration than simulated in 1996 and 2002 (Fig. 6d–f). Soil moisture stock was largely replenished during the period of peak water requirements, whereas in year 2002 there was a steady soil moisture depletion right from the start till values got close to wilting point at the late growth stages. Year 1996 was intermediate in terms of soil moisture conditions; however, the pattern was one of relatively high soil moisture contents till about day 150 (31st of May). Thereafter, we see a rapid decline, and at about the same time, a high number of (five) hot days – unlike in other years where only one such event was observed (Fig. 6d–f).

Simulated TAGB show large variations among models. Simulated soil moisture availability in 1996 was higher than in 2002 but lower than in 1999 (Fig. 6g–i). However, an early depletion of soil water during April (Fig. 6g) led to a reduction in simulated and observed biomass production (Fig. 6a) which is the lowest of all three years. This is also expressed by the distinctly lower LAI (not shown) simulated by nearly all models relative to the other seasons. In all three years one of the nine models (DSSAT-CERES) considerably overestimated TAGB, while another one (CROPSYST) mostly underestimated TAGB. In the favorable year 1999, apart from these two models, others estimated TAGB fairly accurately. In 1996, models showed higher discrepancies to observed biomass and among model estimates than in 2002 (with the exception of DSSAT-CERES). In 2002 all (except for one model) were overestimating biomass which can be attributed to the fact that they also overestimated soil moisture (Fig. 6i).

We calculated statistical performance indicators (MBE, RMSE, IA and ME) and regressed RMSE water on RMSE grain yield and biomass. There was a comparable positive correlation with r^2 values of 0.25 and 0.26, respectively. Model DAISY clearly showed best performance for estimating soil water (RMSE = 14.1 mm/90 cm profile; IA = 0.986), followed by HERMES and WOFOST with IAs of 0.93 and 0.864, respectively. DAISY also clearly performed best in estimating total above-ground biomass (RMSE = 2034 kg ha^{-1}). Furthermore, we found a strong correlation when regressing MBE water on MBE grain yield (r^2 = 0.58).

3.2. *Uncertainties*

A wide range of model estimates of grain yield (Figs. 3, 5a and 7) indicates the magnitude of uncertainty related to model estimates. This also applies to estimates of soil moisture contents, total above-ground biomass and other indicators (see Fig. 6).

There is a considerable spread of simulated yields among the models for most of the 44 growing seasons (Fig. 7). The highest ranges of model-based yield estimates (in the extreme case almost 5000 kg ha^{-1}) can be found at Lednice site. For other sites the range of model estimates is much lower (on average about 2500 kg ha^{-1}). There are six out of 44 studied seasons (14%) in which observed yields are not covered by the range of model estimates: at Lednice, observed yields exceed simulated yields in 1987–89 and in 1998; at Verovany, observed exceeds simulated yield in year 1994, while at Foulum in 2007, all simulated yields exceed the observed.

For the two Czech sites (27 growing seasons) the multi-model mean (MMM) underestimates observed yields with one exception, year 1993, at Lednice. On the contrary, at the other sites MMM overestimates observed yields in most of the 17 seasons.

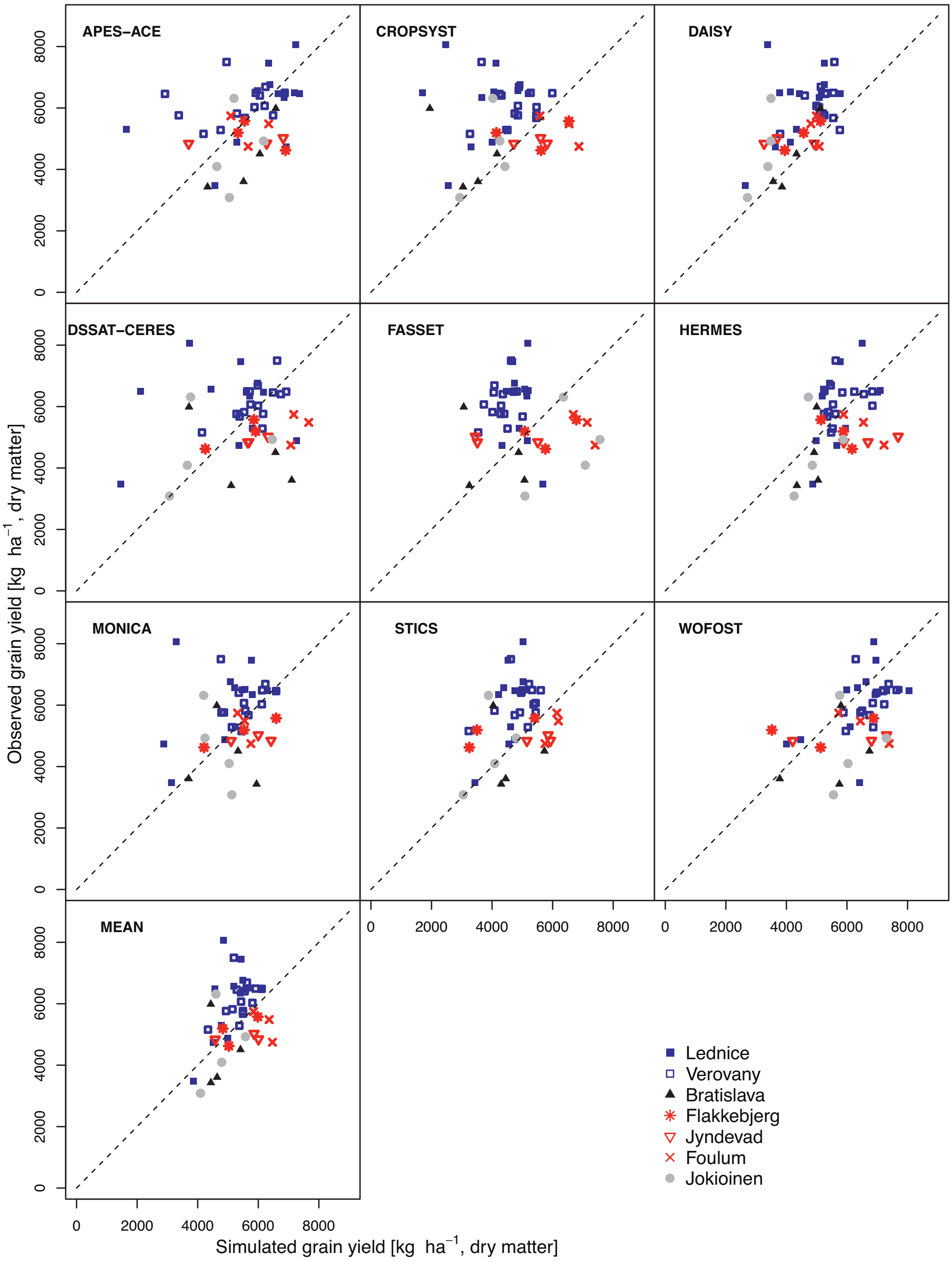

Fig. 3. Simulated and observed grain yield estimates [kg ha^{-1}, dry matter] for 44 studied growing seasons. Simulation results are shown for nine individual models and multi-model means. Different study sites are depicted with different symbols. The 1:1 line is shown, representing perfect agreement.

Exceptions include Bratislava in 1994, Flakkeberg in 2006 and Jokioinen in 2005.

At the two Czech sites, the "best model" HERMES estimates yields slightly better than the MMM (Fig. 7). Overall, however, the MMM is a slighly better predictor than HERMES as indicated by RMSE and IA (Fig. 4). Two other models, DAISY and WOFOST almost perform as well as the "best model". However, their "best performances" look quite different, as we found when plotting yield estimates by the individual models *vis a vis* observed (not shown) as in Fig. 7. Except for Bratislava site, DAISY tends to underestimate observed yields and remains below the MMM. This is most pronounced for the Czech and Finnish sites. WOFOST, on the other hand, in most cases overestimates observed yields, on average by about 1000 kg ha^{-1}.

For all growing seasons, and for Verovany site separately, we also calculated Spearman's rank correlations (not shown) to examine how well the models are in reproducing the order of observed yields. Models DAISY and WOFOST, with Spearman's

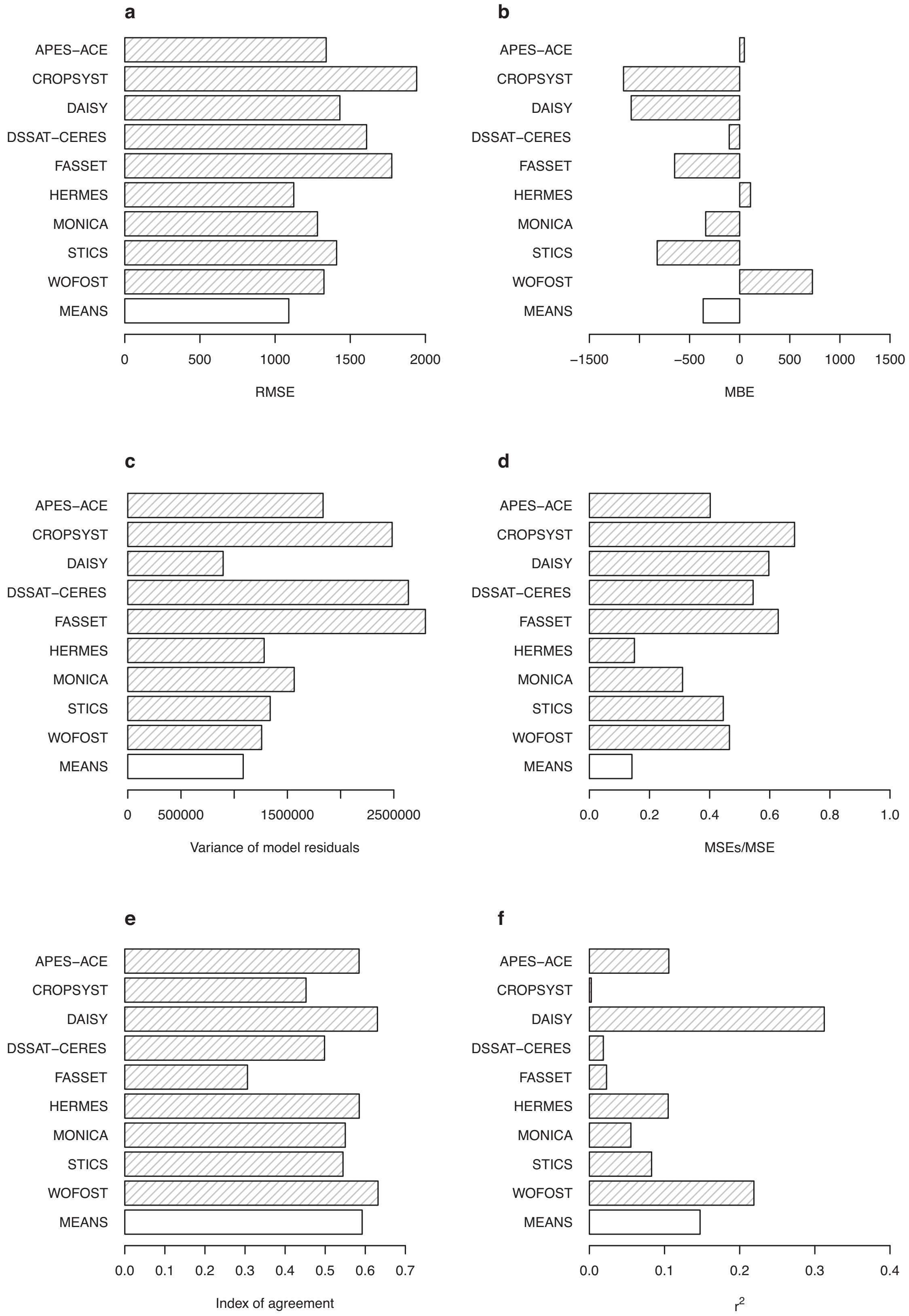

Fig. 4. Graphical representation of statistics describing the performance of individual models in simulating all study sites and growing seasons ($N = 44$) as compared to multi-model means; (a) normalized root mean square error CV(RMSE) [0,1], (b) mean bias error (MBE), (c) variance of model residuals, (d) systematic error (MSE_S/MSE) [0,1], (e) index of agreement (IA) [0,1], and (f) least-squares coefficient of determination (r^2) [0,1].

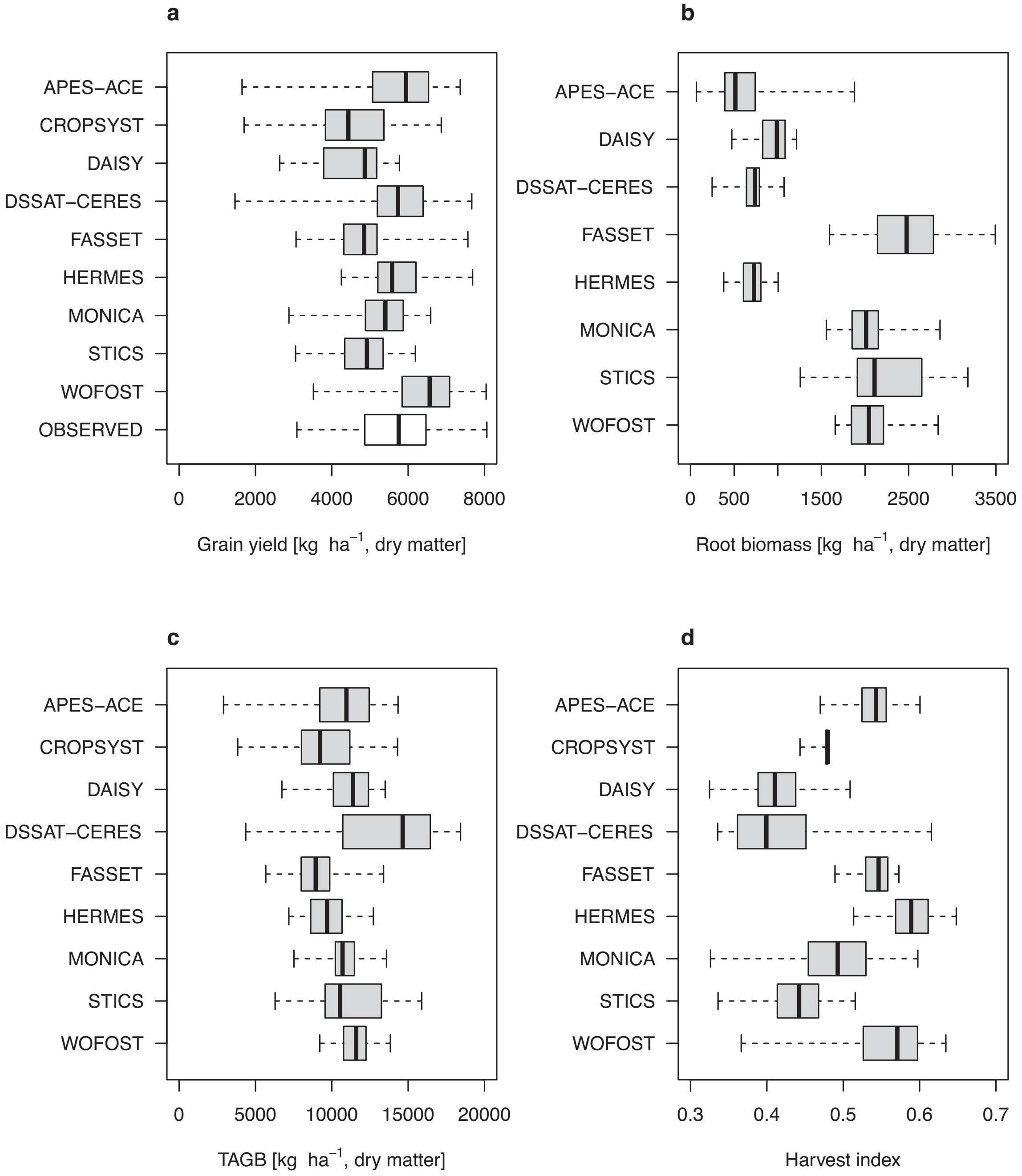

Fig. 5. Box-and-whisker plots of (a) grain yield estimates of models and observations, (b) root biomass estimates, (c) maximum above-ground biomass estimates, and (d) harvest indices of the models – among the simulated sites and years (N = 44). Boxes delimit the inter-quartile range (25–75 percentiles) and whiskers show the high and low extreme values.

rank correlation coefficients of 0.552 and 0.49, respectively, were performing best for all seasons (N = 44), while for Verovany site (N = 14), models DSSAT-CERES, WOFOST and HERMES showed highest rank correlation coefficients (0.539, 0.537 and 0.488, respectively).

4. Discussion

4.1. Uncertainty levels

Our results from this barley model comparison show that simulated grain yields vary widely, ranging from 1700 to 8100 kg ha^{-1} for all sites and seasons, being similar to the observed range (2400–8100 kg ha^{-1}). However, there were considerable differences in estimates for individual sites and years among the models (Figs. 3–5 and 7). Under conditions of limited data available for calibration (as in this blind test), uncertainty ranges in yield estimates from individual models are mostly not acceptable and beyond the measurement error of about 10–15% found in field experiments (Joernsgaard and Halmoe, 2003). This result is similar to the winter wheat study by Palosuo et al. (2011) and confirms that the differences in estimates of grain yield between models, and between the models and field observations have not much decreased when compared to earlier model comparisons for wheat, where yields were off by 20% and more (Goudriaan et al., 1994; Jamieson et al., 1998).

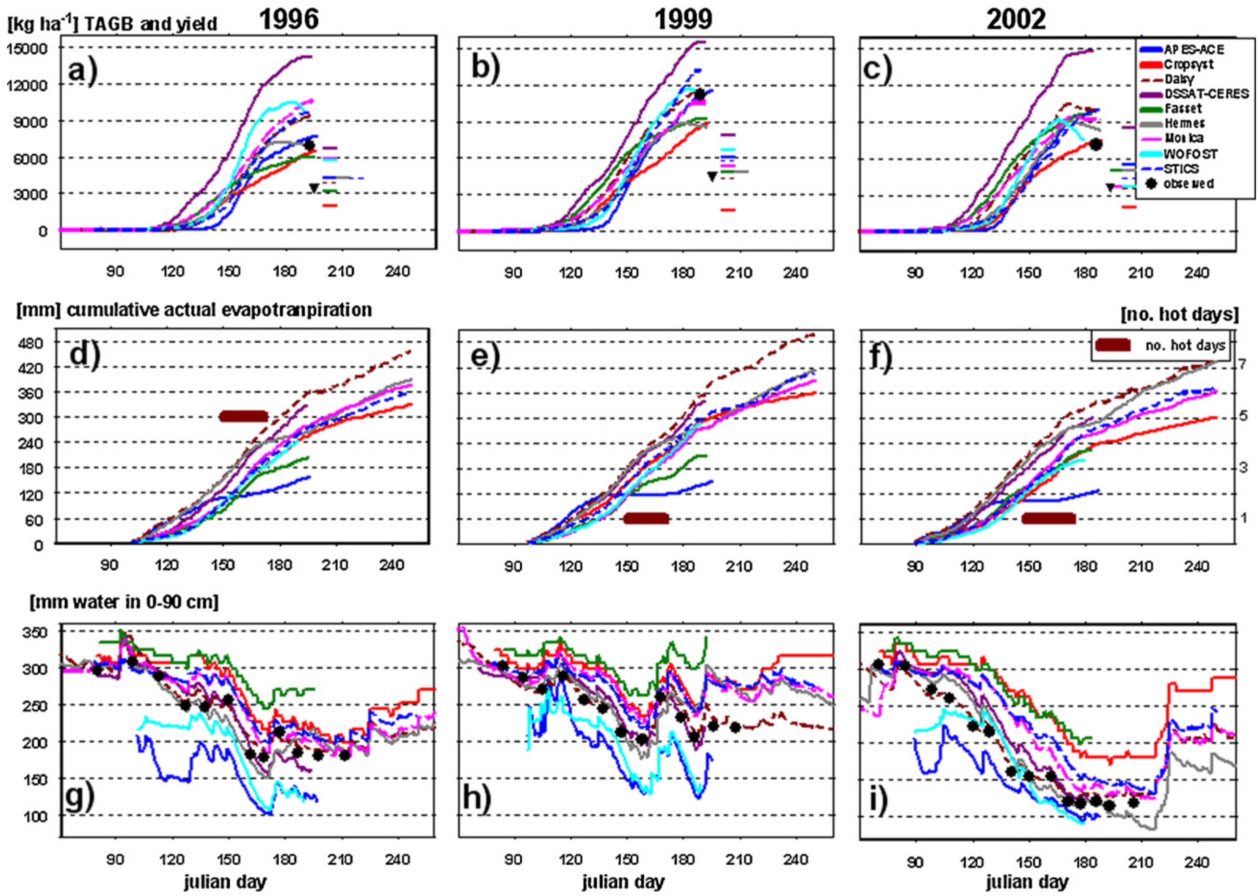

Fig. 6. Bratislava, 1996, 1999 and 2002: comparison of simulated above-ground biomass and grain yield (short lines) with observed biomass and grain yield (▼) at harvest (a–c), simulated cumulative evapotranspiration and observed number of hot days ($T_{\max} > 30\,^{\circ}\mathrm{C}$) between flowering and maturity (length of bar) (d–f), and simulated and observed soil water content in 0–90 cm profile (g–i).

Since our study focusing on spring barley is part of a series of model comparisons conducted in the framework of COST action 734 (www.cost734.eu/), we also aimed to compare results to the study for winter wheat by Palosuo et al. (2011), which was implemented in a similar way as the study here. Regarding that comparison, we hypothesized that uncertainties in model simulation results for barley exceed those for winter wheat, because there has been less modelling efforts and experimental data for barley than there has been for wheat.

For spring barley grain yield, RMSE of model estimates ranged from 1120 to 1940 kg ha^{-1}, while this was 1400–2300 kg ha^{-1} for winter wheat, and IA values for barley model estimates ranged from 0.31 to 0.63, while that range was between 0.40 and 0.74 for wheat yield estimates. If one acknowledges that mean observed yields differed to some degree (5800 and 6100 kg ha^{-1} for barley and wheat, respectively), results look in the end quite similar. In the winter wheat study, the range of simulated yield estimates did not cover the observed yield in 4 out of 49 (9%) seasons – and in those four cases, all models overestimated observed yield. The comparable figure for spring barley was 6 out of 44 (14%) seasons.

For estimating crop phenology there was not much difference in terms of accuracy. For instance, RMSE for estimating maturity of wheat had maximum values of 12.6 days, for barley this was 11.5 days. Such considerable discrepancies between simulated and observed phenology are not very surprising as the models just consider temperature and daylength in calculating numerical phenological development rate. However, a whole range of other factors, such as water deficits or nitrogen deficits can delay or hasten phenological development, and whether it is delaying or hastening depends on the timing of the stress (e.g. Jamieson et al., 1995; Asseng et al., 2011). None of the nine models describes these complex interactions sufficiently.

In summary, although there are some differences in results between spring barley and winter wheat, uncertainties in simulated yields appear to be at a comparable level, which is contrary to our initial hypothesis. The simulation period for winter wheat is longer and contains processes such as vernalization that are not relevant for spring barley. This may make model predictions of winter wheat more difficult. On the other hand, it should be borne in mind that these two model comparisons are not fully comparable: the seven sites in the barley exercise are more homogeneous than the nine sites used for wheat, both in terms of climate and soil conditions. From the 44 barley growing seasons analyzed, for instance, 30 experiments were conducted on Chernozems (Table 3).

Our results also further support the use of multi-crop model estimates in impact assessments. Apart from providing information on uncertainty ranges in model-based yield estimates, similar to the winter wheat study (Palosuo et al., 2011), the MMM appeared to be a better yield predictor than any individual model over all sites and seasons (Figs. 3 and 4) as well as at most individual sites (Fig. 7).

However, unlike in the winter wheat exercise and somewhat unexpectedly, the MMM (5253 kg ha^{-1}) over all sites is approximately 360 kg ha^{-1} lower than the observed mean (5617 kg ha^{-1}) (Figs. 4b and 5a). This overall underestimation can, to a large extent, be attributed to large underestimations of yields in several distinct seasons at Lednice, and to lesser extent at Verovany (Fig. 7).

Here we do not face the situation of one or more 'bad' models affecting the robustness of multi-model ensemble estimates, and

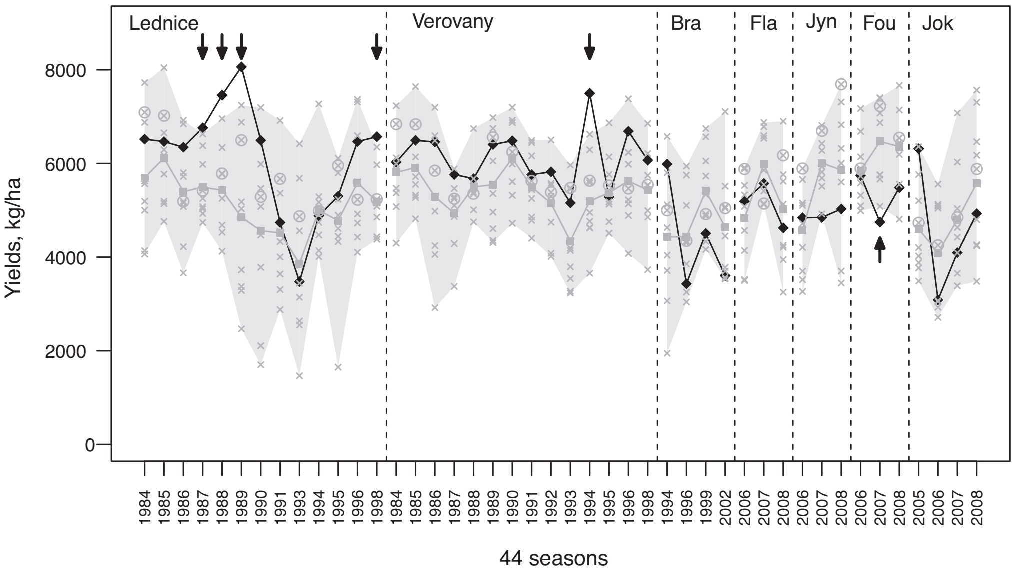

Fig. 7. Simulated and observed grain yield estimates [kg ha^{-1}, dry matter] for 44 studied growing seasons. Simulation results are shown in grey, while observations are shown in black. Multi-model means are connected by grey line, while for nine individual models simulated yields are given as grey crosses. Results for one of the best performing models (HERMES) are encircled. Arrows indicate observed yields that are not covered by model estimates (Bra = Bratislava, Fla = Flakkebjerg, Jyn = Jyndevad, Fou = Foulum, Jok = Jokioinen).

whether to exclude them (see Knutti, 2010), but rather a problem of inadequate parameterization and calibration which has possibly co-determined the underestimation at the Czech sites (for a detailed discussion, see Section 4.2).

4.2. Sources of uncertainties in model simulations

4.2.1. Parameterization and calibration method

Like for the model comparison for winter wheat (Palosuo et al., 2011), the model users were only allowed to calibrate crop phenology-related parameters based on provided phenological observations. Other parameters were taken from default values in the models or from some earlier applications of the models (Table 1). The applicability and quality of these parameters were not systematically tested due to the lack of suitable experimental data. But some part of systematic errors of model estimates is certainly related to these parameters.

As described in Section 3.2, the three models (HERMES, WOFOST and DAISY) that performed best in estimating grain yield, achieved this in different ways. What were the reasons for the differences among the models and the uncertainty sources? While this question goes beyond the scope of the current paper, we present some ideas as an initial contribution to what should become a more general, comprehensive discussion for crop model applications.

The most important reason for fairly systematic overestimation by WOFOST is the assumption that at no time nutrients are yield-limiting. Still, WOFOST underestimates the highest yields observed at Lednice and Verovany. This may be due to the fact that most of the experimental data used to parameterize barley in WOFOST are from the early 1980s (see Boons-Prins et al., 1993).

CROPSYST, on the other hand, has been noted to underestimate crop yields in Europe when applied with the default parameter set (Moriondo et al., 2010). The same was noted in this study with the initial model simulations strongly underestimating the yields, before the team was allowed to re-assess the parameters. They revised two parameters, radiation and water use efficiencies (RUE and WUE, respectively). RUE was set to 4.5 g MJ^{-1} (default = 3) and WUE to 6.5 kPa kg m^{-3} (default = 5) based on an unpublished simulation study for Poland.

As illustrated in Figs. 3, 4b and 5a, on average models showed a tendency to underestimate observed yield – and in particular for the Lednice and Verovany study sites, which covered 27 out of 44 studied seasons. Especially, all models failed to predict the high yields harvested in years 1987–89 and 1998 at Lednice, and in 1994 at Verovany. The underestimation of barley yield at the Lednice and Verovany sites suggests that none of the models' featured crop parameter sets suited to reproduce the growth and yield potential of that spring barley cultivar (cv Orbit) during those seasons (Table 3). This is surprising, as the DSSAT-CERES model had been calibrated for the same cultivar, and at nearby sites (Hlavinka et al., 2010), but obviously not under a 'potential production situation'.

Following the 'Wageningen approach' for calibration (van Ittersum et al., 2003), a crop model is first calibrated using data from experiments in the potential production situation, i.e. where crop growth, production and yield are only defined by radiation, temperature, CO_2 concentration of the atmosphere and crop characteristics. Subsequently, calibration is continued using experiments under (water and/or nitrogen) limited growth conditions, and finally, the gap to actual yields observed on farms can be attributed to yield-reducing factors, such as pest and diseases. Although yields observed on farms are usually subject to water or nutrient limitations (temporarily), and somewhat reduced by pest or disease occurrence, 'near-potential' production situations do exist, not only under irrigated (Dobermann et al., 2000), but also under rainfed conditions (Semenov et al., 2009). If crop models are not calibrated according to the approach sketched above, models easily fail to reproduce yields attainable in very favorable years (van Ittersum et al., 2003).

When we look at the barley cultivars, sites or growing environments the individual models were calibrated (Table 1), we find some similarities to the studied sites (Table 3). For instance, DAISY has been calibrated at sites near Bratislava, however, not for the same years or cultivars. DSSAT-CERES was calibrated for cultivars 'Akcent' and 'Orbit', not for the same, but at some nearby sites. It should be stressed, however, that previous versions of DSSAT-CERES (i.e. 3.0 and 3.5) were far easier to calibrate correctly for the local conditions than version 4.0. used in this study and that the calibration parameters are not easily transferable. WOFOST was calibrated for cultivar 'Scarlett' at Jokioinen, however, in earlier years and on different soils. While most models were calibrated at field scale for a number of experimental sites, models APES-ACE and CROPSYST were calibrated using actual yields at regional level (NUTS-2 level) within Europe. While this, for instance, explains the relatively good performance of DAISY at Bratislava (Fig. 4), and model DSSAT-CERES at Verovany (not shown), the model calibration and evaluation conditions only provide a partial explanation of the (site-specific and overall) differences among models and uncertainty sources.

4.2.2. Model input

While checking the field logbooks of the two Czech sites, we found additional explanations for the high yields that were related to input information provided for the modellers. At those sites spring barley was cultivated as a "second crop" with moderate doses of nitrogen, and yields crucially depended on fertilization of the previous crop. It has been suggested that estimates of initial nitrogen mistakenly have been too low. For example, there was an exceptionally high manure application to the previous (maize crop) in Lednice 1987, high manure applications and improved land preparation (new plough) in 1988; highest nitrogen application rates, and ideal weather with record yields in 1989, and crop residue effects in 1998. These examples highlight the challenges related to correct description of the cultivation conditions in the modelling exercise.

4.2.3. Model structure and complexity

Similar to Lednice 1987–1988, conditions at Verovany in 1994 are described as 'potential production situations', which was also confirmed by examining data from a second independent trial containing the same cultivar, and, by high yields of 20 other cultivars included in the experiment. According to the field logbook descriptions, the weather during the spring 1994 (March and April) was very favorable with the high positive effect on tillering (described as the best within the last decade) (Chmielewski and Köhn, 1999). This was supported by the observed high number of productive tillers. It was about 830 tillers per square meter in 1994 (e.g. in 1993 it was less than 500). Analyzing the weather records showed that in none of the years the amplitude between maximum and minimum temperature was as high as in 1994 in combination with very high irradiation levels. That means, low temperature minima implying low respiration losses during the night while having, at the same time, high assimilation rates during daytime. It is likely that the effects of weather conditions on tillering is not adequately described in our models.

Fig. 6 indicates another deficit in the structure (i.e. the processes covered) of most models, and that is their deficiency in adequately describing heat stress effects on cereal yields (Porter and Semenov, 2005; Semenov and Shewry, 2011), which can have considerable effects for climate change impact assessments (Lobell et al., 2012). Although heat is inherently considered in models with detailed photosynthesis-respiration approach (see Table 2) using optimum functions for temperature-gross assimilation relation and exponential increase of respiration with temperature, stress on grain formation is rarely explicitly considered. Only one model, MONICA, explicitly describes heat stress effects around flowering (Table 2). In MONICA, heat stress is described by a reduction factor that is dependent on calculated day and night temperature with a heat impact factor according to Challinor et al. (2005), using a critical and a limiting temperature threshold on the daytime temperature. Finally, the fraction of open flowers calculated according to Moriondo et al. (2011) is multiplied with the heat impact factor. Similar approaches have been applied to other models not included in the comparison, such as APSIM (Asseng et al., 2011).

As shown in Fig. 6, year 1996 in Bratislava that had the highest count of hot days, also had lowest yield. Nevertheless, it remains difficult to decide to what extent hot days and drought events limited yields – much depends on the exact timing of heat and drought stress (Jamieson et al., 1995; Savin and Nicolas, 1999). The latter authors found for barley that individual grain weight was most sensitive to heat stress and drought imposed early in grain filling, and less sensitive to late drought. As reported by Jamieson et al. (1995), early drought can cause both, long-lasting changes in the radiation use efficiency, as well as a reduction in the radiation intercepted, while, in contrast, later initiated drought usually leads to accelerated leaf senescence, and thus only reduced radiation interception. Although at Bratislava plant available soil moisture in 1996 was on average higher than in 2002, there was a moderate early drought in May followed by a pronounced drop of soil moisture by end of June (just before flowering) (Fig. 6g–i). Both, model simulations and observations showed reduced biomass production (Fig. 6a).

While we definitely do not span the full range of accessible, well-documented and widely applied models in climate change impact studies (see, e.g. White et al., 2011 for an overview), for the European situation we cover almost all of the important "model families". When examining the history of simulation techniques introduced to agricultural research, two schools play a predominant role, that of de Wit (Bouman et al., 1996) and that of DSSAT (Jones et al., 2003). While the 'Wageningen C.T. de Wit School' comprises the moderately complex SUCROS/WOFOST type and the less complex and less data demanding LINTUL-types (van Ittersum et al., 2003), the DSSAT school combines the moderately complex CERES and CROPGRO model families (see Bouman et al., 1996; Stöckle et al., 2003). Both schools have strongly influenced others such as the American CROPSYST, the Australian APSIM, and the French STICS – which all developed during the 1990s. APES-ACE, the most recent development, is largely based on LINTUL. In our set, APES-ACE and CROPSYST are the most simple, and DAISY the most complex. Results from our "blind tests" do not indicate a clear connection between model complexity and prediction error.

4.3. Data quality and requirements

Yield variability due to soil variability is considered an important aspect at several sites, as found in the Palosuo et al. (2011) study and elsewhere (Lawless et al., 2008). Model input on soil characteristics is given as point data, representative for the whole experiment and cannot capture within-field soil variability. For instance, within-field and -season yield variation at Verovany was considerable in some years, with the yield range exceeding 1000 kg ha^{-1} (>15% of mean yield) in 1991 and 1993, which is more than the intra-field variation reported for winter wheat elsewhere by Joernsgaard and Halmoe (2003). When looking for reasons for overestimating yields at Foulum site, it has been suggested that the high within-field variability of soil properties has resulted in overestimation of available soil moisture at that site.

Climate models and crop models have in common that they both suffer from considerable structural and parameter uncertainty and from lack of independent datasets to evaluate them thoroughly (Knutti, 2010). In order not to raise the question of circular reasoning (on why models are getting better), we strictly avoided to use the same data for model calibration and evaluation. In our case, this also resulted in some downsides, such as moderate data quality. As shown in this paper and its sister study (Palosuo et al., 2011), among others, in most European countries it is hardly possible to acquire fresh data from experiments suitable for thorough model calibration, validation and comparison. Apart from the data requirements listed in Table 4, such datasets also should include seasonal dynamics of LAI, above-ground biomass and N-content. Moreover, for comparison of crop models in drought stress or heat stress situations (e.g. Jamieson et al., 1998) additional data are needed and these should be generated through "manipulated environment" experiments. There will be no major progress in improving crop models without such new datasets (Semenov and Shewry, 2011; Lobell et al., 2012).

4.4. Perspectives on assessing uncertainties

A distinct merit of this study is that almost all major crop models applied for barley in Europe participated. To our knowledge, there is no earlier study for barley with such large number of models compared under various climatic conditions. Proposed next steps, that is, performing such comparison with more comprehensive and fresh datasets comprising sequential measurements of key variables, and for conditions of climatic change have already been carried out. That is, for barley at a Finnish site (Salo et al., companion paper, in preparation), and, for current and future climatic conditions for winter wheat in the framework of the Agricultural Model Intercomparison and Improvement project (AgMIP) (www.agmip.org).

Renewed interest and awareness of the need of comparing and improving crop models (see, e.g. Rötter et al., 2011a; White et al., 2011; Lobell et al., 2012) has also led to a couple of new research initiatives in Europe deepening and extending the work initiated under COST action 734 and reported in this paper. In these initiatives model intercomparisons for major food crops and crop rotations are being performed in contrasting locations, extending the number of models involved and comparing model responses to changes in temperature, precipitation and atmospheric CO_2 concentrations. The larger number of models, more contrasting environments and inclusion of sensitivity analyses will allow a more systematic assessment of model uncertainty ranges and treatment of questions related to the use of multi-crop model ensembles.

5. Conclusions

The results obtained suggest that application of crop models with limited calibration leads to high impact (yield, length of growing period) uncertainties. Furthermore, the degree of uncertainty for spring barley does not differ much from that for winter wheat (Palosuo et al., 2011). Another result parallel to the winter wheat comparison is that mean model predictions are in relatively good agreement with observed yields. This again supports the use of multi-model ensembles rather than relying on single models that reportedly perform well for specific regions or agro-ecological conditions. While some models performed better than others in estimating grain yields in specific environments, none was clearly superior or more robust in terms of yield prediction accuracy across all sites, for which the multi-model mean proved as the best predictor.

For both, winter wheat and barley models applied with restricted calibration, we conclude that uncertainty levels are not acceptable and that the models require crop cultivar and region-specific calibration and improvements before being used with confidence in regional climate impact assessments.

Acknowledgements

This study was carried out as a co-operative project under the umbrella of COST action 734 "Impacts of Climate Change and variability on European Agriculture (CLIVAGRI)" and the work of individual researchers was funded by various bodies:

R. Rötter, T. Palosuo: The strategic projects IAM-Tools and MODAGS funded by MTT Agrifood Research Finland, and project A-LA-CARTE, funded by the Academy of Finland (decision no. 140806)
J.E. Olesen, R. Patil: Impacts of climate change on cropping systems funded by the Danish Ministry of Food, Agriculture and Fisheries, Denmark.
P. Hlavinka: Soil Climate Atlas project funded by National Agency for Agricultural Research (QI91C054)
M. Trnka: CLIMSAVE 7 FP project (contract no. 244031) and CzechGlobe Centre (CZ.1.05/1.1.00/02.0073).

Finally, we thank two anonymous reviewers for their valuable comments.

References

Adam, M., et al., 2011. Effects of modelling detail on simulated crop productivity under a wide range of climatic conditions. Ecol. Model. 222, 131–143.
Asseng, S., et al., 2011. The impact of temperature variability on wheat yields. Glob. Change Biol. 17, 997–1012.
Bellocchi, G., et al., 2009. Validation of biophysical models: issues and methodologies. A review. Agron. Sustain. Dev. 30, 109–130.
Berntsen, J., et al., 2004. Modelling dry matter production and resource use in intercrops of pea and barley. Field Crops Res. 88, 69–83.
Boons-Prins, E.R., de Koning, G.H.J., van Diepen, C.A.,;1; 1993. Crop specific simulation parameters for yield forecasting across the European community. Simulation Reports CABO-TT, no. 32, Wageningen, The Netherlands.
Bouman, B.A.M., et al., 1996. The 'School of de Wit' crop growth simulation models: a pedigree and historical overview. Agric. Syst. 52, 171–198.
Brisson, N., et al., 2003. An overview of the crop model STICS. Eur. J. Agron. 18, 309–332.
Brouwer, F.M., van Ittersum, M. (Eds.), 2010. Environmental and Agricultural Modelling. Integrated Approaches for Policy Impact Assessment. Springer, Dordrecht, 322 pp.
Børgesen, C.D., Olesen, J.E., 2011. A probabilistic assessment of climate change impacts on yield and nitrogen leaching from winter wheat in Europe. Nat. Hazards Earth Syst. Sci. 11, 2541–2553.
Challinor, A.J., et al., 2005. Simulation of the impact of high temperature stress on annual crop yields. Agric. Forest Meteorol. 135, 180–189.
Chmielewski, F.-M., Köhn, W., 1999. Impact of weather on yield components of spring cereals over 30 years. Agric. Forest Meteorol. 96, 49–58.
Chirinda, N.N., et al., 2012. Root carbon input in organic and inorganic fertilizer-based systems. Plant Soil, (in press).
Confalioneri, R., et al., 2009. Multi-metric evaluation of the models WARM, CropSyst, and WOFOST for rice. Ecol. Model. 220, 1395–1410.
Corre-Hellou, G., et al., 2009. Adaptation of the STICS intercrop model to simulate crop growth and N accumulation in pea-barley intercrops. Field Crops Res. 113, 72–81.
Dobermann, A., et al., 2000. Reversal of rice yield decline in a long-term continuous cropping experiment. Agron. J. 92, 633–643.
Doltra, J., et al., 2011. Cereal yield and quality as affected by nitrogen availability in organic and conventional arable crop rotations. A combined modeling and experimental approach. Eur. J. Agron. 34, 83–95.
Donatelli, M., et al., 1997. Evaluation of CropSyst for cropping systems at two locations of northern and southern Italy. Eur. J. Agron. 6, 35–45.
Diekkrüger, B., et al., 1995. Validity of agroecosystem models a comparison of results of different models applied to the same data set. Ecol. Model. 81, 3–29.
Eitzinger, J., et al., 2004. Comparison of CERES WOFOST and SWAP models in simulating soil water content during growing season under different soil conditions. Ecol. Model. 171, 223–246.
Eitzinger, J., et al., 2012. Sensitivities of crop models to extreme weather conditions during flowering period demonstrated for maize and winter wheat in Austria. J. Agric. Sci., in press.
EUROSTAT, 2011. Agriculture and fishery statistics. Main Results 2009–10. EUROSTAT Pocketbooks, Belgium. ISBN 978-92-79-20424-1.
Ewert, F., et al., 2002. Effects of elevated CO_2 and drought on wheat Testing crop simulation models for different experimental and climatic conditions. Agric. Ecosyst. Environ. 93, 249–266.
Ewert, F., et al., 2011. Assessing the adaptive capacity of agriculture in the Netherlands to the impacts of climate change under different market and policy scenarios (AgriAdapt) Project of the Research Program Climate Change and Spatial Planning. Scenario development and assessment of the potential impacts of climate and market changes on crops in Europe. AgriAdapt Project Reports No. 2 & 3.
Ferrise, R., et al., 2011. Probabilistic assessments of climate change impacts on durum wheat in the Mediterranean region. Nat. Hazards Earth Syst. Sci. 11, 1293–1302.
Franko, U., Puhlmann, M., Kuka, K., Böhme, F., Merbach, I., 2007. Dynamics of water carbon and nitrogen in an agricultural used Chernozem soil in Central Germany. In: Kersebaum, K.C., Hecker, J.-M., Mirschel, W., Wegehenkel, M. (Eds.), Modelling Water and Nutrient Dynamics in Soil-Crop-Systems. Springer, Dordrecht, pp. 245–258.

Goudriaan, J., et al., 1994. GCTE Focus 3 Wheat Modelling and Experimental Data Comparison Workshop Report, Lunteren, The Netherlands, November 1993. GCTE Focus 3 Office. University of Oxford, Oxford, UK.

Groot, J.J.R., Verberne, E.L.J., 1991. Response of wheat to nitrogen fertilization, a data set to validate simulation models for nitrogen dynamics in crop and soil. Fert. Res. 27, 349–381.

Hlavinka, P., et al., 2010. The performance of CERES-Barley and CERES-Wheat under various soil conditions and tillage practices in Central Europe. Die Bodenkultur 61, 5–17, ISSN 0006-5471.

Hakala, K., et al., 2012. Sensitivity of barley varieties to weather. J. Agric. Sci. 150, 145–160, http://dx.doi.org/10.1017/S0021859611000694.

Hansen, S., et al., 1990. DAISY – A Soil Plant System Model. Danish Simulation Model for Transformation and Transport of Energy and Matter in the Soil–Plant–Atmosphere System. National Agency for Environmental Protection, Copenhagen.

Jamieson, P.D., et al., 1995. Drought effects on biomass production and radiation-use efficiency in barley. Field Crops Res. 43, 77–86.

Jamieson, P.D., et al., 1998. A comparison of the models AFRCWHEAT2 CERES-Wheat, Sirius, SUCROS2 and SWHEAT with measurements from wheat grown under drought. Field Crops Res. 55, 23–44.

Joernsgaard, B., Halmoe, S., 2003. Intra-field yield variation over crops and years. Eur. J. Agron. 19, 23–33.

Jones, J.W., et al., 2003. The DSSAT cropping system model. Eur. J. Agron. 18, 235–265.

Kabat, P., et al. (Eds.), 1995. Modelling and Parameterization of the Soil–Plant–Atmosphere System. A Comparison of Potato Growth Models. Wageningen Press, Wageningen, The Netherlands, 513 pp.

Kersebaum, K.C., et al., 2007. Modelling water and nutrient dynamics in soil–crop systems: a comparison of simulation models applied on common data sets. In: Kersebaum, K.C., Hecker, J., Mirschel, W., Wegehenkel, M. (Eds.), Modelling Water and Nutrient Dynamics in Soil–Crop Systems. Springer, Dordrecht, pp. 1–17.

Kleemola, J., et al., 1995. Modelling the impact of climatic change on growth of spring barley in Finland. J. Biogeogr. 22, 581–590.

Knutti, R., 2010. The end of model democracy? An editorial comment. Climatic Change 102, 395–404.

Kobayashi, K., Salam, M.U., 2000. Comparing simulated and measured values using mean squared deviation and its components. Agron. J. 92, 345–352.

Launay, M., et al., 2009. Exploring options for managing strategies for pea-barley intercropping using a modeling approach. Eur. J. Agron. 31, 85–98.

Lawless, C., et al., 2008. Quantifying the effect of uncertainty in soil moisture characteristics on plant growth using a crop simulation model. Field Crops Res. 106, 138–147.

Lobell, et al., 2012. Extreme heat effects on wheat senescence in India. Nat. Climate Change, published online (29.01.12). http://dx.doi.org/10.1038/nclimate1356.

Metzger, M.J., Bunce, R.G.H., Jongman, R.H.G., Mucher, C.A., Watkins, J.W., 2005. A climatic stratification of the environment of Europe. Glob. Ecol. Biogeogr. 14, 549–563.

Moriondo, M., et al., 2010. Impact and adaptation opportunities for European agriculture in response to climatic change and variability. Mitigation Adaptation Strategies Global Change 15, 657–679.

Moriondo, M., et al., 2011. Climate change impact assessment: the role of climate extremes in crop yield simulation. Climatic Change 104, 679–701.

Nendel, C., et al., 2011. The Monica model testing predictability for crop growth, soil moisture and nitrogen dynamics. Ecol. Model. 222, 1614–1625.

Olesen, et al., 2000. Design of an organic farming crop-rotation experiment. Acta Agric. Scand., Sect. B: Soil Plant Sci. 50, 13–21.

Palosuo, T., et al., 2011. Simulation of winter wheat yield and its variability in different climates of Europe: a comparison of eight crop growth models. Eur. J. Agron. 35, 103–114.

Passarella, V.S., et al., 2005. Breeding effects on sensitivity of barley grain weight and quality to events of high temperature during grain filling. Euphytica 141, 41–48.

Peltonen-Sainio, P., et al., 2008. Variation in harvest index of modern spring barley oats and wheat cultivars adapted to Northern growing conditions. J. Agric. Sci. 146, 35–47.

Porter, J.R., Semenov, M.A., 2005. Crop responses to climatic variation. Philos. Trans. Roy. Soc. B 360, 2021–2035.

Porter, J.R., et al., 1993. Comparison of the wheat simulation models ARFWHEAT2 CERES-Wheat and SWHEAT for non-limiting conditions of crop growth. Field Crops Res. 33, 131–157.

Refsgaard, J.C., et al., 1998. An Integrated Model for the Danubian Lowland – Methodology and Applications. Water Resources Management 12. Kluwer Academic Publishers, pp. 433–465.

Rosenzweig, C., Wilbanks, T.J., 2010. The state of climate change vulnerability, impacts, and adaptation research: strengthening knowledge base and community. Climatic Change 100, 103–106.

Rötter, R.P., et al., 2011a. Crop-climate models need an overhaul. Nat. Climate Change 1, 175–177.

Rötter, R.P., et al., 2011b. What would happen to barley production in Finland if global warming exceeded 4 °C? A model-based assessment. Eur. J. Agron. 35, 205–214.

Salo, T., et al. Comparing the performance of eleven agro-ecosystems models in predicting crop yield response to nitrogen under Finnish weather conditions. Field Crops Res., in preparation.

Sapkota, T.B., et al., 2012. Effects of catch crop type and root depth on nitrogen leaching and yield of spring barley. Field Crops Res. 125, 129–138.

Savin, R., Nicolas, M.E., 1999. Effects of timing of heat stress and drought on growth and quality of barley grains. Aust. J. Agric. Res. 50, 357–364.

Semenov, M.A., et al., 2009. Quantifying effects of simple wheat traits on yield in water-limited environments using a modeling approach. Agric. Forest Meteorol. 149, 1095–1104.

Semenov, M.A., Shewry, P.R., 2011. Modelling predicts that heat stress, not drought, will increase vulnerability of wheat in Europe. Sci. Rep. 1, http://dx.doi.org/10.1038/srep00066.

Smith, P., et al., 1997. A comparison of the performance of nine soil organic matter models using datasets from seven long-term experiments. Geoderma 81, 153–225.

Stöckle, C.O., et al., 2003. CropSyst a cropping systems simulation model. Eur. J. Agron. 18, 289–307.

Svendsen, H., et al., 1995. Simulation of crop production water and nitrogen balances in two German agro-ecosystems using the DAISY model. Ecol. Model. 81, 197–212.

Trnka, M., et al., 2004. Climate change impacts and adaptation strategies in spring barley production in the Czech Republic. Climatic Change 64, 227–255.

Trnka, M., et al., 2011. Agroclimatic conditions in Europe under climate change. Glob. Change Biol. 17, 2298–2318, http://dx.doi.org/10.1111/j. 1365-2486.2011.02396.x.

Tubiello, F.N., Ewert, F., 2002. Simulating the effects of elevated CO_2 on crops: approaches and applications for climate change. Eur. J. Agron. 18, 57–74.

van Ittersum, M.K., et al., 2003. On approaches and applications of the Wageningen crop models. Eur. J. Agron. 18, 201–234.

Wallach, D., et al. (Eds.), 2006. Working with Dynamic Models. Evaluation, Analysis, Parameterization, and Applications. Elsevier, Amsterdam, 462 pp.

White, J.W., et al., 2011. Methodologies for simulating impacts of climate change on crop production. Field Crops Res., http://dx.doi.org/10.1016/j.fcr.2011.07.001.

Willmott, C.J., 1981. On the validation of models. Phys. Geogr. 2, 184–194.

Wolf, J., et al., 1996. Comparison of wheat simulation models under climate change I. Model calibration and sensitivity analyses. Climate Res. 7, 253–270.