CrossMark

Performance of 13 crop simulation models and their ensemble for simulating four field crops in Central Europe

M. Kostková[1,2], P. Hlavinka[1,2] , E. Pohanková[1,2], K. C. Kersebaum[2,3], C. Nendel[2,3,4], A. Gobin[5,6] , J. E. Olesen[2,7], R. Ferrise[8], C. Dibari[8], J. Takáč[9], A. Topaj[10], S. Medvedev[10], M. P. Hoffmann[11] , T. Stella[3], J. Balek[1,2], M. Ruiz-Ramos[12], A. Rodríguez[12,13], G. Hoogenboom[14], V. Shelia[14], D. Ventrella[15] , L. Giglio[15], B. Sharif[7], I. Ozturk[7], R. P. Rötter[11,16], J. Balkovič[17,18], R. Skalský[9,17], M. Moriondo[8,19], S. Thaler[2,20], Z. Žalud[1,2] and M. Trnka[1,2]

[1]Institute of Agriculture Systems and Bioclimatology, Mendel University in Brno, Zemedelska 1, 613 00 Brno, Czech Republic; [2]Global Change Research Institute Academy of Sciences of the Czech Republic, Belidla 986/4b, 603 00 Brno, Czech Republic; [3]Leibniz Centre for Agricultural Landscape Research, Eberswalder Straße 84, 15374 Müncheberg, Germany; [4]University of Potsdam, Am Mühlenberg 3, 14476 Potsdam (Golm), Germany; [5]Flemish Institute for Technological Research (VITO), Boeretang 200, 2400 Mol, Belgium; [6]Department of Earth and Environmental Sciences, Faculty of BioScience Engineering, Celestijnenlaan 200E, 3001 Heverlee, Belgium; [7]Department of Agroecology, Aarhus University, Blichers Allé 20, 8830 Tjele, Denmark; [8]Department of Agriculture, Food, Environment, and Forestry, University of Florence, P.le delle Cascine 18, 50144, Firenze, Italy; [9]National Agricultural and Food Centre, Soil Science and Conservation Research Institute, Trenčianska 55, 821 09 Bratislava, Slovakia; [10]Agrophysical Research Institute, Grazhdansky pr., 14, 195220, Saint-Petersburg, Russia; [11]Agvolution GmbH, Tropical Plant Production and Agricultural Systems Modelling (TROPAGS), Georg-August-University Göttingen, Grisebachstraße 6, 37077, Göttingen, Germany; [12]Ceigram – Research Centre for the Management of Agricultural and Environmental Risks, Universidad Politécnica de Madrid – Ciudad Universitaria, Madrid 28040, Spain; [13]Department of Economic Analysis and Finances, Universidad de Castilla-La Mancha, 45071, Toledo, Spain; [14]Institute for Sustainable Food Systems, University of Florida, 185 Rogers Hall, P. O. Box 110570, Gainesville, Florida 32611, USA; [15]Consiglio per la Ricerca in Agricoltura e l'analisi dell'economia agraria, Centro di Ricerca Agricoltura e Ambiente, (CREA-AA), Via Celso Ulpiani 5, 70125 Bari, Italy; [16]Centre for Biodiversity and Sustainable Land Use (CBL), Georg-August-University Göttingen, Buesgenweg 1, 37077 Göttingen, Germany; [17]International Institute for Applied Systems Analysis (IIASA), Biodiversity and Natural Resources Program (BNR), Schlossplatz 1, A-2361 Laxenburg, Austria; [18]Faculty of Natural Science, Comenius University, Ilkovičova 6, 842 15 Bratislava, Slovakia; [19]CNR IBE, via Madonna del Piano 10, 50019, Firenze, Italy and [20]Institute of Meteorology and Climatology (BOKU-Met), University of Natural Resources and Life Sciences, Vienna, Gregor-Mendel-Straße 33, 1180 Vienna, Austria

Abstract

The main aim of the current study was to present the abilities of widely used crop models to simulate four different field crops (winter wheat, spring barley, silage maize and winter oilseed rape). The 13 models were tested under Central European conditions represented by three locations in the Czech Republic, selected using temperature and precipitation gradients for the target crops in this region. Based on observed crop phenology and yield from 1991 to 2010, performances of individual models and their ensemble were analyzed. Modelling of anthesis and maturity was generally best simulated by the ensemble median (EnsMED) compared to the ensemble mean and individual models. The yield was better simulated by the best models than estimated by an ensemble. Higher accuracy was achieved for spring crops, with the best results for silage maize, while the lowest accuracy was for winter oilseed rape according to the index of agreement (IA). Based on EnsMED, the root mean square errors (RMSEs) for yield was 1365 kg/ha for winter wheat, 1105 kg/ha for spring barley, 1861 kg/ha for silage maize and 969 kg/ha for winter oilseed rape. The AQUACROP and EPIC models performed best in terms of spread around the line of best fit (RMSE, IA). In some cases, the individual models failed. For crop rotation simulations, only models with reasonable accuracy (i.e. without failures) across all included crops within the target environment should be selected. Application crop models ensemble is one way to increase the accuracy of predictions, but lower variability of ensemble outputs was confirmed.

Introduction

Process-based crop models are commonly used to simulate growth, development, yield and other characteristics of field crops as well as related soil processes such as the soil, water and nitrogen balances. These models simulate such interconnected soil–plant–atmosphere system that is

influenced by agronomic practices (e.g. Ritchie, 1981; Jones *et al.*, 2003). The use of different crop models for assessing possible future climate change impacts, including analyses oriented to derive adaptation strategies, is increasing (Corbeels *et al.*, 2018; Rötter *et al.*, 2018). Due to the nature of modelling, there is always a certain degree of uncertainty (e.g. due to the need for certain simplification against the real system) within results which may raise doubts about the reliability of the results (Rötter *et al.*, 2011). The use of crop model ensembles (against only one model with its specifics) is among the recommended approaches (Martre *et al.*, 2015; Wallach *et al.*, 2016) enabling estimation of the range of outputs, leading to more robust predictions and the possibility of reducing uncertainties. Several studies that applied model ensembles across different conditions have shown that the ensemble median (EnsMED) or mean (EnsAVG) provide more robust assessments than individual models (e.g. Rötter *et al.*, 2012; Asseng *et al.*, 2013). Simultaneously, compliance of individual models can be shown from ensemble simulations in relation to monitored variables (e.g. Palosuo *et al.*, 2011). Usually, model ensembles were tested only for one individual crop, for example, the ensemble of 23 corn models presented by Bassu *et al.* (2014), the ensemble of eight models for winter wheat by Palosuo *et al.* (2011) or the ensemble of 11 models for spring barley (Salo *et al.*, 2016). The focus on individual crops separately is due to the fact that ensemble modelling is a technical and time-demanding procedure concerning data availability, technical aspects and personnel requirements. On the other hand, model ensemble evaluation focused simultaneously on several crops, is desired as the first step to enable simulating crop sequences and crop rotations by an ensemble approach. The simulations of crop sequences or rotations are closer to real-world conditions, with higher relevance of results (Kollas *et al.*, 2015), because carry-over effects between seasons can be considered. At the same time, the long-term effects of weather, cultivated crops and agronomic management on soil properties can be assessed (Kersebaum, 2007; Hlavinka *et al.*, 2014), or climate change adaptation options can be tested (Hlavinka *et al.*, 2015). However, there is generally lower availability of data for model calibration from uninterrupted observations within crop rotation experiments. To cover the sufficient number of samples of each individual crop (or certain cultivars) of rotations, this could be alternatively solved through the data from individual crops experiments under defined conditions, which was applied in the current study.

The current paper focuses on simultaneous modelling of four separate crops (winter wheat, spring barley, silage maize and winter oilseed rape) based on field experiments with comparable methodology and compares 13 crop growth models. These comparisons were made under contrasting climatic conditions represented by three different sites in Central Europe. The main objective of the current study was to identify potentially best-performing models and to compare them with the ensemble approach, which could be used, for example, for impact and adaptation measures assessment studies both under present and expected climatic conditions. The present study is an essential first step for selecting the most proper models that would be suitable for modelling uninterrupted crop sequences or rotations by ensemble or individual models.

Materials and methods

Models

The current study included 13 crop growth models varying in complexity and functionality. Concurrently, two models were operated by different users and were independently evaluated. The DAISY model was operated by two independent modelling groups (marked as DAISY_1 and DAISY_2) and the DSSAT model was operated by three independent modelling groups (marked as DSSAT_1, DSSAT_2 and DSSAT_3). Specifically, 28 modellers from 10 countries participated in the study. The models that were used and their appropriate references are listed in Table 1, and the major characteristics of the participating models are summarized in Table 2.

Experimental sites and crop selection

The test sites selected in the current study (spatial locations in Fig. 1) represent substantial temperature and precipitation gradients for winter wheat, spring barley, silage maize and winter oilseed rape cultivation within the Czech Republic, and the sites are also representative of wider Central European conditions. The Lednice experimental site represents a warm and relatively dry region, Věrovany is a production area with a fertile soil, where warm and mostly sufficient rainy conditions prevail and Domanínek is characterized as a colder and wetter production area (Table 3).

Selected crops represent more than 0.6 of arable land in the Czech Republic and are crucial also within neighbouring countries. Winter wheat (variety Samanta), spring barley (variety Tolar), silage maize (variety Cefran in Lednice and Věrovany and variety Cingaro in Domanínek) and winter oilseed rape (variety Artus) were selected as representative varieties for each experimental site:

- Samanta (registration 2006) ranks among the semi-early winter wheat varieties with medium-grain and medium tillering. Resistance to overwintering and tolerance to late sowing are advantages of this variety.
- Spring barley variety Tolar (registration 1997) represents the semi-early malting variety preferred by malt-houses to produce Czech beer. For the warmer production area, Tolar has moderately high grain yield. The plants have medium to high heights and medium resistance to lodging (CISTA 2007).
- The Cefran variety (registration 2003) is a medium-late maize hybrid suitable for silage within the warm production area. The high yield and starch content in total dry matter are an advantage of Cefran (CISTA 2004). The variety Cingaro (registration 2006) is among the very early hybrids achieving above-average dry matter yield. Cingaro is resistant to cold weather, and its fibre is well digestible (Dobos, 2004).
- The Artus variety (registration 1999) of winter oilseed rape is a medium early hybrid variety. Plant height is medium to high, which causes the crop to have low to medium resistance to lodging. It tolerates overwintering and is characterized by high grain yield and very low glucosinolate content (CISTA 2006).

Model calibration and criteria for evaluation

The results of models and their ensemble were evaluated on the basis of acquired available data sets from rainfed variety trials conducted by the Central Institute for Supervising and Testing in Agriculture (CISTA). Data from the period 1991 to 2010 for the three selected experimental sites and for each crop (winter wheat, spring barley, silage maize and winter oilseed rape) were used. There was the comparable experimental protocol along all

Table 1. List of models, their abbreviations used in the current study and references

Model	Version	Abbreviation	References
AGROTOOL	3.4	AG	Poluektov *et al.* (2002)
APSIM	7.7	AP	Holzworth *et al.* (2014)
AQUACROP	6.0	AQ	Steduto *et al.* (2009)
CROPSYST	3.04	CS	Stöckle *et al.* (2003)
DAISY_1	5.24	DA1	Hansen *et al.* (1990)
DAISY_2	4.1	DA2	Hansen *et al.* (1990)
DSSAT_1	4.5, 4.6	DS1	Jones *et al.* (2003), Hoogenboom *et al.* (2015)
DSSAT_2	4.6	DS2	Jones *et al.* (2003), Hoogenboom *et al.* (2015)
DSSAT_3	4.6	DS3	Jones *et al.* (2003), Hoogenboom *et al.* (2015)
EPIC	0801	EP	Williams *et al.* (1989)
FASSET	2.1	FA	Olesen *et al.* (2000)
HERMES	4.26	HE	Kersebau. (2007)
MONICA	1.2.9	MO	Nendel *et al.* (2011)

In the case of DAISY_1 and DAISY_2, the same model was used, but the simulations were performed by two different modelling teams. Analogously, in the case of DSSAT_1 (v4.5: spring barley, winter wheat, silage maize, v4.6 winter oilseed rape), DSSAT_2 and DSSAT_3, the simulations were performed by three separate teams.

Table 2. Parameters and modelling approaches of the included models

Parameters										
	AG	AP	AQ	CS	DA1, DA2	DS1, DS2, DS3	EP	FA	HE	MO
Leaf area, development and LI[a]	S	S	S	S	D	S	S	D	D	D
Light utilization[b]	P-R	RUE	TE	RUE	P-R	RUE	RUE	RUE	P-R	P-R
Crop phenology[c]	T, O	T, DL, V, O	T, O, V	T, DL, V	T, DL, V	T, DL, V	T, DL	T, DL	T, DL, V, O	T, DL,V, O
Yield formation[d]	B	Prt	HI, B	HI, B	Prt	HI(Gn), B	HI, B	HI,B	Prt	Prt
Water dynamics[e]	R	C	C	C	R	C	C	C	C	C
Evapo-transpiration[f]	mPM	PT	PM	PT	PM	PT	PM	Mak	PM	PM
Soil CN-model[g]	CN, P(3)	CN, P(3)	C, P(5), B	N	CN, P(6), B	CN, P(4), B	CN, P(5)	CN, P(6), B	N, P(2)	CN, P(6), B

[a]Leaf area development and light interception: S – simple or D – detailed approach.
[b]Light utilization/biomass growth: RUE (simple approach) – Radiation use efficiency, P-R (detailed approach) – Gross photosynthesis minus respiration, and TE – transpiration efficiency biomass growth.
[c]Crop phenology is a function of T – temperature, DL – photoperiod (day length), V – vernalization, and O – other water (nutrient) stress effect considered.
[d]Yield formation depending on HI – harvest index, B – total (above-ground) biomass, Gn – number of grains, and Prt – partitioning during reproductive stages.
[e]Water dynamics approach (app.): C – capacity app. and R – Richards app.
[f]Evapo-transpiration estimation methods: P – Penman, PM – Penman-Monteith, PT – Priestley-Taylor, Mak – Makking, and mPM – modified Penman-Monteith.
[g]Soil CN model, C – C model, N – N model, P(*x*) – *x* = organic matter pools number, and B – microbial biomass.

sites. Moreover, constant cultivars of selected crops were included throughout the whole period, so the technological trend did not have to be considered. The observed experimental data that were available included the dates of sowing and harvest, amounts of seeds sown per m², observed phenological phases (emergence, tillering, shooting, heading, flowering and maturity), the number of tillers per m², the weight of 1000 seeds, fertilizer application data (timing and amount fully representing real field experiment management) and information on the previous crop. Additionally, the information about texture, bulk density, total pores, hydrolimits (field capacity, wilting point), the content of organic carbon, total nitrogen and pH for defined layers of

included soil profiles (one specific soil profile for each station) were available. Not known from the trials was information on tillage and was thus not available to the modellers. Initial conditions for the available soil water (% volumetric) at depths of 0.0–0.4 and 0.4–1.5 m and soil mineral nitrogen contents (kg N/ha) at depths 0.0–0.4 and 0.0–0.8 m were set. For winter oilseed rape, the initial conditions were defined on 1 August, for winter wheat 1 September, for spring barley and silage maize 1 November of each year (constant dates within the years for individual crops). As soil moisture was not observed, the SoilClim model (Hlavinka *et al.*, 2011) was used to estimate soil water content at the beginning of each simulation based on preceding

Fig. 1. Right: The Czech Republic within Europe (in red). Left: Locations of the Lednice, Věrovany and Domanínek experimental sites in the Czech Republic.

Table 3. Experimental sites' characteristics (the reference period for the average annual temperature and precipitation is 1981–2010)

	Locality (latitude/longitude)		
Characteristics	Lednice (48°48′N/ 16°48′E)	Věrovany (49°28′N/17°17′E)	Domanínek (49°31′N/16°14′E)
Altitude (m a. s. l.)	171	210	560
Soil type	Chernozem	Chernozem	Dystric cambisol
Soil plant available water (mm per 1.5 m soil depth)	260	269	180
Mean soil wilting point (%)	15.5	15.0	13.1
Mean soil field capacity (%)	32.9	32.9	25.1
Average annual temperature (°C)	9.9	9.0	7.2
Average annual precipitation (mm)	527	553	626

meteorological conditions. Further, the observed initial soil mineral nitrogen content was not available. It was alternatively estimated from the average of the measured amounts at each of included sites and from parallel CISTA field experiments focused on nitrogen balance, ranging from 85 to 140 kg of mineral N in the soil per hectare. The sowing dates within the database differed from year to year according to the suitability of the current weather conditions and soil moisture. The mean date of sowing was on 28 August for winter oilseed rape, on 2 October for winter wheat, on 3 April for spring barley and 30 April for silage maize.

The concept of the study was not in the form of a classical division into a calibration part of the database and a subsequent validation on an independent data sample. Rather only a minimal calibration concept was adopted with subsequent evaluation of results based on the whole data set (without independent data sample). This approach was used due to the absence of more detailed observations (such as the development of aboveground biomass, leaf area, nitrogen content in biomass, measured initial

soil conditions before sowing, etc.), which would be required for a detailed calibration (Kersebaum *et al.*, 2015). This mimics the typical situation of a regional parameter adjustment and assumed, that models were already parameterized for the specific crop and require only a minimum adjustment to reflect the regional varieties. This procedure was already applied by several model ensemble studies on climate change impacts (e.g. Asseng *et al.*, 2013; Bassu *et al.*, 2014). So, it was an adjustment of temperature sums to mimic the phenological development of an included variety of each crop, cultivar coefficients (i.e. in case of DSSAT models), adjustment of the assimilates partitioning and harvest index. In the case of the DSSAT and DAISY models, the parameter settings differed from group to group, as each was calibrated independently. Each crop was represented by a constant cultivar throughout the whole period and therefore, there was a requirement to use only one parameter set per crop/cultivar by each model. The results of each model after minimum calibration were evaluated on the basis of observed

phenological phases (namely, anthesis and maturity) and to fit observed average levels and variability of yearly crop yield. Because specific additional observations were not available such as for leaf area index and total aboveground biomass, the range of outputs for these variables were evaluated on the basis of reasonable range.

To evaluate the simulations, three statistical indexes were used. Mean bias error (MBE; Addiscott and Whitmore 1987) indicates positive and negative deviations (i.e. average systematic error). Root mean square error (RMSE, Fox 1981) is the standard deviation of residuals or prediction errors and describes the spread of residuals. If the RMSE values are lower, the simulations are concentrated around the best-fit line. Index of agreement (IA; Willmott 1982) can detect additive and proportional differences in the observed and simulated means and variances (Legates and McCabe, 1999). IA evaluates modelling interpretation results in a range between 0 and 1. An IA closer to 1 indicates a better simulation quality, similar to the coefficient of determination (Davies and McKay, 1988).

$$MBE = \sum_{i=1}^{n} \frac{S_i - O_i}{n} \qquad (1)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (S_i - O_i)^2}{n}} \qquad (2)$$

$$IA = 1 - \frac{\sum_{i=1}^{n} (S_i - O_i)^2}{\sum_{i=1}^{n} (|S_i - \bar{O}| + |O_i - \bar{O}|)^2} \qquad (3)$$

where S_i is the simulated value of the variable, O_i is the observed value of the variable, $\bar{O}$ is the mean value of the observed data and n is the number of pairs of observed/estimated values.

In addition to the evaluations of individual models, the median (abb. EnsMED) and the mean (abb. EnsAVG) from all model simulations for each combination of crop, year, station and output parameters were analyzed as separate indicators.

Results

Anthesis

The simulations for winter wheat anthesis across all the sites showed (see Table 4 and Fig. 2) that CROPSYST and DAISY_1 simulated anthesis very well, with an IA of 0.96, followed by all other models whose results were satisfactory (IA varied from 0.84 to 0.95, except -FASSET with an IA of 0.37). Regarding the absolute error, CROPSYST and DAISY_1 achieved RMSEs equal to 2.9 days across all the sites, and the remaining models (except FASSET, which had an RMSE equal 12.6 days) achieved RMSEs ranging from 3.2 to 8.3 days. Fully comparable results with the best models for all sites together were achieved by EnsMED (IA = 0.97, MBE = 0.8 days, and RMSE = 2.9 days). EnsAVG showed a slightly lower accuracy (IA = 0.96, MBE = 1.2 days, and RMSE = 3.2 days).

For spring barley anthesis, the best agreement was achieved by AQUACROP (all sites considered), with an IA of 0.96, MBE of 1.6 days and RMSE of 4.0 days. Models CROPSYST, DAISY_2, DSSAT_2, DSSAT_3, EPIC and MONICA achieved very similar IAs (0.95). Again, the FASSET model gave the least precise results for spring barley, but was more accurate than for winter wheat, with an IA equal to 0.77, an RMSE of 9.5 days, and simulated anthesis was systematically 7.6 days earlier (based on MBE and across all sites). EnsMED and EnsAVG provided results comparable to the best models for spring barley (IAs = 0.96, MBEs −0.6 and 0.0 days, respectively, RMSEs 4.0 and 3.7 days, respectively).

For silage maize anthesis, the AQUACROP and DAISY_1 models had the highest IA of 0.98 for all the sites. The remaining models (except FASSET) also had very satisfactory results, with IAs varying from 0.97 to 0.94 (for all the sites). The lowest value of IA (0.85) corresponded to the FASSET model. The systematic error MBEs varied from −3.0 to 1.6 days, and the absolute errors (RMSE) varied from 2.8 to 9.1 days based on individual models and considering all the sites together. EnsMED and EnsAVG had the highest indexes of agreement, 0.98 and 0.99, respectively, considering all the sites, with systematic errors of 0.1 and −0.2 days, respectively, and RMSEs of 3.0 and 2.7 days, respectively.

Although flowering was successfully estimated for the previous three crops (except for FASSET in the case of winter wheat), for winter oilseed rape, lower accuracy was generally achieved, either in terms of individual models or based on EnsMED and EnsAVG. For all the sites, the highest IAs were achieved by MONICA (0.91), AQUACROP (0.90) and FASSET (0.89). AGROTOOL had the lowest accuracy, with an IA equal to 0.38 and systematic and absolute errors of ~1 month. The general feature was a later estimate of the time of flowering (more pronouncedly in the cooler location of Domanínek). Only in the case of winter oilseed rape did EnsMED and EnsAVG produce worse results than the best individual models. Using EnsMED and EnsAVG, the indexes of the agreement were 0.76 and 0.73, respectively, the MBEs were 7.7 and 9.2 days, respectively, and the RMSEs were 9.8 and 11.5 days, respectively (considering all the sites).

Generally, the anthesis dates for winter wheat, spring barley and winter rape were overestimated (later than observed) in the case of the coolest station, Domanínek.

Maturity

In terms of winter wheat maturity, the highest IA (equal to 0.98) was achieved for model DAISY_1, with a low MBE at 1.1 days and an RMSE of 3.9 days (see Table 5). The IAs for the remaining models, except FASSET, which in this case also provided the least accurate results, varied from 0.86 to 0.96 (MBEs from −7.0 to 5.9 days and RMSEs from 4.6 to 10.8 days). Repeatedly, the FASSET model showed the poorest agreement (IA = 0.42, MBE = −16.5 days and RMSE = 24.2 days). EnsMED and EnsAVG had IAs of 0.96, MBEs in the range of −0.6 to −1.6 days and RMSEs from 4.8 to 5.0 days, respectively. Both EnsMED and EnsAVG were also able to produce reasonable values when obvious model outliers appeared (see Fig. 2, winter wheat).

Similar results were achieved for the simulations of spring barley maturity but with more accurate results in the case of FASSET (against winter wheat). Regarding the IA, the APSIM and CROPSYST models achieved the highest values (0.96) for all the sites. AQUACROP along with DSSAT_2, DSSAT_3 and MONICA reached IAs equal to 0.95. The least accurate results were attained by DSSAT_1, with an IA of 0.77 connected with a systematically later prediction of anthesis (MBE = 6.4 days) and high absolute error (RMSE = 14.0 days). The high values of MBE (11.9 days) and RMSE (14.4 days) were also achieved by

Table 4. The evaluation according to the statistical parameters MBE (mean bias error), RMSE (root mean square error) and IA (index of agreement) for anthesis dates of winter wheat, spring barley, silage maize and winter oilseed rape

Crop	Models	Anthesis											
		Lednice			Věrovany			Domanínek			All sites		
		MBE (days)	RMSE (days)	IA	MBE (days)	RMSE (days)	IA	MBE (days)	RMSE (days)	IA	MBE (days)	RMSE (days)	IA
Winter wheat	AGROTOOL	−6.1	7.1	0.61	−3.1	5.1	0.78	3.6	5.7	0.82	−2.3	6.1	0.89
	APSIM	2.2	4.6	0.81	5.3	6.2	0.75	10.6	11.5	0.52	5.6	7.6	0.84
	AQUACROP	−0.5	3.6	0.82	5.0	7.3	0.69	1.2	5.3	0.86	1.9	5.6	0.88
	CROPSYST	−1.6	3.0	0.88	0.1	2.3	0.94	2.1	3.6	0.85	0.0	2.9	0.96
	DAISY_1	0.6	2.4	0.92	2.3	3.1	0.91	2.5	3.2	0.91	1.7	2.9	0.96
	DAISY_2	−4.1	7.0	0.62	−1.8	7.9	0.52	2.6	4.7	0.76	−1.4	6.8	0.85
	DSSAT_1	−2.1	5.5	0.79	2.6	5.4	0.84	12.6	12.9	0.49	3.7	8.3	0.85
	DSSAT_2	−1.4	3.6	0.80	0.6	2.4	0.93	0.6	3.5	0.87	−0.1	3.2	0.95
	DSSAT_3	−5.1	7.3	0.70	−0.4	4.8	0.86	9.2	9.8	0.57	0.6	7.4	0.87
	EPIC	−0.6	4.5	0.83	0.2	3.6	0.90	4.0	5.5	0.79	1.0	4.5	0.93
	FASSET	1.9	3.5	0.83	17.6	18.0	0.37	−11.5	11.9	0.52	3.6	12.6	0.37
	HERMES	1.1	2.9	0.89	2.8	3.8	0.85	4.9	7.6	0.65	2.8	4.9	0.91
	MONICA	−3.2	4.2	0.81	−0.8	2.8	0.92	0.2	3.5	0.87	−1.4	3.5	0.95
	EnsMED	−1.2	2.8	0.90	1.4	2.3	0.94	2.8	3.6	0.88	0.8	2.9	0.97
	EnsAVG	−1.4	2.9	0.90	2.3	2.9	0.92	3.3	3.8	0.87	1.2	3.2	0.96
Spring barley	AGROTOOL	4.6	6.4	0.83	5.7	9.1	0.61	13.0	13.1	0.50	7.9	10.0	0.81
	APSIM	−7.4	7.9	0.78	−6.1	7.8	0.75	0.6	1.4	0.99	−4.2	6.4	0.92
	AQUACROP	2.9	5.0	0.91	1.1	4.1	0.90	1.0	2.4	0.96	1.6	4.0	0.96
	CROPSYST	−1.8	3.6	0.93	−1.4	5.6	0.79	3.6	4.4	0.91	0.0	4.6	0.95
	DAISY_1	−1.4	4.7	0.85	−2.4	5.6	0.79	−0.3	2.1	0.97	−1.4	4.4	0.94
	DAISY_2	−0.8	4.5	0.87	−1.6	5.1	0.82	1.3	2.5	0.96	−0.3	4.2	0.95
	DSSAT_1	−0.6	4.3	0.88	13.2	14.1	0.51	2.6	3.2	0.93	5.3	8.8	0.81
	DSSAT_2	−3.1	4.9	0.89	−2.8	5.6	0.84	3.6	3.9	0.90	−0.7	4.9	0.95
	DSSAT_3	−1.1	3.3	0.94	−0.9	4.9	0.84	4.2	4.5	0.88	0.8	4.3	0.95
	EPIC	−3.9	4.7	0.89	−3.5	6.1	0.79	1.6	2.4	0.96	−1.9	4.7	0.95
	FASSET	−9.4	10.7	0.62	−2.6	6.2	0.73	−10.4	10.6	0.59	−7.6	9.5	0.77
	HERMES	1.9	5.5	0.82	0.7	5.4	0.80	2.3	4.3	0.90	1.6	5.1	0.93
	MONICA	−2.3	4.1	0.90	−3.3	5.8	0.80	0.5	1.8	0.98	−1.7	4.2	0.95

	EnsMED	−1.4	3.8	0.92	−1.8	5.3	0.81	1.4	2.3	0.96	−0.6	4.0	0.96
	EnsAVG	−1.7	3.8	0.92	−0.2	4.7	0.85	1.8	2.3	0.96	0.0	3.7	0.96
Silage maize	AGROTOOL	0.0	3.6	0.91	−1.0	5.4	0.78	1.9	7.5	0.66	0.2	5.5	0.95
	APSIM	−0.1	3.2	0.90	−0.8	3.4	0.85	3.0	5.6	0.77	0.5	4.1	0.97
	AQUACROP	1.8	3.4	0.90	1.9	3.8	0.84	−0.4	2.1	0.97	1.2	3.2	0.98
	CROPSYST	0.6	3.3	0.92	1.8	3.7	0.88	2.3	5.9	0.96	1.5	4.6	0.96
	DAISY_1	−0.2	2.6	0.93	0.3	2.9	0.84	0.4	3.2	0.93	0.1	2.8	0.98
	DAISY_2	2.2	5.5	0.69	0.5	4.1	0.67	2.0	3.6	0.91	1.6	4.6	0.95
	DSSAT_1	1.5	4.8	0.86	0.5	2.6	0.92	−1.3	4.8	0.84	0.4	4.3	0.96
	DSSAT_2	0.8	4.9	0.84	−1.4	4.0	0.87	−0.1	5.3	0.79	−0.1	4.8	0.96
	DSSAT_3	1.7	4.8	0.84	0.5	2.4	0.94	9.4	11.5	0.65	0.6	3.9	0.97
	EPIC	0.6	2.7	0.94	−2.0	3.1	0.89	−2.7	5.2	0.86	−1.1	3.7	0.97
	FASSET	−6.8	10.7	0.38	8.3	9.1	0.61	5.6	6.7	0.69	−1.6	9.1	0.85
	HERMES	−3.9	5.8	0.82	−3.3	5.0	0.82	22.7	23.2	0.36	−3.0	5.8	0.94
	MONICA	−1.8	3.2	0.93	−2.5	3.5	0.89	2.6	4.7	0.84	−2.7	4.2	0.96
	EnsMED	0.2	2.7	0.94	0.4	2.6	0.92	−0.4	3.8	0.89	0.1	3.0	0.98
	EnsAVG	−0.3	2.2	0.96	0.2	2.1	0.95	−0.5	3.7	0.91	−0.2	2.7	0.99
Winter oil seed rape	*AGROTOOL*	20.3	20.6	0.18	–	–	–	41.3	41.4	0.24	31.6	33.4	0.38
	APSIM	−0.7	5.2	0.62	–	–	–	12.1	15.1	0.39	6.2	11.6	0.69
	AQUACROP	−0.2	1.1	0.94	–	–	–	1.7	7.8	0.83	0.8	5.8	0.90
	CROPSYST	16.8	17.3	0.23	–	–	–	25.6	25.9	0.32	21.5	22.3	0.47
	DAISY_1	19.4	19.7	0.19	–	–	–	22.7	23.0	0.35	21.2	21.5	0.46
	DAISY_2	4.6	6.5	0.52	–	–	–	17.0	17.3	0.44	11.3	13.4	0.67
	DSSAT_1	−21.6	25.0	0.15	–	–	–	6.3	9.0	0.74	−6.6	18.2	0.60
	DSSAT_2	−12.6	16.4	0.24	–	–	–	11.0	12.5	0.60	0.1	14.5	0.69
	DSSAT_3	−18.6	23.0	0.17	–	–	–	9.4	11.5	0.65	−3.5	17.8	0.62
	EPIC	7.4	9.4	0.35	–	–	–	11.4	12.1	0.55	9.6	10.9	0.71
	FASSET	0.3	3.0	0.80	–	–	–	5.6	6.7	0.69	3.1	5.3	0.89
	HERMES	19.1	19.6	0.19	–	–	–	22.7	23.2	0.36	21.0	21.6	0.47
	MONICA	2.6	3.4	0.72	–	–	–	2.6	4.7	0.84	2.6	4.1	0.91
	EnsMED	3.1	4.4	0.65	–	–	–	11.7	12.7	0.53	7.7	9.8	0.76
	EnsAVG	2.8	5.4	0.55	–	–	–	14.6	14.8	0.50	9.2	11.5	0.73

The evaluation involves individual experimental sites Lednice, Věrovany, Domanínek (abb. LED, VER, and DOM), respectively, and results from all the sites together. The medians (abb. EnsMED) and means (abb. EnsAVG) derived from all the model results were compared with observed data as well. Winter oilseed rape simulations within Věrovany were unavailable. When the RMSE of a model for all sites is higher than the mean RMSE plus 1.5 times the standard deviation for a certain crop, the name of a model is written in italics.
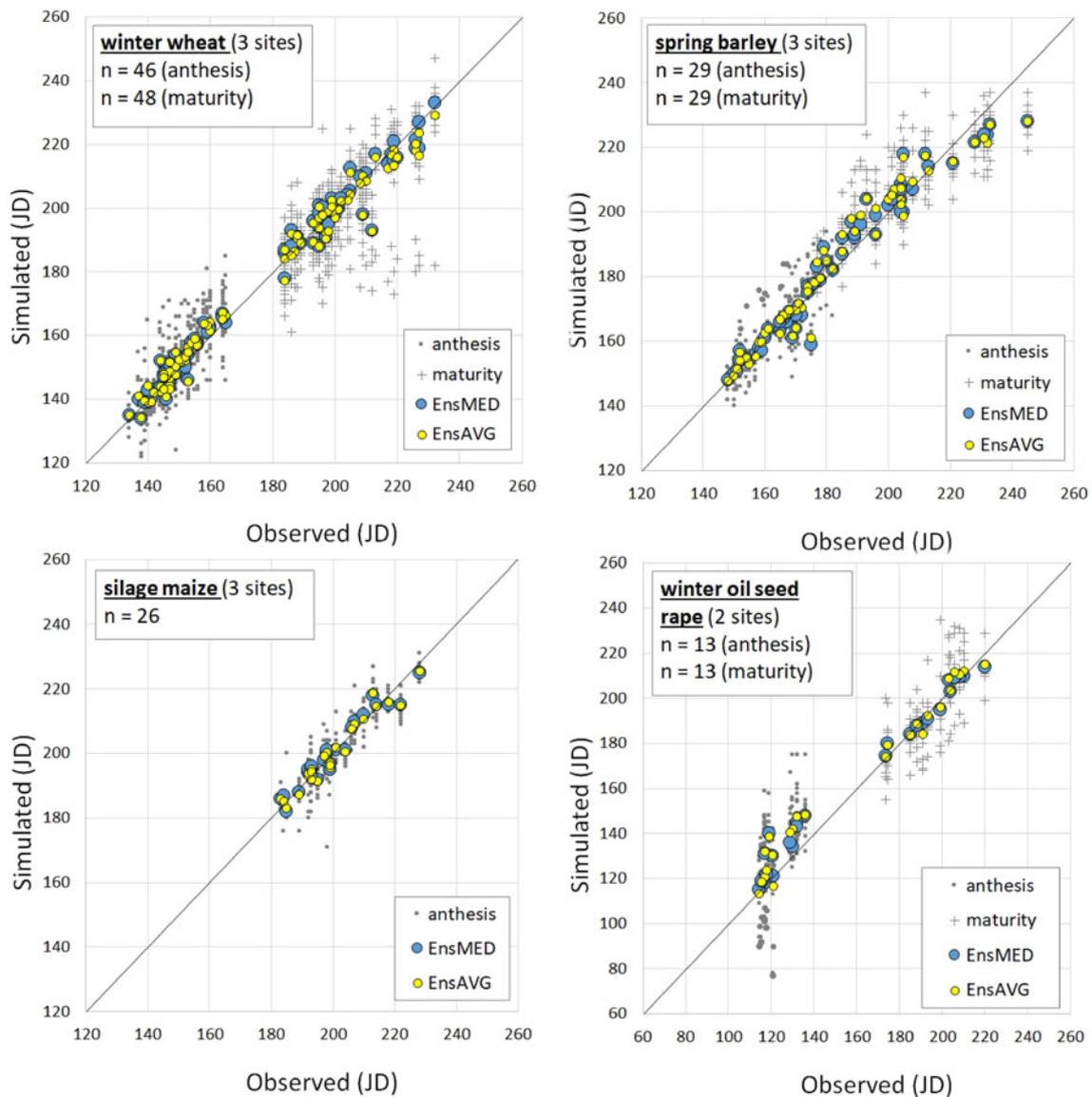
Fig. 2. Comparisons of observed and estimated anthesis (circles) and maturity (crosses) in days of the year (JD). The results based on individual model simulations are depicted in grey (or are apparent from Supplements 1–7), while the medians (EnsMED) and means (EnsAVG) from all the models for specific site-year combinations are in blue and yellow, respectively. Maturity for silage maize was not considered.

the AGROTOOL model. Using EnsMED and EnsAVG, the IAs were 0.95 and 0.94, the MBEs were 1.1 and 1.7 days, and the RMSEs were 6.5 and 7.1 days, respectively (considering all the sites); this indicates slightly lower accuracy, as in the case of anthesis. A majority of the models and, consequently, the ensemble tools had problems reproducing very late spring barley maturity dates in the case of Domanínek (see Fig. 2 and Supplement 6).

In contrast to anthesis, winter oilseed rape maturity was better estimated by the models. For CROPSYST, the IA was 0.98, and for DAISY_1, DAISY_2, EPIC, HERMES and MONICA, the IAs was equal to 0.97 considering all the sites together. The other models IA ranged from 0.62 to 0.93. For most of the models, the MBEs varied from −6.0 to 3.0 days, and the RMSEs from 3.3 to 9.3 days (considering all the sites). AGROTOOL, APSIM and FASSET presented systematic errors ranged from −18.9 to 18.8 days, and absolute errors ranged

from 14.5 to 20.4 days. Using EnsMED and EnsAVG, the IAs of 0.98 for both methods were achieved with MBEs of 0.0 and 0.3 days and RMSEs from 3.2 to 3.7 days, respectively. The observed date of maturity for maize was unavailable because it was harvested for silage.

Yield

The yields were simulated with considerably lower accuracy (compared to phenology) by both the models and the ensemble methods, as is apparent from Fig. 3 and Table 6. For the winter wheat yield, the models reached IAs ranging from 0.24 to 0.60 considering all sites together, where the highest value was recorded for the AQUACROP model and the lowest IA by MONICA (0.24). The MBEs varied from −1274 kg/ha (AGROTOOL) to 1057 kg/ha (DAISY_2), whereas the lowest MBE was achieved by CROPSYST at a level of −4 kg/ha.

Table 5. The evaluation of models according to the statistical parameters MBE (mean bias error), RMSE (root mean square error) and IA (index of agreement) for the maturity of winter wheat, spring barley and winter oilseed rape

| | | Maturity | | | | | | | | | | | |
| | | Lednice | | | Věrovany | | | Domanínek | | | All sites | | |
Crop	Models	MBE (days)	RMSE (days)	IA	MBE (days)	RMSE (days)	IA	MBE (days)	RMSE (days)	IA	MBE (days)	RMSE (days)	IA
Winter wheat	AGROTOOL	2.0	4.5	0.85	−0.6	5.2	0.87	6.0	11.5	0.52	2.4	7.6	0.92
	APSIM	4.0	6.3	0.78	−6.8	53.7	0.65	6.5	8.2	0.81	5.9	8.0	0.91
	AQUACROP	−3.8	8.1	0.68	−12.8	14.4	0.56	−4.3	8.9	0.83	−7.0	10.8	0.86
	CROPSYST	0.8	4.0	0.87	1.6	5.0	0.83	−2.3	4.6	0.84	0.3	4.6	0.96
	DAISY_1	0.5	3.1	0.93	0.8	4.8	0.86	2.1	3.6	0.96	1.1	3.9	0.98
	DAISY_2	−3.4	6.6	0.77	−0.3	11.8	0.30	0.2	4.8	0.93	−1.2	8.3	0.90
	DSSAT_1	−8.9	10.8	0.61	−5.8	8.6	0.74	−3.0	5.5	0.90	−6.0	8.6	0.91
	DSSAT_2	0.7	4.3	0.85	1.9	5.4	0.78	−1.9	5.5	0.87	0.3	5.1	0.95
	DSSAT_3	−3.9	7.2	0.75	−0.8	6.3	0.82	2.4	5.1	0.91	−0.9	6.3	0.95
	EPIC	3.4	5.6	0.83	2.5	6.8	0.76	1.3	4.6	0.93	2.5	5.8	0.95
	FASSET	−12.7	13.4	0.52	4.9	10.0	0.54	−36.6	37.5	0.31	−16.5	24.2	0.42
	HERMES	−2.5	4.9	0.86	−0.6	5.7	0.81	−1.3	4.7	0.93	−1.5	5.1	0.96
	MONICA	−2.8	4.9	0.85	−0.8	6.0	0.78	−2.4	5.3	0.91	−2.0	5.4	0.95
	EnsMED	−1.6	4.6	0.86	0.5	5.5	0.80	−0.6	4.3	0.93	−0.6	4.8	0.96
	EnsAVG	−2.0	4.7	0.86	−0.4	5.4	0.82	−2.5	4.8	0.91	−1.6	5.0	0.96
Spring barley	*AGROTOOL*	12.3	14.3	0.59	14.3	15.3	0.49	9.1	13.5	0.64	11.9	14.4	0.81
	APSIM	2.2	5.1	0.91	3.0	5.0	0.86	−1.0	8.1	0.81	1.4	6.3	0.96
	AQUACROP	5.7	8.1	0.78	−1.0	2.8	0.94	−3.2	7.8	0.85	0.3	6.6	0.95
	CROPSYST	1.3	5.7	0.87	1.1	2.9	0.95	−2.3	8.3	0.85	0.2	5.9	0.96
	DAISY_1	3.1	5.4	0.88	0.6	2.8	0.94	−8.1	12.0	0.69	−1.6	7.8	0.92
	DAISY_2	5.0	7.3	0.79	1.9	3.7	0.90	−7.4	11.9	0.68	−0.3	8.4	0.90
	DSSAT_1	4.4	9.1	0.70	19.5	20.3	0.41	−4.8	9.2	0.78	6.4	14.0	0.77
	DSSAT_2	2.1	5.2	0.90	2.1	4.7	0.87	−2.4	8.7	0.79	0.6	6.5	0.95
	DSSAT_3	2.3	5.9	0.86	2.2	3.7	0.90	−3.5	9.2	0.76	0.3	6.7	0.95
	EPIC	6.7	8.4	0.77	5.6	6.5	0.80	1.8	8.7	0.79	4.6	7.9	0.93
	FASSET	−4.3	7.5	0.81	7.8	9.3	0.48	−14.0	16.1	0.63	−5.8	12.1	0.83
	HERMES	3.3	7.9	0.74	3.8	6.1	0.78	−3.3	8.1	0.84	1.2	7.4	0.93
	MONICA	2.8	6.3	0.84	1.6	3.8	0.89	−2.4	8.5	0.80	0.5	6.6	0.95

(Continued)

Table 5. (*Continued.*)

		Maturity												
		Lednice			Věrovany			Domanínek			All sites		
Crop	Models	MBE (days)	RMSE (days)	IA	MBE (days)	RMSE (days)	IA	MBE (days)	RMSE (days)	IA	MBE (days)	RMSE (days)	IA
	EnsMED	3.2	5.9	0.87	3.1	4.8	0.85	−2.8	8.6	0.79	1.1	6.5	0.95
	EnsAVG	3.6	6.5	0.84	4.9	5.9	0.80	−3.3	8.7	0.79	1.7	7.1	0.94
Winter oil seed rape	AGROTOOL	3.4	5.3	0.86	–	–	–	18.1	19.2	0.36	11.3	14.5	0.82
	APSIM	18.2	19.5	0.46	–	–	–	19.3	21.1	0.31	18.8	20.4	0.62
	AQUACROP	3.1	5.6	0.79	–	–	–	−4.0	8.2	0.76	−0.7	7.1	0.91
	CROPSYST	−1.1	4.0	0.91	–	–	–	−0.3	2.6	0.94	−0.7	3.3	0.98
	DAISY_1	−0.8	3.3	0.93	–	–	–	−0.6	5.1	0.79	−0.7	4.3	0.97
	DAISY_2	−5.8	6.8	0.83	–	–	–	0.0	3.3	0.92	−2.7	5.2	0.97
	DSSAT_1	−9.3	11.4	0.66	–	–	–	−3.3	7.1	0.74	−6.0	9.3	0.90
	DSSAT_2	−3.9	7.9	0.77	–	–	–	1.3	7.3	0.72	−1.1	7.6	0.93
	DSSAT_3	−3.6	8.1	0.77	–	–	–	4.4	8.5	0.66	0.7	8.3	0.92
	EPIC	5.6	6.1	0.85	–	–	–	0.7	2.1	0.97	3.0	4.4	0.97
	FASSET	−17.6	18.0	0.51	–	–	–	−20.0	20.2	0.39	−18.9	19.2	0.68
	HERMES	1.6	4.2	0.90	–	–	–	−1.9	3.9	0.90	−0.3	4.1	0.97
	MONICA	2.3	3.8	0.92	–	–	–	−1.0	3.2	0.94	0.5	3.5	0.97
	EnsMED	0.1	2.6	0.96	–	–	–	−0.2	3.6	0.90	0.0	3.2	0.98
	EnsAVG	−0.6	3.5	0.93	–	–	–	1.1	3.9	0.89	0.3	3.7	0.98

The evaluation involved individual experimental sites Lednice, Věrovany and Domanínek (abb. LED, VER and DOM, respectively) and the results from all the sites together. The medians (abb. EnsMED) and means (abb. EnsAVG) derived from all the model results were compared with observed data as well. Winter oilseed rape simulations for the Věrovany site were unavailable. When the RMSE of a model for all sites is higher than the mean RMSE plus 1.5 times the standard deviation for a certain crop, the name of a model is written in italics.
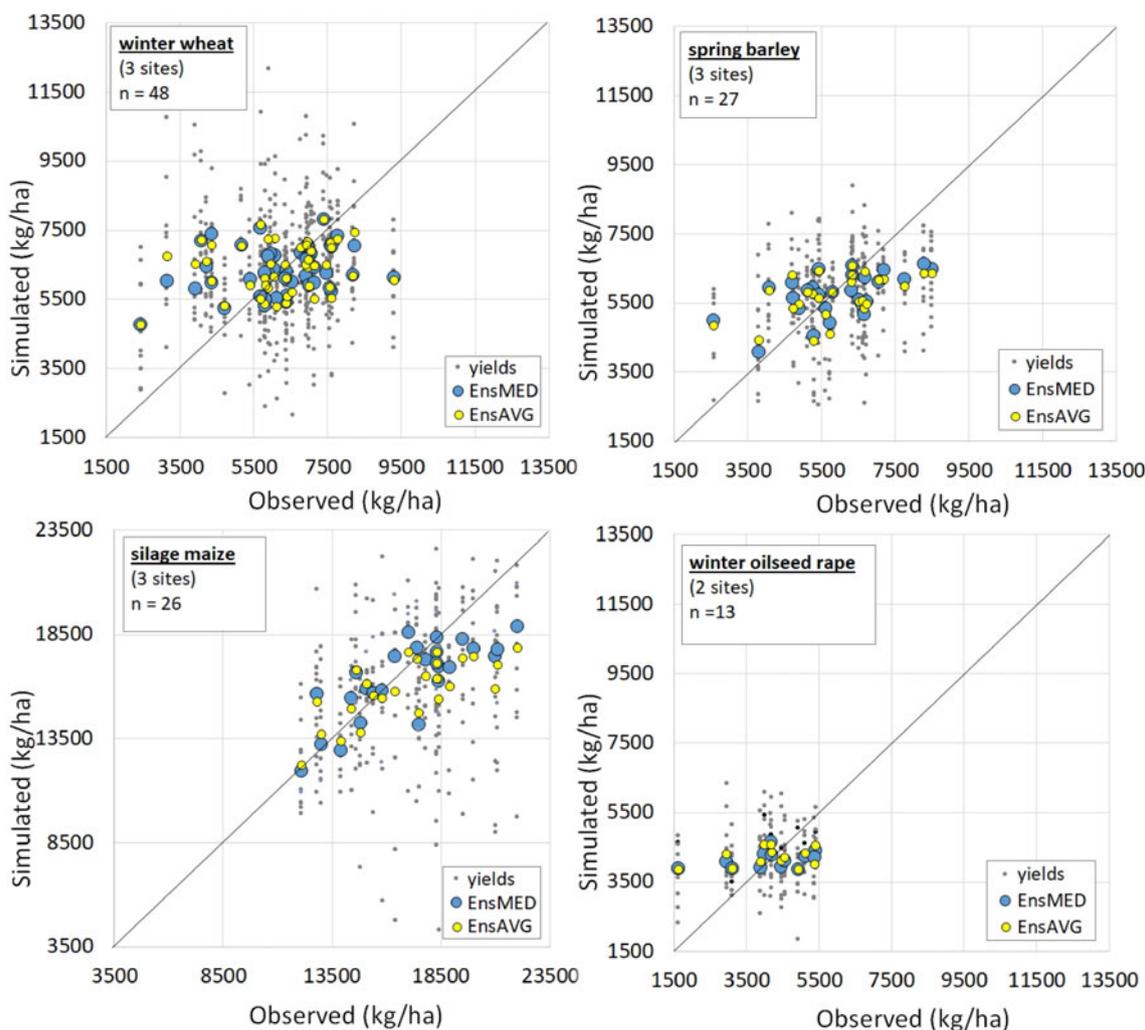
Fig. 3. Comparisons of observed and simulated yield (in kg/ha) within 13 crop growth models for winter wheat, spring barley, silage maize and winter oilseed rape. The results based on individual model simulations are depicted in grey (or are apparent from Supplements 8–11), while the medians (EnsMED) and means (EnsAVG) from all the models for specific site-year combinations are in blue and yellow, respectively.

The RMSEs varied from 1326 kg/ha (AQUACROP) to 2828 kg/ha (DAISY_2). Six out of all 13 models achieved greater or equal IAs with respect to EnsMED (IA = 0.48). The IA for EnsAVG was 0.46. Thus, from the perspective of MBE and RMSE, EnsMED and EnsAVG still performed reasonably and are comparable to the best models (MBEs of 88 and 141 kg/ha, RMSEs of 1365 and 1439 kg/ha for EnsMED and EnsAVG, respectively). Although the individual runs resulted in highly variable winter wheat yield outputs (∼10 tons/ha based on the difference between the minimal and maximal values of all runs) compared to the observed variability (about 7 t/ha), using EnsMED and EnsAVG for the seasonal yield resulted in an even lower variability of ∼3 t/ha. This feature is similar for all crops, as can be seen in Fig. 4, where a comparison of the observed yield variability with each model and its ensemble (both as EnsAVG and EnsMED) can be seen. This figure shows that the higher variability of simulated yields (when considering all runs together) is mainly due to the differences between individual models than to the variability within the outputs of the single models.

For spring barley yield, the models achieved better estimates than for winter wheat (Table 6); the AQUACROP model obtained the highest IA (0.87), with an MBE equal to −151 kg/ha and an

RMSE equal to 807 kg/ha when all the sites are considered. The lowest IA was observed for the MONICA model, i.e. 0.32. For the other models, the IAs varied from 0.36 to 0.74 for all the sites. Three models (AQUACROP, EPIC, CROPSYST) achieved better results for annual yield estimates with respect to using EnsMED, which had an IA of 0.64. The IA for EnsAVG was 0.61. The ensemble MBEs were −148 and −190 kg/ha, the RMSEs ranged from 1105 to 1136 kg/ha.

For silage maize yield, large differences also existed among the models, as is apparent from Table 6. Specifically, the AQUACROP model provided the highest value of IA, which was equal to 0.93 if all the sites were considered together. The other models had IAs ranging from 0.32 (AGROTOOL) to 0.84 (CROPSYST), MBE varied from −6784 to 716 kg/ha, and the RMSEs varied from 1355 to 7674 kg/ha. Among the individual localities, the AQUACROP model simulated Domanínek very accurately (IA = 0.96; MBE = 216 kg/ha, and RMSE = 1325 kg/ha). On the other hand, almost all the models underestimated the observed highest yield (Fig. 3) from the most fertile station, Věrovany. Based on IA, two of the models were better than EnsMED, and seven of the models were better than EnsAVG. Based on the IA, the ensemble models were able to most

Table 6. Model evaluation according to the statistical parameters MBE (mean bias error), RMSE (root mean square error) and IA (index of agreement) for yields of winter wheat, spring barley, silage maize and winter oilseed rape

| | | Yields | | | | | | | | | | | |
| | | Lednice | | | Věrovany | | | Domanínek | | | All sites | | |
Crop	Models	MBE (kg/ha)	RMSE (kg/ha)	IA	MBE (kg/ha)	RMSE (kg/ha)	IA	MBE (kg/ha)	RMSE (kg/ha)	IA	MBE (kg/ha)	RMSE (kg/ha)	IA
Winter wheat	AGROTOOL	−1064	1837	0.54	−2001	2304	0.39	−707	1719	0.28	−1274	1977	0.42
	APSIM	498	1486	0.72	251	1462	0.37	568	1652	0.25	435	1534	0.55
	AQUACROP	326	1321	0.77	−83	1178	0.44	683	1473	0.33	300	1326	0.60
	CROPSYST	−501	1510	0.70	−396	1376	0.26	1140	1868	0.33	−4	1573	0.48
	DAISY_1	573	1888	0.58	−1110	2030	0.55	1949	2722	0.30	436	2239	0.40
	DAISY_2	791	2364	0.64	59	1614	0.49	2322	3975	0.16	1057	2828	0.39
	DSSAT_1	−215	1569	0.66	−702	1366	0.46	187	1623	0.34	−253	1520	0.52
	DSSAT_2	1376	2679	0.33	−66	1577	0.48	−360	2152	0.20	308	2172	0.39
	DSSAT_3	284	2133	0.46	314	1467	0.45	−129	1592	0.37	160	1749	0.50
	EPIC	−309	1708	0.63	22	914	0.57	835	1728	0.42	179	1488	0.55
	FASSET	3	1737	0.30	−205	1361	0.55	−71	1369	0.27	−93	1497	0.46
	HERMES	1086	2144	0.36	292	1197	0.53	398	2428	0.09	585	1980	0.34
	MONICA	−195	1848	0.29	−375	1088	0.50	869	2004	0.20	90	1684	0.24
	EnsMED	41	1542	0.57	−230	1068	0.52	474	1455	0.21	88	1365	0.48
	EnsAVG	204	1544	0.59	−318	1095	0.53	567	1640	0.20	141	1439	0.46
Spring barley	*AGROTOOL*	109	959	0.56	−2596	2789	0.40	−483	1330	0.66	−1068	1926	0.44
	APSIM	807	1234	0.49	−325	986	0.66	921	1430	0.58	422	1214	0.62
	AQUACROP	−158	467	0.95	−436	786	0.85	215	1086	0.57	−151	807	0.87
	CROPSYST	303	769	0.84	−251	1217	0.46	1698	2245	0.49	444	1427	0.65
	DAISY_1	−1171	1616	0.46	−1506	1897	0.41	−631	1529	0.39	−1135	1702	0.52
	DAISY_2	−389	1560	0.27	−1953	2494	0.37	285	1065	0.65	−769	1858	0.36
	DSSAT_1	154	1592	0.26	−786	1473	0.42	881	1508	0.39	21	1524	0.37
	DSSAT_2	569	1851	0.12	−809	1276	0.67	921	1516	0.65	163	1557	0.55
	DSSAT_3	−89	1695	0.23	−1199	1465	0.65	127	890	0.82	−436	1410	0.62
	EPIC	726	1190	0.58	−474	943	0.67	−62	1042	0.71	48	1060	0.74
	FASSET	136	1157	0.41	−1410	2243	0.47	361	1495	0.59	−370	1724	0.47
	HERMES	713	1079	0.66	−404	996	0.64	879	1733	0.50	348	1283	0.63
	MONICA	629	1304	0.32	−1217	1475	0.59	1028	1482	0.55	63	1422	0.32

	EnsMED	242	1027	0.46	−959	1240	0.63	428	1006	0.70	−148	1105	0.64
	EnsAVG	180	1014	0.53	−1033	1317	0.61	448	1015	0.68	−190	1136	0.61
Silage maize	*AGROTOOL*	−5346	5688	0.31	−10 894	11 040	0.26	−4346	5349	0.48	−6784	7674	0.32
	APSIM	−2294	2663	0.55	−3749	4359	0.38	−2434	3876	0.52	−2780	3591	0.57
	AQUACROP	−474	993	0.91	−882	1753	0.77	216	1325	0.96	−414	1355	0.93
	CROPSYST	−910	2199	0.71	−643	2104	0.74	−249	1896	0.87	−580	2083	0.84
	DAISY_1	−400	1780	0.65	−1039	1802	0.66	2340	3259	0.71	141	2280	0.68
	DAISY_2	−720	1985	0.78	−6117	6577	0.33	973	2660	0.78	−1925	4109	0.47
	DSSAT_1	23	1554	0.82	−610	1923	0.76	−478	3505	0.51	−307	2338	0.77
	DSSAT_2	754	1757	0.64	−540	2874	0.46	243	2649	0.72	218	2395	0.67
	DSSAT_3	479	2499	0.58	829	2662	0.47	959	3091	0.79	716	2719	0.75
	EPIC	406	2104	0.68	−1620	2802	0.66	−929	2515	0.79	−577	2448	0.77
	FASSET	−2137	3335	0.30	−873	4200	0.27	1478	2822	0.69	−775	3504	0.48
	HERMES	379	2121	0.61	−2883	3930	0.52	−1754	3361	0.55	−1199	3114	0.63
	MONICA	−56	2710	0.58	−1542	2760	0.38	3479	4510	0.57	100	3305	0.48
	EnsMED	−830	1503	0.80	−1756	2394	0.58	−201	1661	0.91	−946	1861	0.83
	EnsAVG	−1818	2202	0.66	−3298	3624	0.51	−936	2378	0.75	−2036	2760	0.67
Winter oil seed rape	AGROTOOL	219	953	0.77	–	–	–	1325	1640	0.35	815	1366	0.53
	APSIM	−62	1416	0.08	–	–	–	−861	1211	0.37	−492	1310	0.29
	AQUACROP	158	1329	0.40	–	–	–	−435	922	0.46	−161	1128	0.46
	CROPSYST	22	1402	0.35	–	–	–	843	1446	0.32	464	1426	0.31
	DAISY_1	−163	1340	0.32	–	–	–	796	979	0.61	353	1159	0.44
	DAISY_2	−1148	1645	0.56	–	–	–	23	673	0.52	−518	1221	0.54
	DSSAT_1	−230	1275	0.32	–	–	–	255	1003	0.24	31	1137	0.27
	DSSAT_2	1074	1567	0.46	–	–	–	−731	1254	0.38	102	1407	0.47
	DSSAT_3	225	1778	0.00	–	–	–	−472	1134	0.40	−151	1466	0.20
	EPIC	40	731	0.84	–	–	–	103	708	0.39	74	719	0.74
	FASSET	58	1134	0.40	–	–	–	127	989	0.23	95	1058	0.32
	HERMES	713	1422	0.38	–	–	–	7	895	0.40	333	1168	0.42
	MONICA	−230	1180	0.41	–	–	–	244	947	0.16	25	1061	0.29
	EnsMED	−50	1170	0.35	–	–	–	39	754	0.31	−2	969	0.33
	EnsAVG	52	1150	0.41	–	–	–	94	823	0.21	75	987	0.35

The evaluation involved the individual Lednice, Věrovany and Domanínek (abb. LED, VER and DOM, respectively) experimental sites and results from all the sites together. Median (abb. EnsMED) and mean (abb. EnsAVG) derived from all the model results were compared with observed data as well. When the RMSE of a model for all sites is higher than the mean RMSE plus 1.5 times the standard deviation for a certain crop, the name of a model is written in italics.
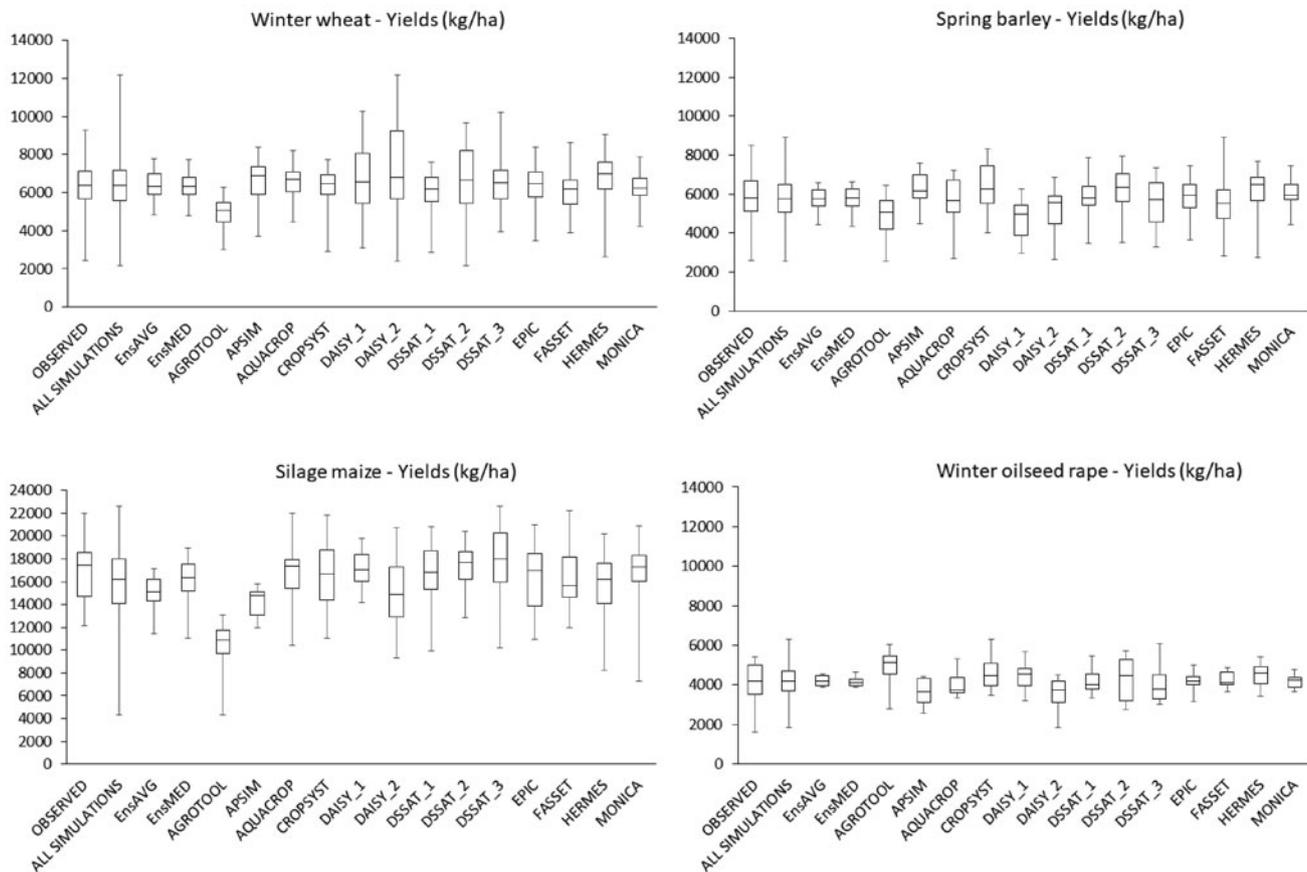
Fig. 4. Boxplots for values of observed and estimated yields by the models and on the ensemble basis. The boxes delimit the medians and interquartile ranges (25–75 percentiles), and the whiskers link the high and low extreme values.

accurately simulate (compared to all the tested crops) silage maize yield.

The winter oilseed rape yield results are shown in Table 6 and Figs 3 and 4. Across all the sites, the highest value of IA was recorded for the EPIC model (0.74). On the other hand, the DSSAT_3 model had the lowest overall value of IA (0.20). Among the individual sites, the Lednice experimental site, especially when using the EPIC model, achieved the highest value of IA (0.84). The MBE and RMSE for the EPIC model and Lednice were 40 and 731 kg/ha, respectively. Generally, MBE and RMSE, based on individual models, varied from −518 to 815 kg/ha and from 719 to 1466 kg/ha, respectively (all sites considered together). Although the variabilities among the models are clear, EnsMED and EnsAVG produced quite narrow and low variable estimates close to 4 t/ha on average (see Fig. 4 for winter oilseed rape). Based on IA, the lowest agreements between the simulated (both by individual models and ensemble products) and observed yield were achieved for winter oilseed rape. For winter wheat, spring barley and silage maize, EnsMED produced more accurate yield estimates than EnsAVG (based on MBE, RMSE and IA). For winter oilseed rape, a lower IA for EnsMED was achieved (against EnsAVG) but also with lower MBE and RMSE.

Table 7 depicts a general overview of the observed and simulated yield of the four crops, including the ranges of errors, indexes of agreement and best performing simulations/models (based on all the statistical parameters individually). Overall and considering RMSE and IA, AQUACROP performed best for winter wheat, spring barley and silage maize, whereas EPIC performed best for winter oilseed rape. Although these models performed well overall, individual models outperformed the obtained results for specific locations or different statistical metrics. For example, considering MBE, CROPSYST performed best overall for winter wheat, DSSAT1 performed best for spring barley and MONICA performed best for silage maize and winter oilseed rape.

Figure 5 depicts the arrangement of models in order of their accuracies (considering average order based on RMSE and IA) for phenology, yields and their combination. The most successful models are AQUACROP and EPIC, but the best predictions were based on ensemble products. Considering all parameters and crops, EnsMED achieved the best results according to this evaluation. This was true at each individual station (so regardless of the climatic conditions) as well as considering all stations together. The EnsAVG indicator was in the second place if all stations are considered together, and in the individual stations, it was twice in the second place (Lednice, Věrovany) and once on the fourth place (Domanínek). CROPSYST is fine overall but not for estimated winter oilseed rape yield, low leaf area and high harvest index (see Figs 6 and 7). The AGROTOOL model resulted in less accurate results in general, and FASSET experienced some difficulties when reproducing crop phenology. For some models (AQUACROP, CROPSYST, DSSAT_3) a different behaviour (lower order) was found in the case of winter oilseed rape yields compared to the remaining crops in the case of the achieved order of models (Fig. 5b).

Table 7. Observed and simulated ranges of yields for all the sites and seasons using mean and ± standard deviation and agreements between simulated and observed yields using MBE, RMSE and IA

	Field crops yields			
	Winter wheat	Spring barley	Silage maize	Winter oilseed rape
Observed (kg/ha)	6117 ± 1371	5902 ± 1316	17 008 ± 2626	4134 ± 1035
Simulated (kg/ha)	6393 ± 1375	5703 ± 1137	15 918 ± 3081	4066 ± 844
MBE range (kg/ha)	−1274 to 1057	−1135 to 444	−6784 to 716	−518 to 815
Best MBE (kg/ha)	−4 (CS)	21 (DS1)	100 (MO)	25 (MO)
RMSE range (kg/ha)	1326 to 2828	807 to 1926	1355 to 7674	719 to 1466
The lowest RMSE	(AQ)	(AQ)	(AQ)	(EP)
IA	0.24 to 0.60	0.32 to 0.87	0.32 to 0.93	0.20 to 0.74
The highest IA	(AQ)	(AQ)	(AQ)	(EP)

The most precise models are indicated using abbreviations within brackets
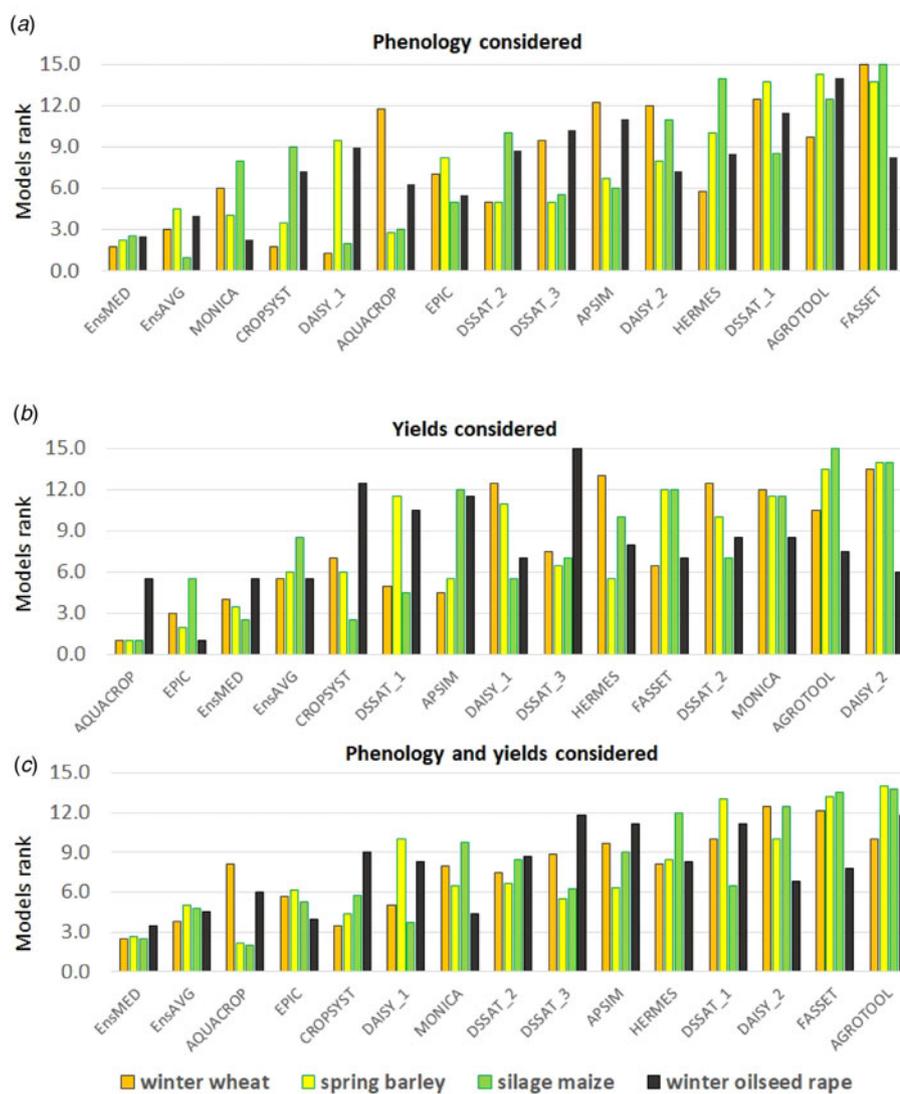
Fig. 5. Arrangement of the crop model tools based on RMSE and IA for phenology (a), yield (b) and combined for both phenology and yields (c) through the tested four crops and within the three experimental sites.

Although the study is not based on data from crop rotation experiments, the 112 seasons (sum through all crops) with observed yield bring the opportunity to evaluate the set of crops and models under uniform methodology. In the current study, IA was higher when simulating the development stages than when simulating yield. In the case of anthesis, the timing was
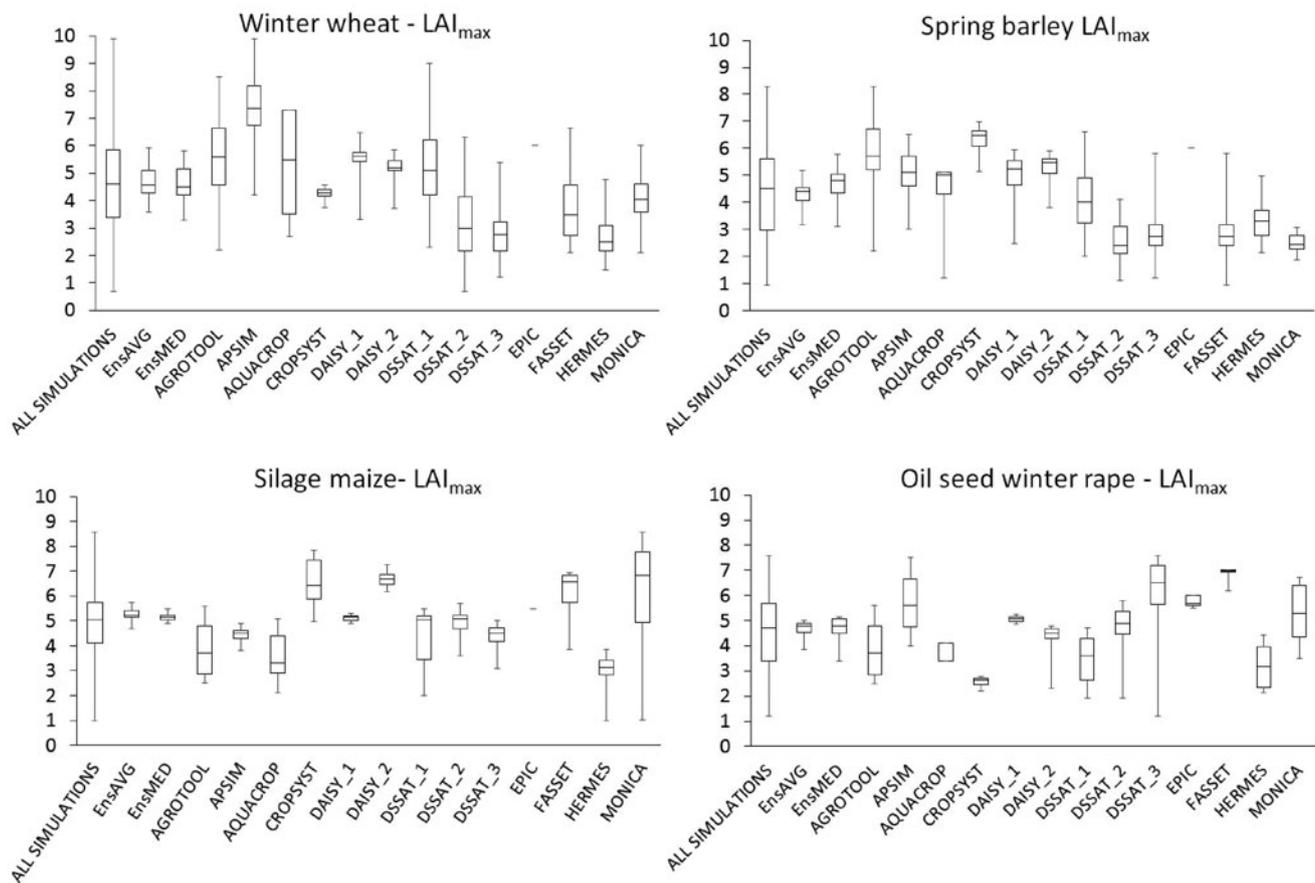
Fig. 6. Boxplots for maximum values of leaf area indexes during the season (LAImax) by the models. The comparisons include medians from single models, processed data from all the simulations, and averages and medians of the model ensemble for all the crops. The boxes delimit the interquartile ranges (25–75 percentiles), and the whiskers link the high and low extreme values.

simulated at similar levels of accuracy for winter wheat, spring barley and silage maize but less satisfactory for winter oilseed rape. On the other hand, IA for the maturity of winter oilseed rape was higher. In the cases of winter wheat and spring barley, the results were worse for maturity compared to anthesis. For most of the crops, no significant feature was identified between simulations, except at the Lednice station (warmer and drier), where the models' performances were better for the anthesis of winter oilseed rape when compared to those for Domanínek (colder and wetter).

Discussion

Crop yield prediction is at a less accurate level compared to phenology. One of the reasons is that the yield formation process is naturally more complex than phenology, as soil conditions and processes as well as biomass production and allocation, including root growth, which has large effects on water and nutrient availability, are estimated. One explanation for weaker model performances in simulating interannual and site variabilities could be related to the fact that crops within the experimental sites were not always grown on exactly the same plots over the years. Although the representative soil profiles for each trial site in the current study used estimated initial conditions as inputs to the crop growth models, slight deviations in the defined soil properties between years could affect the results. Moreover, spatial

variations of soil conditions could exist within plots affecting the representation of soil model inputs. These factors can be one source of uncertainty, which may partly explain the spread of model outputs and correspondence with observed values. At the same time, however, it can be said, that the quality of the soil input data can be considered as completely appropriate to the concept of minimum calibration, also with respect that they are used by CISTA for description of field experiment conditions. Furthermore, although some models responded very sensitively to differences in site conditions, e.g. soil, water and nitrogen supply, others are less sensitive (Zhao *et al.*, 2016; Wallor *et al.*, 2018). Another point is that yield extremes, especially at the lower end, are often caused by environmental conditions that most models have no available algorithm to represent their impact. For example, the influences of pests, diseases and other adverse conditions such as lodging, harvest losses and damage caused by rodents (e.g. Nendel *et al.*, 2013; Gobin, 2018) were not directly considered by the crop models. To understand some model overestimations, modellers need to have a good description or database of how observed yield decreases can be explained. No information or evidence was available in the study database on the actual occurrence of phenomena such as lodging, hail, pest and diseases, etc. which could cause some reduction in the observed yield.

Considering yield combined for all sites, the best models (e.g. AQUACROP, EPIC) produced higher IAs than EnsMED and
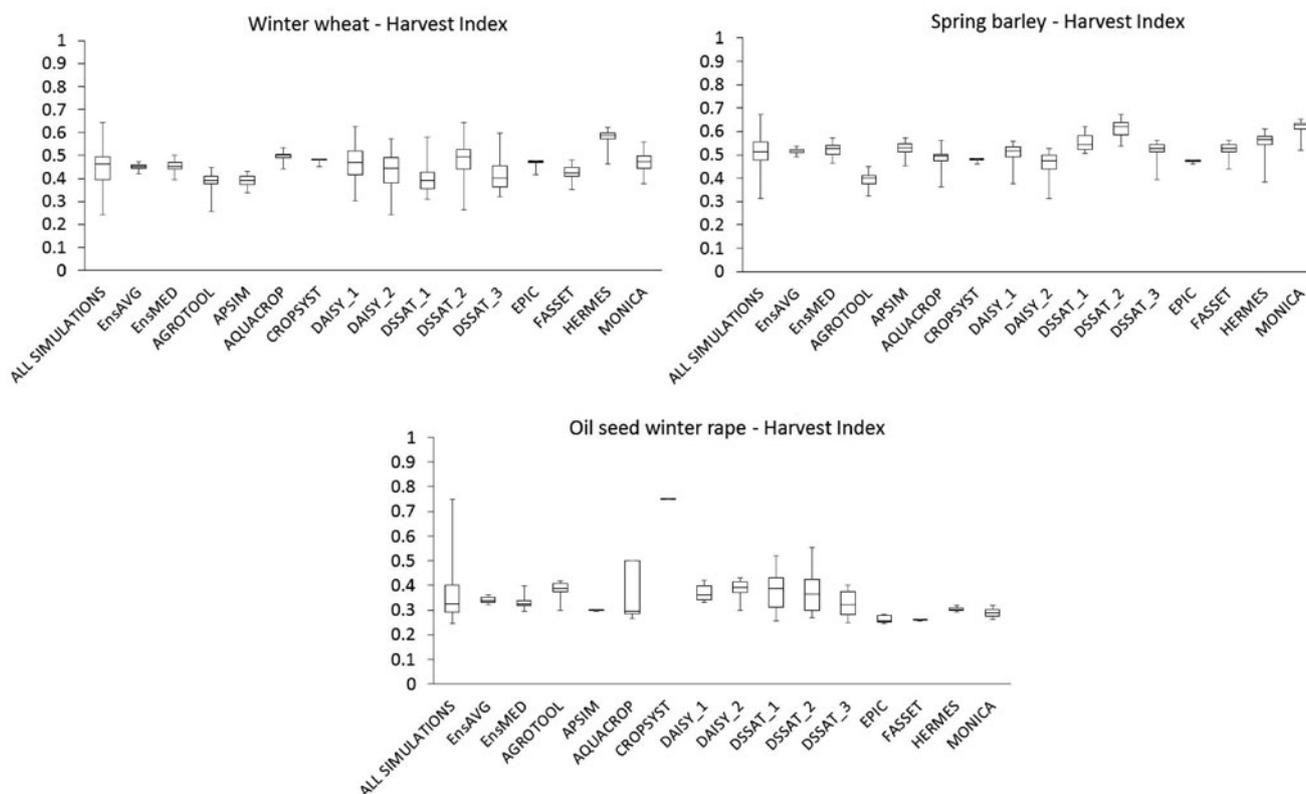
Fig. 7. Boxplots for values of harvest indexes within the employed models. The comparisons include medians from the single models, processed data from all the simulations, and the averages and medians of the model ensemble for all the crops. The boxes delimit the interquartile ranges (25–75 percentiles), and the whiskers link the high and low extreme values.

EnsAVG. This is in agreement with Soltani and Sinclair (2015), as simpler models tend to be more robust. The better results from some individual models (against the ensemble outputs) are not in line with the results of previous studies (e.g. Asseng *et al.*, 2015), which concluded that EnsMED is more accurate than any individual member in simulating the response of crop (as determined from experiments and observations across a wide range of environments). In addition, Martre *et al.* (2015) found both median and mean estimates to be better predictors than any individual model outcomes. However, collecting a high number of members within the ensemble still does assure that certain ensemble outcomes or their interpretation might not change if some members were added or removed from the ensemble (e.g. Rodriguez *et al.*, 2018).

Another point to consider is model accuracy across all the sites together or at each single site where other models may achieve better results. For instance, for winter wheat at Věrovany, EPIC can be ranked first (lowest RMSE and highest IA), whereas AQUACROP's IA is fourth lowest. Therefore, when it comes to selecting the best model(s) or ensemble composition (based on a set of models with the best performance) for specific conditions or range of conditions, the priority between accuracy and robustness (maintaining good performance in different environments) should be resolved. This can be supported by a good calibration and validation procedure, for which detailed and extensive databases are absolutely essential. Especially in the case of more complex models, this plays an important role, where detailed information is crucial for calibration. Based on the results of global large model ensemble studies, pre-selection of models would not be possible due to their complexity and it is not always true

that higher complexity means higher accuracy. For example, a model such as Aquacrop may be preferred when information is limited (Confalonieri *et al.*, 2016).

In addition to the models' characteristics, the variation in modellers' knowledge, experiences, parametrization and subjective approaches for study-specific conditions and target crops could be factors affecting accuracy, as the same models (DAISY_1 v. DAISY_2 and DSSAT_1 v. DSSAT_2 v. DSSAT_3) did not produce the same phenology or yield (Figs 4 and 5). The input data was the same, but the parameter settings differed for each model. Generally, for phenology, lower differences between the same models (except DAISY for winter oilseed rape anthesis) resulted with respect to yield. The differences between the same models are also evident within the maximum estimated values of leaf area index (Fig. 6), especially for DSSAT models. On the other hand, the harvest index results were more stable between the same models as well as through all the models (Fig. 7).

Generally, the highest variability in modelled annual yield was achieved considering individual models. For example, in our study, simulated winter wheat yield ranged from 2164 to 12 185 kg/ha, whereas the observed yield range was narrower, 2442 to 9297 kg/ha, across all the sites. A similar trend was achieved by Palosuo *et al.* (2011), in which the modelled winter wheat yield results ranged from 1800 to 12 000 kg/ha compared against a similar real yield range. Here, the same field experiments with Samanta variety in Lednice and Věrovany were included. The different feature of the mentioned study was that several of their models showed a larger yield range than the observed data. Palosuo *et al.* (2011) also listed a wide range of IA (0.40–0.74)

and RMSE (1400–2300 kg/ha) for winter wheat and similarly Rötter *et al.* (2012) for spring barley (RMSE from 1120 to 1940 kg/ha and IA from 0.31 to 0.63). However, the observed yield was not known to the modellers for the model's setting.

The feature about lower reported variability by EnsMED and EnsAVG is not valid for phenology (Fig. 2), perhaps as an effect of the higher accuracies of individual models in phenology simulations compared to yield, where the errors of individual models are equally reduced. Based on the IAs for winter wheat, spring barley and silage maize, EnsMED produced more accurate estimates than EnsAVG, whereas, in the case of winter oilseed rape, that was reversed. The better performance of EnsMEDs is in accordance with results from previous multimodel ensemble studies and is reflected within the impact and adaptation studies (Pirttioja *et al.*, 2015; Ruiz-Ramos *et al.*, 2018). In the case of an ensemble based on verified and suitable models for target crops, EnsAVG likely provides higher informative value, but when there is uncertainty regarding some of the models, less experience and a smaller number of studies for a crop, EnsAVG is more prone to bias due to the higher impact of a model failure (Wallach *et al.*, 2018).

Using the ensemble, better yield results (IA-based) were achieved for spring crops, and worse results were obtained for winter crops. This is in contrast with Rötter *et al.* (2012) who compared models for spring barley (represented by cultivar Orbit from the Lednice and Věrovany experiments conducted between 1984 and 1998) and concluded, using RMSE and IA, that models performed slightly, but not significantly, better for winter wheat. For instance, the worst results were achieved for yield of winter oilseed rape. This could be explained by the influence of overwintering conditions and modelling of connected impacts (low-temperature stresses, presence and influence of snow cover). Moreover, modellers' experience when simulating winter oilseed rape is usually lower, and the amount of data for calibration as well as the number of studies are much smaller than for main staple crops, e.g. wheat and maize (see e.g. Kollas *et al.*, 2015) as it is obvious from the different number of studies focused on modelling of individual field crops (e.g. Web of Science, Orlandini *et al.*, 2008). The relatively poor performance of most models for oilseed rape yields indicates that the database for many models for parameterizing oilseed rape was not sufficient so far or important crop-specific responses, e.g. frost damage and recovery, were not sufficiently considered in some models. In principle, however, the experience of modellers can influence simulation accuracy for all the crops involved, when the calibration procedure could by burdened to some extent due to the user's subjectivity. The example could be the possibility to simulate similar phenological development with different combinations of cardinal temperatures and thermal sums or due to the large variability in the values available in the literature for the parameters involved (Confalonieri *et al.*, 2016). In consequence, extending research to less modelled crops such as winter oilseed rape is strongly recommended (e.g. Rötter *et al.*, 2018). In the case of CROPSYST, the robustness of winter oilseed rape should be reconsidered due to the very high harvest index results from the current study (Fig. 7). On the other hand, errors within some models under such conditions are quite surprising, and their reparameterization or improvement should be considered. Especially in the case of winter crops, the relative RMSEs (normalized by the average observed yield) for a majority of the models exceeded 20%, which could be considered as a threshold for successful calibration (e.g. Ruiz-Ramos *et al.*, 2018).

The ability of models to explain yield variability could be attributed to the differences within their characteristics (Table 2). Consequences could be evident also from individual growth outputs such as the simulated maximum values of leaf area index during the season (LAI$_{MAX}$) (Fig. 6). Relevant differences in leaf area index were reported for various models for winter wheat in Palosuo *et al.* (2011) as well. On the other hand, the reported harvest indexes by models were at a stable level (except CROPSYST for winter oilseed rape) in the current study (Fig. 7).

In some cases, the individual models failed completely, even for variables with known target values, e.g. FASSET for winter crops maturity dates and AGROTOOL for silage maize yield. Failure is defined here if the RMSE of a model to a specific target variable is higher than the mean RMSE of all models plus 1.5 times of the standard deviation (as it is indicated within Tables 4–6). This is another argument for preferring EnsMED to EnsAVG (or use at least both of them) to avoid effects due to simulated outliers in multi-model ensembles (Wallach *et al.*, 2018). Using EnsMED and EnsAVG for seasonal yield estimates results in lower variability than the observed values. This is somewhat problematic in terms of the desired capability of model ensembles to simulate extremely low yield under adverse conditions (e.g. future droughts and heat stress). Simultaneously simpler models could also bring important robustness, but it is necessary to balance such approaches by the parallel use of more complex models, which allows more processes to be analyzed. In some cases, the yields formation based on the harvest index approach brings problems (e.g. in the case of winter oilseed rape by CROPSYST), but a similar approach was also used within the most successful of included models EPIC and AQUACROP (see Table 2). Fronzek *et al.* (2018) provided insight into differences within wheat models behaviour and concluded about the closer correspondence of sensitivity to temperature and precipitation change in models using partitioning schemes for yield formation than in those using a harvest index approach.

Conclusions

The responses of the single models and the multi-model ensemble were not found to be consistent through all the tested crops, primarily because although there are countless calibration studies for winter wheat, spring barley and silage maize, a smaller number of studies have focused on winter oilseed rape, and modellers have less experience with modelling this crop. Hence, further calibration works and connected research are recommended for crops, that are less common for crop modellers. Also, such crops are necessary for successful simulations of crop rotations and more complex soil–atmosphere–crop–farmer interaction assessments.

It can be concluded that even in cases when yields are known, there are significant differences in results between individual models. In general, spring crop yield was simulated more satisfactorily than winter crops. The poorest results were obtained for winter oilseed rape.

For correct ensemble crop rotation simulations, only models with reasonable accuracy (i.e. without failures) across all included crops and investigated variables within the target environment (which is not automatic) should be selected and based on robust calibration/verification studies. Modelling of anthesis and maturity was generally best simulated by the ensemble compared to the individual model results, whereas EnsMED is better than EnsAVG. The yield was better estimated by the best models than by the tested ensemble, which nevertheless provided robust results.

Considering RMSE and IA, and all the sites together, AQUACROP resulted in the best model for simulating the yield of winter wheat, spring barley, and silage maize, and EPIC resulted in the best model for winter oilseed rape. Regarding the MBE metric, CROPSYST was best for winter wheat, DSSAT1 was best for spring barley, and MONICA was best for silage maize and winter oilseed rape. Taking into account phenology and yields together, EnsMED was identified as the most universal predictor (compared to EnsAVG and individual crop models). This was also proved within the individual stations, so across the tested climatic conditions. This is despite ensemble composition and the inclusion of less accurate models. Some degree of simulations uncertainty could be connected to the user´s subjectivity within the calibration process. Also, in this case, the ensemble approach (including ensembles of the same models) can help increase the accuracy and robustness of estimates and with quantification of uncertainty. Achieved results together with the characteristic lower variability of ensemble outputs against the observed values can be used for planning and interpretation of studies focused on the impacts of changed agrometeorological conditions on individual crops as well as crop rotations.

References

Addiscott TM and Whitmore AP (1987) Computer simulation of changes in soil mineral nitrogen and crop nitrogen during autumn, winter and spring. *Journal of Agricultural Science* **109**, 141–157.

Asseng S, Ewert F, Rosenzweig C, Jones JW, Hatfield JL, Ruane AC, Boote KJ, Thorburn PJ, Rötter RP, Cammarano D, Brisson N, Basso B, Martre P, Aggarwal PK, Angulo C, Bertuzzi P, Biernath C, Challinor AJ, Doltra J, Gayler S, Goldberg R, Grant R, Heng L, Hooker J, Hunt LA, Ingwersen J, Izaurralde RC, Kersebaum KC, Müller C, Naresh Kumar S, Nendel C, O'Leary G, Olesen JE, Osborne TM, Palosuo T, Priesack E, Ripoche D, Semenov MA, Shcherbak I, Steduto P, Stöckle C, Stratonovitch P, Streck T, Supit I, Tao F, Travasso M, Waha K, Wallach D, White JW, Williams JR and Wolf J (2013) Uncertainty in simulating wheat yields under climate change. *Nature Climate Change* **3**, 827–832.

Asseng S, Ewert F, Martre P, Rötter RP, Lobell DB, Cammarano D, Kimball BA, Ottman MJ, Wall GW, White JW, Reynolds MP, Alderman PD, Prasad PVV, Aggarwal PK, Anothai J, Basso B, Biernath C, Challinor AJ, De Sanctis G, Doltra J, Fereres E, Garcia-Vila M, Gayler S, Hoogenboom G, Hunt LA, Izaurralde RC, Jabloun M, Jones CD, Kersebaum KC, Koehler A-K, Müller C, Naresh Kumar S, Nendel C, O'Leary G, Olesen JE, Palosuo T, Priesack E, Eyshi Rezaei E, Ruane AC, Semenov MA, Shcherbak I, Stöckle C, Stratonovitch P, Streck T, Supit I, Tao F, Thorburn PJ, Waha K, Wang E, Wallach D, Wolf J, Zhao Z and Zhu Y (2015) Rising temperatures reduce global wheat production. *Nature Climate Change* **5**, 143–147.

Bassu S, Brisson N, Durand JL, Boote K, Lizaso J, Jones JW, Rosenzweig C, Ruane AC, Adam M, Baron C, Basso B, Biernath C, Boogaard H, Conijn S, Corbeels M, Deryng D, De Sanctis G, Gayler S, Grassini P, Hatfield J, Hoek S, Izaurralde C, Jongschaap R, Kemanian AR, Kersebaum KC, Kim SH, Kumar NS, Makowski D, Müller C, Nendel C, Priesack E, Pravia MV, Sau F, Shcherbak I, Tao F, Teixeira E, Timlin D and Waha K (2014) How do various maize crop models vary in their responses to climate change factors? *Global Change Biology* **20**, 2301–2320.

CISTA Central Institute for Supervising and Testing in Agriculture (2004) Přehledy odrůd: Kukuřice 2004 (Variety overviews: Maize 2004). The National Plant Variety Office. ISBN 80-86548-61-9. Available at http://www.agris.cz/Content/files/main_files/57/135173/PO_kukurice04.pdf (accessed 7 November 2018).

CISTA Central Institute for Supervising and Testing in Agriculture (2007) Seznam doporučených odrůd 2007 (List of recommended varieties 2007). The National Plant Variety Office. ISBN 80-86548-92-9. Available at http://eagri.cz/public/web/file/112995/SDO_PO_Obilniny_07.pdf (accessed 7 November 2018).

CISTA Cetral Institute for Supervising and Testing in Agriculture (2006) Seznam doporučených odrůd 2006: Řepka olejka (List of recommended varieties 2006: Oilseed rape). The Nationl Plant Variety Office. ISBN 80-86548-75-9. Available at http://eagri.cz/public/web/file/112934/SDO_PO_Olejniny.pdf (accessed 7 November 2018).

Confalonieri R, Orlando F, Paleari L, Stella T, Gilardelli C, Movedi E, Pagani V, Cappelli G, Vertemara A, Alberti L, Alberti P, Atanassiu S, Bonaiti M, Cappelletti G, Ceruti M, Confalonieri A, Corgatelli G, Corti P, Dell'Oro M, Ghidoni A, Lamarta A, Maghini A, Mambretti M, Manchia A, Massoni G, Mutti P, Pariani S, Pasini D, Pesenti A, Pizzamiglio G, Ravasio A, Rea A, Santorsola D, Serafini G, Slavazza M and Acutis M (2016) Uncertainty in crop model predictions: what is the role of users? *Environmental Modelling & Software* **81**, 165–173.

Corbeels M, Berre D, Rusinamhodzi L and Lopez-ridaura S (2018) Can we use crop modelling for identifying climate change adaptation options ? *Agricultural and Forest Meteorology* **256–257**, 46–52.

Davies JA and McKay DC (1988) Evaluation of selected models for estimating solar radiation on horizontal surfaces. *Solar Energy* **43**, 153–168.

Dobos L (2004) Listina registrovaných odrôd. Vestník Ministerstva pôdohospodárstva Slovenskej republiky (List of registered varieties. Bulletin of the Ministry of Agriculture of the Slovak Republic), p. 216.

Fox DG (1981) Judging air quality model performance: a summary of the AMS workshop on dispersion model performance. *Bulletin of the American Meteorological Society* **62**, 599–609.

Fronzek S, Pirttioja N, Carter TR, Bindi M, Ho H, Palosuo T, Ruiz-ramos M, Tao F, Trnka M, Acutis M, Asseng S, Baranowski P, Basso B, Bodin P, Buis S, Cammarano D, Deligios P, Destain M, Dumont B, Ewert F, Ferrise R, François L, Gaiser T, Hlavinka P, Jacquemin I, Christian K, Kollas C, Krzyszczak J, Lorite IJ, Minet J, Minguez MI, Montesino M,

Moriondo M, Müller C, Nendel C, Öztürk I, Perego A, Rodríguez A, Ruane AC, Ruget F, Sanna M, Semenov MA, Slawinski C, Stratonovitch P, Supit I, Waha K, Wang E, Wu L, Zhao Z and Rötter RP (2018) Classifying multi-model wheat yield impact response surfaces showing sensitivity to temperature and precipitation change. *Agricultural systems* **159**, 209–224.

Gobin A (2018) Weather related risks in Belgian arable agriculture. *Agricultural Systems* **159**, 225–236.

Hansen S, Jensen HE, Nielsen NE and Svendsen H (1990) DAISY: Soil Plant Atmosphere System Model. NPO Report No. A 10. The National Agency for Environmental Protection, Copenhagen, 272 pp.

Hlavinka P, Trnka M, Balek J, Semerádová D, Hayes M, Svoboda M, Eitzinger J, Možný M, Fischer M, Hunt E and Žalud Z (2011) Development and evaluation of the SoilClim model for water balance and soil climate estimates. *Agricultural Water Management* **98**, 1249–1261.

Hlavinka P, Trnka M, Kersebaum KC, Čermák P, Pohanková E, Orság M, Pokorný E, Fischer M, Brtnický M and Žalud Z (2014) Modelling of yields and soil nitrogen dynamics for crop rotations by HERMES under different climate and soil conditions in the Czech Republic. *Journal of Agricultural Science* **152**, 188–204.

Hlavinka P, Kersebaum KC, Dubrovský M, Fischer M, Pohanková E, Balek J, Žalud Z and Trnka M (2015) Water balance, drought stress and yields for rainfed field crop rotations under present and future conditions in the Czech Republic. *Climate Research* **65**, 175–192.

Holzworth DP, Huth NI, Peter G, Zurcher EJ, Herrmann NI, Mclean G, Chenu K, Van Oosterom EJ, Snow V, Murphy C, Moore AD, Brown H, Whish JPM, Verrall S, Fainges J, Bell LW, Peake AS, Poulton PL, Hochman Z, Thorburn PJ, Gaydon DS, Dalgliesh NP, Rodriguez D, Cox H, Chapman S, Doherty A, Teixeira E, Sharp J, Cichota R, Vogeler I, Li FY, Wang E, Hammer GL, Robertson MJ, Dimes JP, Whitbread AM, Hunt J, Van Rees H, Mcclelland T, Carberry PS, Hargreaves JNG, Macleod N, Mcdonald C, Harsdorf J, Wedgwood S and Keating BA (2014) Environmental modelling & software APSIM e evolution towards a new generation of agricultural systems. *Environmental Modelling and Software* **62**, 327–350.

Hoogenboom G, Jones JW, Wilkens PW, Porter CH, Boote KJ, Hunt LA, Singh U, Lizaso JI, White JW, Uryasev O, Ogoshi R, Koo J, Shelia V and Tsuji GY (2015) Decision Support System for Agrotechnology Transfer (DSSAT) Version 4.6. DSSAT Foundation, Prosser, Washington.

Jones J, Hoogenboom G, Porter C, Boote K, Batchelor W, Hunt L, Wilkens P, Singh U, Gijsman A and Ritchie J (2003) The DSSAT cropping system model. *European Journal of Agronomy* **18**, 235–265.

Kersebaum KC (2007) Modelling nitrogen dynamics in soil-crop systems with HERMES. *Nutrient Cycling in Agroecosystems* **77**, 39–52.

Kersebaum KC, Boote KJ, Jorgenson JS, Nendel C, Bindi M, Frühauf C, Gaiser T, Hoogenboom G, Kollas C, Olesen JE, Rötter RP, Ruget F, Thorburn PJ, Trnka M and Wegehenkel M (2015) Analysis and classification of data sets for calibration and validation of agro-ecosystem models. *Environmental Modelling & Softwar* **72**, 402–441.

Kollas C, Kersebaum KC, Nendel C, Manevski K, Müller C, Palosuo T, Armas-Herrera CM, Beaudoin N, Bindi M, Charfeddine M, Conradt T, Constantin J, Eitzinger J, Ewert F, Ferrise R, Gaiser T, De Cortazar-Atauri IG, Giglio L, Hlavinka P, Hoffmann H, Hoffmann MP, Launay M, Manderscheid R, Mary B, Mirschel W, Moriondo M, Olesen JE, Öztürk I, Pacholski A, Ripoche-Wachter D, Roggero PP, Roncossek S, Rötter RP, Ruget F, Sharif B, Trnka M, Ventrella D, Waha K, Wegehenkel M, Weigel H-J and Wu L (2015) Crop rotation modelling – a European model intercomparison. *European Journal of Agronomy* **70**, 98–111.

Legates DR and McCabe Jr GJ (1999) Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water Resources Research* **35**, 233–241.

Martre P, Wallach D, Asseng S, Ewert F, Jones JW, Rötter RP, Boote KJ, Ruane AC, Thorburn PJ, Cammarano D, Hatfield JL, Rosenzweig C, Aggarwal PK, Angulo C, Basso B, Bertuzzi P, Biernath C, Brisson N, Challinor AJ, Doltra J, Gayler S, Goldberg R, Grant RF, Heng L, Hooker J, Hunt LA, Ingwersen J, Izaurralde RC, Kersebaum KC, Müller C, Kumar SN, Nendel C, o'Leary G, Olesen JE, Osborne TM, Palosuo T, Priesack E, Ripoche D, Semenov MA, Shcherbak I, Steduto P, Stöckle CO, Stratonovitch P, Streck T, Supit I, Tao F, Travasso M, Waha K, White JW and Wolf J (2015) Multimodel ensembles of wheat growth: many models are better than one. *Global Change Biology* **21**, 911–925.

Nendel C, Berg M, Kersebaum KC, Mirschel W, Specka X, Wegehenkel M, Wenkel KO and Wieland R (2011) The MONICA model: testing predictability for crop growth, soil moisture and nitrogen dynamics. *Ecological Modelling* **222**, 1614–1625.

Nendel C, Wieland R, Mirschel W, Specka X, Guddat C and Kersebaum KC (2013) Simulating regional winter wheat yields using input data of different spatial resolution. *Field Crop Research* **145**, 67–77.

Olesen JE, Askegaard M and Rasmussen IA (2000) Design of an organic farming crop-rotation experiment. *Acta Agriculturae Scandinavica, Section B – Soil & Plant Science* **50**, 13–21.

Orlandini S, Nejedlik P, Eitzinger J, Alexandrov V, Toulios L, Calanca P, Trnka M and Olesen JE (2008) Impacts of climate change and variability on European agriculture: results of inventory analysis in COST 734 countries. *Annals of the New York Academy of Sciences* **1146**, 338–333.

Palosuo T, Kersebaum KC, Angulo C, Hlavinka P, Moriondo M, Olesen J, Patil R, Ruget F, Rumbaur C, Takac J, Trnka M, Bindi M, Caldag B, Ewert F, Ferrise R, Mirschel W, Saylan L, Siska B and Rötter R (2011) Simulation of winter wheat yield and its variability in different climates of Europe: a comparison of eight crop growth models. *European Journal of Agronomy* **35**, 103–114.

Pirttioja N, Carter TR, Fronzek S, Bindi M, Hoffmann H, Palosuo T, Ruiz-Ramos M, Tao F, Trnka M, Acutis M, Asseng S, Baranowski P, Basso B, Bodin P, Buis S, Cammarano D, Deligios P, Destain M-F, Dumont B, Ewert F, Ferrise R, Francois L, Gaiser T, Hlavinka P, Jacquemin I, Kersebaum KC, Kollas C, Krzyszczak J, Lorite IJ, Minet J, Minguez MI, Montesino M, Moriondo M, Muller C, Nendel C, Öztürk I, Perego A, Rodríguez A, Ruane AC, Ruget F, Sanna M, Semenov MA, Slawinski C, Stratonovitch P, Supit I, Waha K, Wang E, Wu L, Zhao Z and Rötter RP (2015) A crop model ensemble analysis of temperature and precipitation effects on wheat yield across a European transect using impact response surfaces. *Climate Research* **65**, 87–105.

Poluektov RA, Fintushal SM, Oparina IV, Shatskikh DV, Terlleev VV and Zakharova ET (2002) AGROTOOL – a system for crop simulation. *Archiv für Acker- und Pflanzenbau und Bodenkunde* **48**, 609–635.

Ritchie JT (1981) Water dynamics in the soil-plant-atmosphere system. *Plant and Soil* **58**, 81–96.

Rodríguez A, Ruiz-ramos M, Palosuo T, Carter TR, Fronzek S, Lorite IJ, Ferrise R, Pirttioja N, Bindi M, Baranowski P, Buis S, Cammarano D, Chen Y, Dumont B, Ewert F, Gaiser T, Hlavinka P, Ho H, Höhn JG, Jurecka F, Kersebaum KC, Trnka M, De Wit A and Rötter RP (2018) Implications of crop model ensemble size and composition for estimates of adaptation effects and agreement of recommendations. *Agricultural and Forest Meteorology* **264**, 351–362.

Rötter RP, Carter TR, Olesen JE and Porter JR (2011) Crop – climate models need an overhaul. *Nature Climate Change* **1**, 175–177.

Rötter RP, Palosuo T, Christian K, Angulo C, Bindi M, Ewert F, Ferrise R, Hlavinka P, Moriondo M, Nendel C and Olesen JE (2012) Simulation of spring barley yield in different climatic zones of Northern and Central Europe: a comparison of nine crop models. *Field Crop Research* **133**, 23–36.

Rötter RP, Appiah M, Fichtler E, Kersebaum KC, Trnka M and Ho MP (2018) Linking modelling and experimentation to better capture crop impacts of agroclimatic extremes – a review. *Field Crop Research* **221**, 142–156.

Ruiz-ramos M, Ferrise R, Rodríguez A, Lorite IJ, Bindi M, Carter TR, Fronzek S, Palosuo T, Pirttioja N, Baranowski P, Buis S, Cammarano D, Chen Y, Dumont B, Ewert F, Gaiser T, Hlavinka P, Hoffmann H, Höhn JG, Jurecka F, Kersebaum KC, Krzyszczak J, Lana M, Mechiche-alami A, Minet J, Montesino M, Nendel C, Porter JR, Ruget F, Semenov MA, Steinmetz Z, Stratonovitch P, Supit I, Tao F, Trnka M, De Wit A and Rötter RP (2018) Adaptation response surfaces for managing wheat under perturbed climate and CO_2 in a Mediterranean environment. *Agricultural Systems* **159**, 260–274.

Salo TJ, Palosuo T, Kersebaum KC, Nendel C, Angulo C, Ewert F, Bindi M, Calanca P, Klein T, Moriondo M, Ferrise R, Olesen JE, Patil RH, Ruget F, Takáč J, Hlavinka P, Trnka M and Rötter RP (2016) Comparing the

performance of 11 crop simulation models in predicting yield response to nitrogen fertilization. *Journal of Agricultural Science* **154**, 1218–1240.

Soltani A and Sinclair TR (2015) A comparison of four wheat models with respect to robustness and transparency: simulation in a temperate, sub-humid environment. *Field Crop Research* **175**, 37–46.

Steduto P, Hsiao TC, Raes D and Fereres E (2009) AquaCrop: the FAO crop model to simulate yield response to water: I. Concepts and underlying principles. *Agronomy Journal* **101**, 426–437.

Stöckle CO, Donatelli M and Nelson R (2003) CropSyst, a cropping systems simulation model. *European Journal of Agronomy* **18**, 289–307.

Wallach D, Mearns LO, Ruane AC, Rötter RP and Asseng S (2016) Lessons from climate modeling on the design and use of ensembles for crop modeling. *Climate Change* **139**, 551–564.

Wallach D, Martre P, Liu B, Asseng S, Ewert F, Thorburn PJ, van Ittersum M, Aggarwal PK, Ahmed M, Basso B, Biernath C, Cammarano D, Challinor AJ, De Sanctis G, Dumont B, Rezaei EE, Fereres E, Fitzgerald GJ, Gao Y, Garcia-Vila M, Gayler S, Girousse C, Hoogenboom G, Horan H, Izaurralde RC, Jones CD, Kassie BT, Kersebaum KC, Klein C, Koehler AK, Maiorano A, Minoli S, Müller C, Naresh Kumar S, Nendel C, O'Leary GJ, Palosuo T, Priesack E, Ripoche D, Rötter RP, Semenov MA, Stöckle C, Stratonovitch P, Streck T, Supit I, Tao F, Wolf J and Zhang Z (2018) Multi-model ensembles improve predictions of crop-environment-management interactions. *Global Change Biology* **24**, 5072–5083.

Wallor E, Kersebaum KC, Ventrella D, Bindi M, Cammarano D, Coucheney E, Gaiser T, Garofalo P, Giglio L, Giola P, Hoffmann MP, Iocola I, Lana M, Lewan E, Maharjan GR, Moriondo M, Mula L, Nendel C, Pohankova E, Roggero PP, Trnka M and Trombi G (2018) The response of process-based agro-ecosystem models to within-field variability in site conditions. *Field Crops Research* **228**, 1–19.

Williams JR, Jones CA, Kiniry JR and Spanel DA (1989) The EPIC crop growth model. *Transactions of the ASAE* **32**, e511.

Willmott CJ (1982) Some comments on the evaluation of model performance. *Bulletin of the American Meteorological Society* **64**, 1309–1313.

Zhao G, Hoffmann H, Yeluripati J, Specka X, Nendel C, Coucheney E, Kuhnert M, Tao F, Constantin J, Raynal H, Teixeira E, Grosz B, Doro L, Kiese R, Eckersten H, Haas E, Cammarano D, Kassie B, Moriondo M, Trombi G, Bindi M, Biernath C, Heinlein F, Klein C, Priesack E, Lewan E, Kersebaum KC, Rötter RP, Roggero PP, Wallach D, Asseng S, Siebert S, Gaiser T and Ewert F (2016) Evaluating the precision of eight spatial sampling schemes in estimating regional means of simulated yield for two crops. *Environmental Modelling and Software* **80**, 100–112.