Field Crops Research 111 (2009) 290–302

Contents lists available at ScienceDirect

Field Crops Research

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

Using pattern recognition for estimating cultivar coefficients of a crop simulation model

Mohammad Bannayan [a,b,*], Gerrit Hoogenboom [b]

[a] Ferdowsi University of Mashhad, Faculty of Agriculture, P.O. Box 91775-1163, Mashhad, Iran
[b] Department of Biological and Agricultural Engineering, University of Georgia, Griffin, GA 30223, USA

ARTICLE INFO

Article history:
Received 30 August 2008
Received in revised form 16 January 2009
Accepted 19 January 2009

Keywords:
Process-based simulation model
Parameter estimation
Pattern recognition
k-NN approach
CSM-CERES-Maize model
DSSAT

ABSTRACT

The introduction of a new cultivar in a process-based crop simulation model requires the estimation of cultivar coefficients that define its growth and development characteristics. An accurate estimation of these coefficients requires replicated field experiments that, in many cases, are not available to crop model users. The objective of this study was to employ a pattern recognition approach to estimate cultivar coefficients from a minimum set of experimental data for use with a crop simulation model. The pattern recognition approach is based on similarity measures. Its main goal is to classify groups of data or patterns based on either a priori knowledge or on statistical information extracted from the patterns. Based on the similarity measure as the central calculation of the pattern recognition approach, the algorithm searches the space of features of other cultivars in the database to find the most similar cultivar as the *best match* to the target cultivar. The approach of this study was based on a few key characteristics of maize crop growth and development, including anthesis and harvest maturity dates, maximum leaf area index (LAI_{max}), final above ground biomass, and grain yield, which were used as the features vector. To construct the feature database, 27,789 hypothetical cultivars were constructed by combining different values of the six cultivar coefficients of the Cropping System Model (CSM)-CERES-Maize. Experiments performed in Florida (FL) and Iowa (IA) USA, Spain, central Punjab, Pakistan, and in Piracicaba, SP, Brazil were selected and later modified to provide a full potential production environment. The crop model was run for potential production for all 27,789 hypothetical cultivars and the outputs of these simulations were used as the feature database. For evaluation of this approach, we used the features for 29 different maize cultivars as reported from field experiments that are available in DSSAT maize cultivar database and also for four additional cultivars of which two had not been used in any aspect of this study. The model was run for all 33 cultivars, using the *best match* cultivar coefficients, for the conditions of the three study sites and locations where the latter four cultivars have been grown. The simulated crop characteristics were compared with the same simulated crop characteristics based on the original coefficients used to run the simulation model. We found that the approach based on pattern recognition was able to estimate the cultivar coefficients with a reasonable accuracy. The coefficient of determination (r^2), root mean square difference (RMSD), and relative root mean square of difference (RMSDr) confirmed that this approach provided reliable estimates for the maize cultivar coefficients. The highest R^2 (0.98) was obtained for anthesis in Florida and the lowest (0.57) was obtained for grain yield in Spain. The highest RMSD (8.8) was obtained for maturity in Spain, while the lowest RMSD (1.1) was obtained for aboveground biomass in Florida. Although the values for RMSD were different across the different sites, this approach provided a level of accuracy that might be acceptable, especially for users who only have one year of experimental data and demand the best possible initial guess for the coefficients of their specific cultivar. This approach has been implemented in a simple tool that can be easily applied by users of DSSAT and the CSM-CERES-Maize model.

1. Introduction

1.1. Pattern recognition

There has been a significant interest in similarity-based techniques such as the *k*-nearest neighbor (*k*-NN) approach, in various scientific disciplines, (Davis and Nihan, 1991; Jagtap et al.,

* Corresponding author at: Ferdowsi University of Mashhad, Faculty of Agriculture, P.O. Box 91775-1163, Mashhad, Iran. Tel.: +1 770 229 3436; fax: +1 770 228 7218.
E-mail address: bannayan@uga.edu (M. Bannayan).

0378-4290/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.fcr.2009.01.007

Table 1
Description of the cultivar coefficients of the CSM-CERES-Maize model.

P1 (°C day)	Thermal time from seedling emergence to the end of the juvenile phase (expressed in degree days, °C day, above a base temperature of 8 °C) during which the plant is not responsive to changes in photoperiod.
P2 (days)	Extent to which development (expressed as days) is delayed for each hour increase in photoperiod above the longest photoperiod at which development proceeds at a maximum rate (which is considered to be 12.5 h).
P5 (°C day)	Thermal time from silking to physiological maturity (expressed in degree days above a base temperature of 8 °C).
G2 (Nr)	Maximum possible number of kernels per plant.
G3 (mg day^{-1})	Kernel filling rate during the linear grain filling stage and under optimum conditions (mg day^{-1}).
PHINT (°C day)	Phyllochron interval; the interval in thermal time (degree days) between successive leaf tip appearances.

2004; Chi and Bruzzone, 2005; Wu et al., 2005; Nemes et al., 2006; Bannayan and Hoogenboom, 2008a,b). The *k*-NN approach is one of the most attractive pattern classification algorithms (Yakowitz, 1987; Buishand and Brandsma, 2001). Pattern recognition aims to classify data (patterns) based on either a priori knowledge or on statistical information extracted from the patterns. Similarity measures are central to most pattern recognition techniques. Similarity can be determined through comparison, which requires considering the same characters in two individuals that are subject to the comparison. Analogy techniques that require comparison, use the observed behavior in one class of phenomena and relates this to a different class of phenomena (Goswami, 1995; Chiu and Huang, 2007). However, this does not mean that the two classes function exactly the same way. When all kinds of classifications of measurements or observations, e.g. crop biomass, grain yield, etc., are known, then the pattern recognition based on similarity decides whether any given object, e.g. another set of crop biomass, grain yield, etc., belongs to that kind of classification, e.g. cultivar coefficients. We set C as the space of objects, e.g. cultivars, and each object among C has P character indices, as $C_1, C_2, \ldots, C_p$. As a feature vector these indices describe a pattern of specific characters of the object, e.g. crop biomass, grain yield, etc. Every set of $C_{i.n}$ can be classified into one or more kinds and every kind can be associated with another set of data that characterize a pattern of $S_1, S_2, \ldots, S_n$, e.g. crop cultivar coefficients. Thus, if a given object with the measured or observed feature vector (C_{new}) shows the highest similarity with C_x from the database, then the $S_{x1}, S_{x2}, \ldots, S_{xn}$ pattern can be attributed as the *best match* for the pattern of the

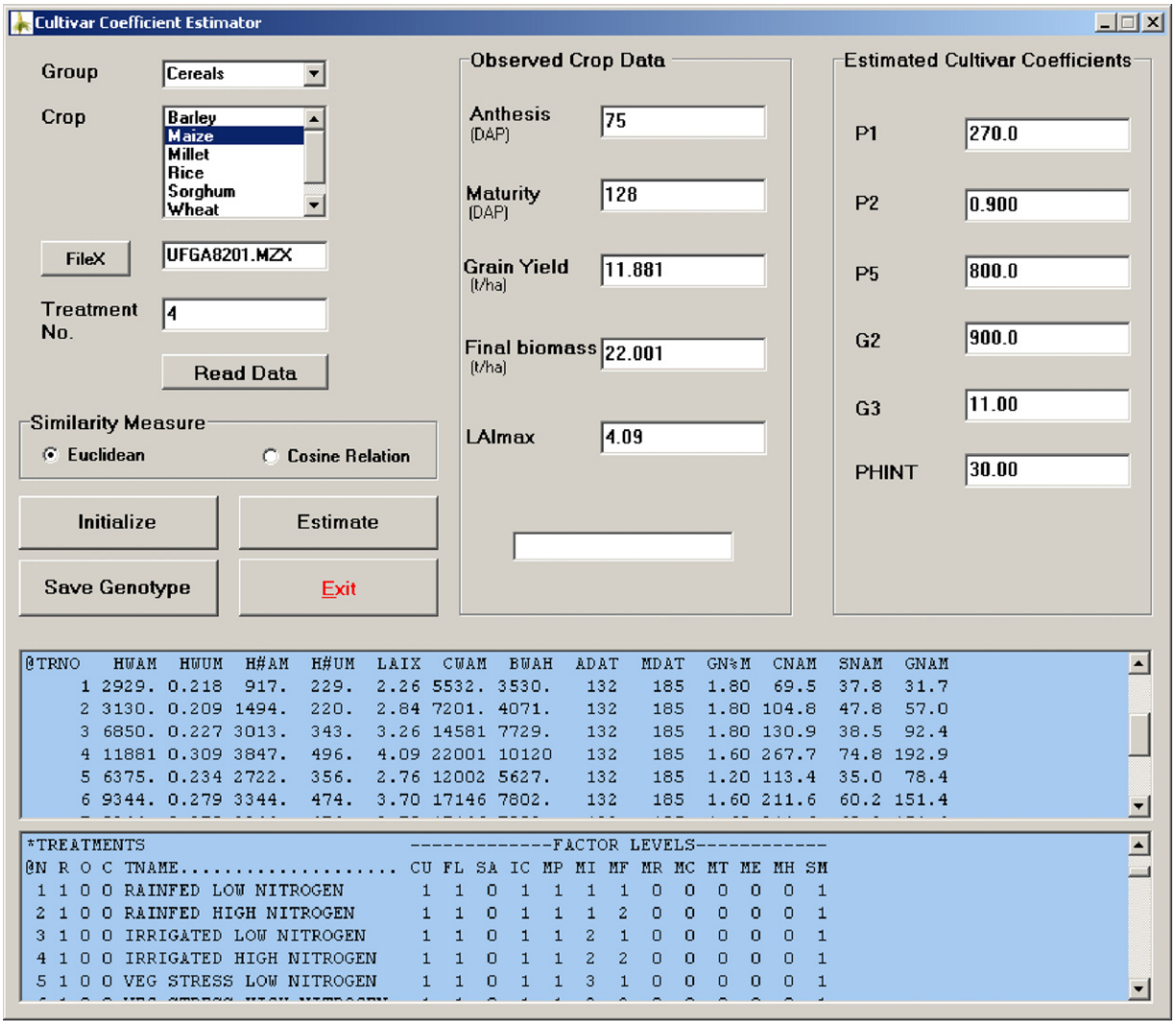

Fig. 1. Interface of the cultivar coefficient estimator for maize.

given object. This approach requires a large database to be representative of all possible combinations of objects, e.g. cultivars. If the database represents only a subset of the population, then the pattern estimation cannot be extended beyond this subset. If the database can be considered as representative, then it is also expected that the given object be uniform to the population of objects.

1.2. Crop model and cultivar coefficients

The development and application of complex ecophysiological simulation models are an integral part of agriculture. Process-based crop simulation models are valuable tools to help represent our understanding and knowledge of a current or future cropping system (Tsuji et al., 1998). These models support biological research and can be used to identify knowledge gaps (Kropff et al., 1994; Bannayan et al., 2007), generating new experimental hypotheses (Azam-Ali et al., 2001) and, once tested, can be used for refinement and improvement that can ultimately lead to more sophisticated hypotheses (Loomis et al., 1979; Hammer et al., 2002; Jones et al., 2003). Crop modeling systems are designed to assist in analyzing the growth and development of crops in response to environmental variables and are able to predict how changes in environment will affect growth and yield (Bannayan et al., 2005). It is also possible to change one or multiple environmental variables in order to predict the response of a target crop in various environments (Bannayan et al., 2004). The cultivar coefficients employed in these models are usually constant values that enable the model to realize the phenological and physiological differences among cultivars of a given plant species. The relationship of these cultivar coefficients to field performance and linkage to genomics was studied by White and Hoogenboom (1996), Boote et al. (2003) and Hoogenboom et al. (2004a). They found that a very limited numbers of crop productivity characters that are normally obtained in Crop Performance Trials can be used to obtain genomic information for each character. However, such an approach demands more detailed information from both controlling genes and associated coefficients (Bannayan et al., 2007). White and Hoogenboom (1996) used simple linear effects of seven genes to replace the empirically determined coefficients of BEANGRO (Hoogenboom et al., 1994a) to develop the GeneGro model. There was a good performance of the GeneGro model when compared to the observed and simulated data for 20 cultivars, but this approach was unable to represent gene action at the process level. Measuring crop cultivar coefficients is, in fact, often very difficult and the data needed are often not readily available (Banterng et al., 2004, 2006; Suriharn et al., 2007). Cultivar coefficients are normally determined based on data collected in experiments that are conducted under optimum conditions over a range of environments and free from water, nutrient, and pest stresses. It is highly recommended that the experiments for determining cultivar coefficients be conducted over several planting dates at the same location or for the same planting dates across multiple locations (Hoogenboom et al., 1999). Because data collection for this type of experiment is elaborate and intensive, this recommendation is difficult to implement for breeding applications that involve a large number of genotypes. This is a major limitation to the application of crop simulation models.

Estimation of cultivar coefficients for the various process-based crop simulation models within the Decision Support System for Agrotechnology Transfer (DSSAT) suite of models (Jones et al., 2003; Hoogenboom et al., 2004b) is an initial step in crop model use. In earlier versions of DSSAT, GENCALC (Hunt et al., 1993; Hunt and Pararajasingham, 1994) was used to calculate the required cultivar coefficients. GENCALC estimates the coefficients via iterations of deterministic processes. The determination of cultivar coefficients with GENCALC requires inputs about the crop, weather, and soil data for the target genotype. The coefficients are adjusted by varying their values according to the realistic physiological ranges of the crop, i.e., flowering date, physiological maturity date, critical daylength, seed size, etc. The coefficients are calculated in a specified sequence, with a user-defined step size over the interval, starting with those that relate to development performance (Hunt et al., 1993). However, the coefficients derived by this approach use data from one set of environments and often change when they are derived using data from a different environment, e.g. different years or sites. Such variation in coefficients might be due to the employed formulation of relationship between variables within the tool. For example, Xue et al. (2004) showed in their evaluation of GENCALC that, using the final leaf number as a measure of the vernalization requirement, resulted in a linear relationship, while experimental evidence indicates a curvilinear relationship between the final leaf number and accumulated vernalization days. The curve becomes asymptotic to the final leaf number as the vernalization requirement is met (Brooking, 1996; Fowler et al., 1996). Such a linear relationship between temperature and development permitted the use of the concept of thermal time, the accumulation of daily temperature above a specified base temperature. In many cases, however, the use of a nonlinear relationship that allows for reduced activity at temperatures near the base and for a smooth transition to the optimum is preferable, both for development per se and vernalization (Yan and Hunt, 1999a,b).

Currently, the estimation of crop model coefficients within DSSAT is a fitting process that is conducted systematically using a rather labor-intensive, somewhat subjective, trial and error procedure (Boote et al., 1998; Mavromatis and Hansen, 2001). To obtain the best estimate for the cultivar coefficients, the user normally changes the coefficients of a similar cultivar or default values until the simulated outputs are in good agreement with

Table 2
An example file with observed yield, yield components and development data in the DSSAT format (FileA) for an experiment conducted in 1982 at the University of Florida in Gainesville, consisting of two nitrogen fertilizer levels and three irrigation levels for a total of six treatments.

*Exp. data (A): UFGA8201MZ Z $N \times$ irrigation, Gainesville

@TRNO	HWAM	HWUM	H#AM	H#UM	LAIX	CWAM	BWAH	ADAT	MDAT	GN%M	CNAM
1	2,929	0.218	917	229	2.26	5,532	3,530	132	185	69.5	37.8
2	3,130	0.209	1494	220	2.84	7,201	4,071	132	185	104.8	47.8
3	6,850	0.227	3013	343	3.26	14,581	7,729	132	185	130.9	38.5
4	11,881	0.309	3847	496	4.09	22,001	10,120	132	185	267.7	74.8
5	6,375	0.234	2722	356	2.76	12,002	5,627	132	185	113.4	35.0
6	9,344	0.279	3344	474	3.70	17,146	7,802	132	185	211.6	60.2

@TRNO: treatment number; HWAM: Mat Yield kg/ha yield at harvest maturity (kg [dm]/ha); H#AM: number #/m^2 Number at maturity (no/m^2); H#UM: number #/unit number at maturity (no/unit); LAIX: LAI maximum Leaf area index; CWAM: tops weight at maturity (kg [dm]/ha); BWAH: by-product kg/ha by-product removed during harvest (kg [dm]/ha); ADAT: anthesis date (YrDoy); MDAT: physiological maturity date (YrDoy); GN%M: grain *N* at maturity (%); CNAM: tops *N* at maturity (kg/ha).

Table 3
The range of values employed for the cultivar coefficient database.

P1																							
50	70	90	110	130	150	170	190	210	230	250	270	290	310	330	350	370	390	410	430	450	470	490	500
P2																							
0.1				0.3				0.5				0.6				0.8				0.9			
P5																							
400					500					600					700					800			
G2																							
400				500				600				700				800				900			
G3																							
5				8				11				14				17				20			
PHINT																							
30								50								60							

The cultivar coefficients are defined in Table 1.

field observations for phenology, growth, and ultimately yield. While growth analysis has been successful for a limited number of cultivars, new approaches must be developed (Boote et al., 2003). With the wealth of experimental data and the large number of generated databases, novel approaches together with computer analyses can provide an opportunity to develop new tools. It is well known that a substantial competitive advantage can be obtained by pattern recognition (Bannayan and Hoogenboom, 2008b). Pattern recognition can be regarded as a collection of methods for drawing inferences from data (Bannayan and Hoogenboom, 2008a). The overall goal of this study was to evaluate the use of pattern recognition for estimating the cultivar coefficients of a crop simulation model. The specific objective was to use *k*-NN procedure based on pattern recognition for estimating the maize cultivar coefficients for the CSM-CERES-Maize model. If successful, this approach could help with extending the application of crop models when limited crop growth and development data of a new cultivar are available.

2. Materials and methods

2.1. CSM-CERES-Maize model

The CSM-CERES-Maize model is a dynamic crop simulation model that simulates the complex management strategies of maize cropping systems for a wide range of weather and soil conditions and that can be used to analyze the interactions of these strategies with environmental conditions (Jones and Kiniry, 1986; Ritchie et al., 1998). Potential growth is a function of photosynthetically active radiation (PAR), the interception of radiation related to leaf area index, row spacing, plant population, and the conversion efficiency of radiation to biomass. The phenological phase determines assimilate partitioning for growth of roots, leaves, stem, and grains. The model quantifies the major processes governing growth and development of maize crop including phenological and reproductive development, vegetative and canopy development, organ formation, photosynthesis, assimilate allocation, and the dynamics of the soil and plant water and

Table 4
Comparison of the original (ST) and *best match* cultivar coefficients based on the *k*-NN approach for 29 maize cultivars for the simulated experiments for Florida (FL), Iowa (IA) and Spain (SP).

Cultivar	P1				P2				P5				G2				G3				PHINT			
	ST	FL	IA	SP	ST	FL	IA	SP	ST	FL	IA	SP	ST	FL	IA	SP	ST	FL	IA	SP	ST	FL	IA	SP
CP170	120	90	70	70	0	0.9	0.9	0.9	685	600	700	800	907.9	700	800	900	10	11	14	17	38.9	30	30	30
LG11	125	110	70	90	0	0.9	0.9	0.9	685	600	700	800	907.9	700	800	700	10	11	14	17	38.9	30	30	30
PIO 3995	130	150	50	110	0.3	0.5	0.1	0.9	685	700	600	700	907.9	800	700	900	8.6	8	17	14	38.9	30	30	30
Dekalbxl71	140	170	50	130	0.3	0.9	0.6	0.9	685	600	700	700	907.9	700	900	900	10.5	11	17	14	38.9	30	30	30
DEKALBXL45	150	170	90	130	0.4	0.9	0.9	0.9	685	600	600	700	907.9	700	900	900	10.2	11	14	14	38.9	30	30	30
B59*OH43	162	170	110	170	0.8	0.9	0.9	0.5	685	700	600	700	862.4	700	800	800	6.9	8	11	17	38.9	30	30	30
F16 X F19	165	150	50	130	0	0.9	0.6	0.9	685	600	700	700	907.9	700	900	900	10	11	17	14	38.9	30	30	30
WASHINGTON	165	170	90	150	0.4	0.1	0.9	0.1	715	800	700	800	825	700	900	800	11	11	14	14	38.9	30	30	30
B14XOH43	172	170	110	130	0.3	0.1	0.3	0.9	685	700	700	700	907.9	700	800	900	8.5	11	17	14	38.9	30	30	30
B60*R71	172	170	150	170	0.8	0.9	0.1	0.5	685	700	800	700	781.4	700	700	800	7.7	8	11	17	38.9	30	30	30
WF9*B37	172	170	150	150	0.8	0.1	0.9	0.1	685	700	800	800	907.9	700	800	800	10.2	11	14	14	38.9	30	30	30
Garst 8702	175	190	190	210	0.2	0.6	0.6	0.9	960	800	800	800	855.8	700	700	700	6	14	14	17	38.9	50	50	30
B14*C103	180	190	150	150	0.5	0.9	0.9	0.1	685	600	800	800	907.9	700	800	800	10.2	11	14	14	38.9	30	30	30
WASH/GRAIN1	185	190	210	210	0.4	0.1	0.9	0.9	775	600	800	800	836	800	700	700	12	11	17	17	38.9	30	30	30
PIO 3475	200	250	330	310	0.7	0.9	0.9	0.9	800	800	700	700	797.5	700	900	900	8.6	17	14	14	38.9	30	30	30
GL 582	200	190	270	230	0.7	0.9	0.1	0.9	750	800	800	800	750	700	800	900	8.6	8	14	11	38.9	30	30	30
GL 482	240	270	330	410	0.7	0.1	0.9	0.9	990	800	700	800	907	800	900	700	8.8	14	14	17	38.9	30	30	30
AGETI76	325	350	430	470	0.1	0.1	0.5	0.1	625	600	800	800	580	900	600	700	7.3	5	5	11	50	50	60	30
DK 611	260	250	270	330	0.1	0.1	0.5	0.9	800	800	800	700	850	900	500	900	6	5	17	14	48	50	60	30
PIO 33Y09	245	250	270	330	0.5	0.1	0.9	0.9	905	800	800	700	780	900	800	900	6	5	11	14	48	50	60	30
PIO 3563	216	190	270	310	0.6	0.6	0.1	0.9	830	800	800	700	860	700	800	900	8.8	14	20	14	48	50	60	30
C/LOL 499	182	190	150	170	0.5	0.5	0.1	0.5	650	600	800	700	750	700	700	800	8.7	14	11	17	46	50	30	30
Prisma GC Avg	280	330	310	410	0.78	0.6	0.6	0.9	789	800	800	800	650	900	700	700	6	5	14	17	38.9	30	50	30
OBA SUPER 2	270	310	330	410	0.6	0.9	0.9	0.9	780	800	700	800	840	700	900	700	7.8	8	11	17	45	30	60	30
EV8728-SR	265	270	330	330	0.6	0.9	0.9	0.9	800	800	700	700	900	700	900	900	7.2	8	11	14	45	30	60	30
OBA S2 Benin	170	130	150	210	0.6	0.6	0.6	0.9	760	800	800	800	800	600	700	700	8	11	14	17	50	60	50	30
EV-8449_TG	260	370	430	410	0.6	0.6	0.5	0.6	630	800	800	800	900	700	600	800	9	8	5	14	45	60	60	30
ASK740	215	190	270	270	0.75	0.1	0.1	0.1	850	800	800	800	700	900	800	700	5	5	8	11	48	50	60	30
Mokwa 87TZPB-SR	305	350	330	410	0.6	0.6	0.9	0.9	765	800	800	800	810	900	800	700	8	5	11	17	45	30	60	30

Refer to Table 1 for the definition of the P1, P2, P5, G2, G3, and Phint.

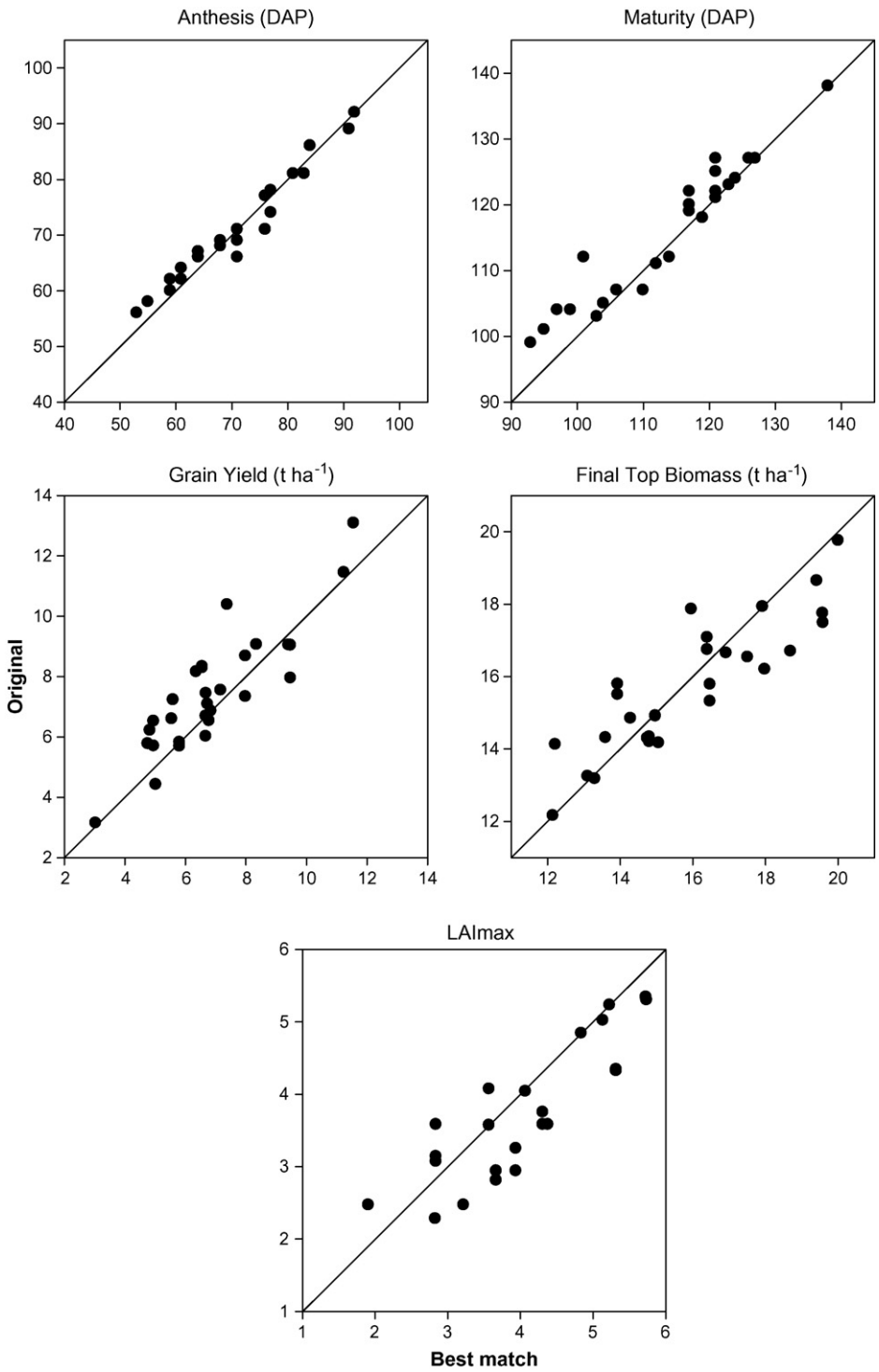

Fig. 2. Comparison of anthesis, maturity, grain yield, final biomass and LAI_{max} for 29 cultivars for Gainesville, Florida (a), Ames, Iowa, (b), and Spain (c). DAP: days after planting.

nitrogen processes. Thus, the model can simulate the impact of weather, and soil water and nitrogen dynamics on growth and yield. The model and its components have been well documented (Jones and Kiniry, 1986; Godwin and Jones, 1991; Hoogenboom et al., 1994b; Ritchie et al., 1998). The CSM-CERES-Maize and other DSSAT models have been used for a wide range of applications that simulate the response of maize to crop management and environmental factors for many regions across the world (Ritchie et al., 1998; Tsuji et al., 1998; Gungula et al., 2003; Jagtap and Abamu, 2003; Jones et al., 2003).

A number of cultivar-specific parameters are used by the CSM-CERES-Maize model to realize the responses of a specific cultivar to weather conditions, soil characteristics, and management actions. These cultivar coefficients describe (i) the cultivar sensitivity to daylength, (ii) durations of life cycle phases, (iii) vegetative growth traits, and (iv) reproductive growth traits (e.g. potential seed size) (Table 1). The life cycle phase coefficients relate to life cycle timing and are measured in "photothermal days." The latter is a unit that combines the standard concept of degree-days with a measure of daylength.

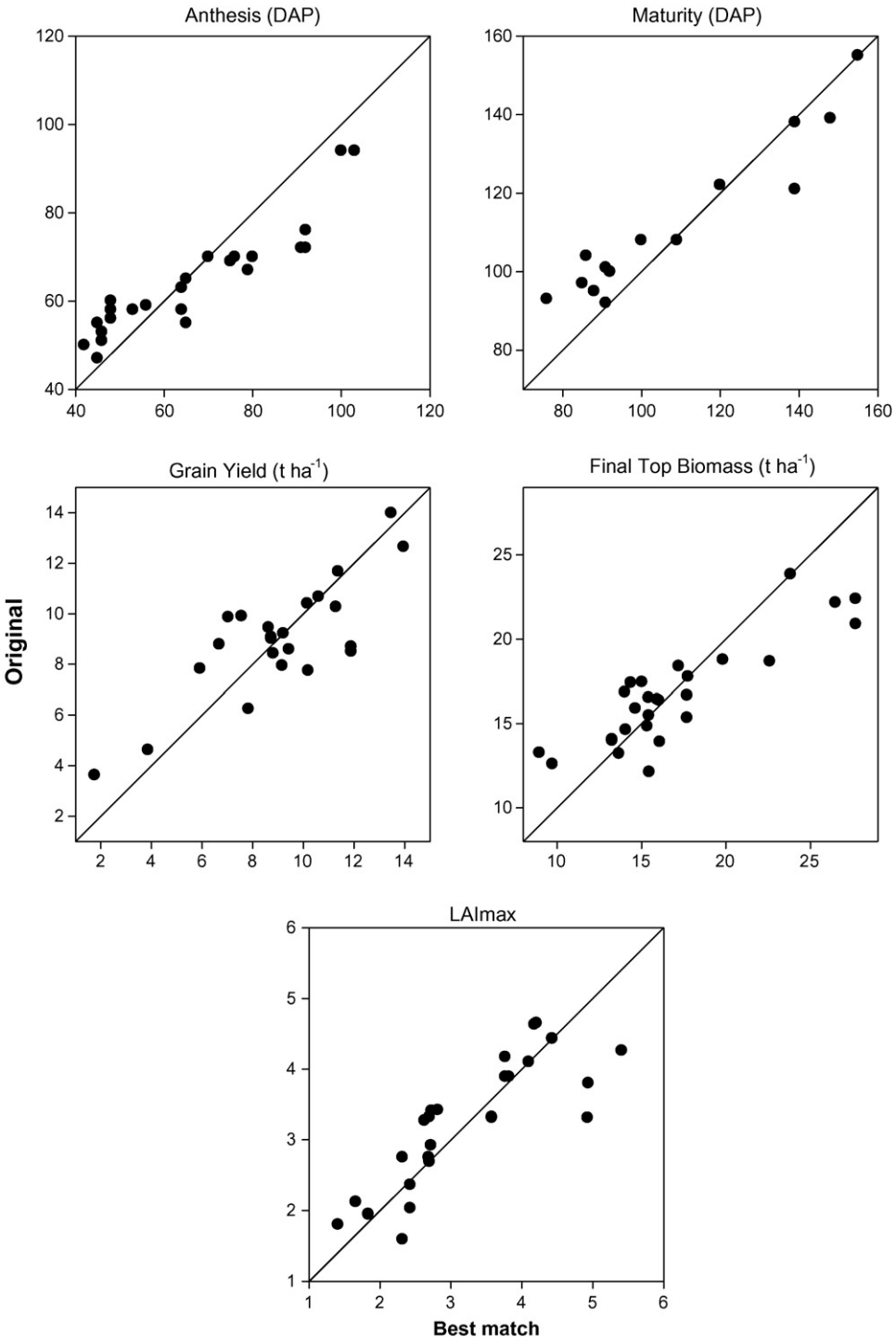

Fig. 2. (*Continued*).

2.2. k-Nearest neighbor approach

The *k*-nearest neighbor approach is an adaptation of standard bootstrap for resampling from time series data (Buishand and Brandsma, 2001). The *k*-nearest neighbor (*k*-NN) method has its origin as a non-parametric statistical pattern recognition procedure, aiming at distinguishing between different patterns according to chosen criteria. Yakowitz (1987) and Karlsson and Yakowitz (1987) constructed a robust theoretical base for the *k*-NN method. Among non-parametric approaches, the *k*-NN approach has shown the most promising and has been applied in various disciplines, including remote sensing (Chi and Bruzzone, 2005), traffic forecasting (Davis and Nihan, 1991), molecular biology (Wu et al., 2005), soil science (Nemes et al., 2006), forest science (LeMay and Hailemariam, 2005) and hydrology (Todini, 2000). The original algorithm has been explained in detail by Brandsma and Buishand (1998), Rajagopalan and Lall (1999) and Gangopadhyay et al. (2005) and was employed here using the following steps:

1- In a typical situation, it is expected that anthesis and maturity dates, maximum leaf area index (LAI_{max}), final above ground biomass, and final grain yield are measured under optimum

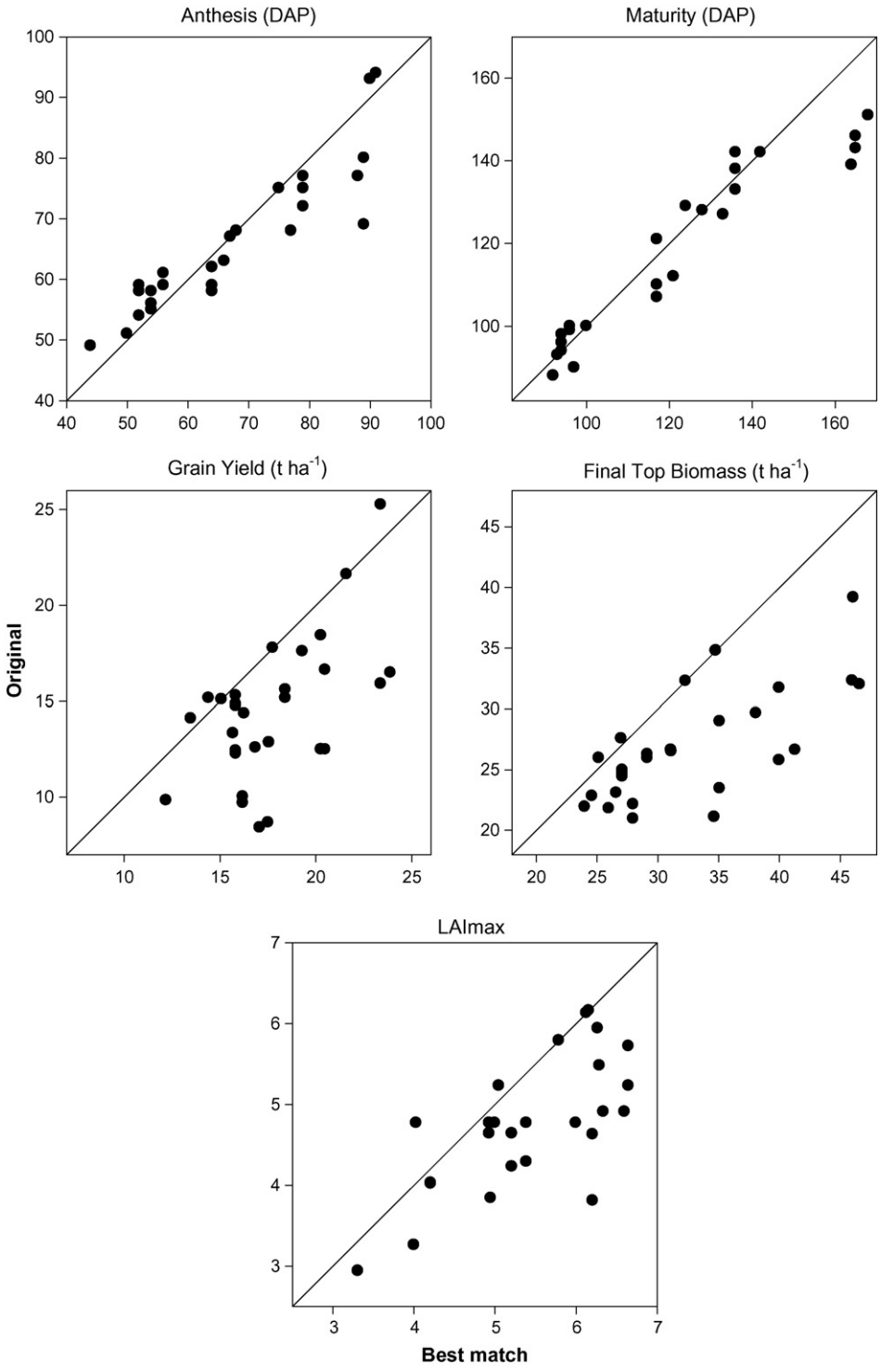

Fig. 2. (*Continued*).

conditions, which is described as the absence of any management and environmental stresses. This set of basic crop data can be considered as the target set of the data or the feature vector. The *k*-NN approach searches the crop characters database which consists of anthesis and maturity dates, LAI_{max}, final above ground biomass and final grain yield and which were constructed based on the simulated data of the CSM-CERES-Maize model. The goal of the search is to find if a cultivar has a similar pattern within *k* nearest neighbors.

2- The number of neighbors (k) practically varies from 1 to $\sqrt{N}$, where N is the total number of cultivars within the crop database. Assuming $k = 1$, for a target cultivar with a set of reported crop characters, e.g. anthesis, maturity dates, LAI_{max}, final above ground biomass and grain yield, our tool calculates the Euclidian distances (d_j) from each cultivar within the crop database. The cultivar that has the lowest d_j would be considered as the most similar cultivar and coefficients associated with this cultivar (within the coefficient database

file) will be attributed as the "*best match*" for the new or target cultivar.

3- The distance (d_j) is calculated for the similarity of the crop characters as

$$d_j = \sqrt{\left[\sum_{j=1}^{d} S_i (V_{ij} - V_{mj})^2\right]} \quad (1)$$

where d_j is Euclidean distances, d is the number of variables, e.g. crop characters, and V_{ij} and V_{mj} are the jth component, such as anthesis date, of the target and ith neighbor, respectively. S_i is the scaling weight for the ith component.

Due to different magnitudes of the crop characters, an approach similar to that of Jagtap et al. (2004) was employed to provide the same influence of all crop variables data on the distance calculation. Therefore, a scaling factor was introduced that determines the relative importance of different crop characters of the observed dataset. The scaling factor for any crop character, e.g. anthesis, was calculated by dividing the maximum range, i.e., the maximum value—the minimum value, among all crop characters divided by the range of each character separately. As example, the range for anthesis is defined as

$$\text{Anthesis} = \frac{\max[(\text{range of anthesis}), (\text{range of maturity}), (\text{range of grain yield}), (\text{range of biomass}), (\text{range of LAI}_{\max})]}{(\text{range of anthesis})}$$

The k-NN approach selects from the Euclidian distances and assigns probability weights to a subset of k distances with smallest to largest Euclidean distance to the target. Euclidean distances, d_j, sorted in ascending order. The weights of the k neighbors are based on their rank distance to the value of the target. The weight function (P_j) was calculated as

$$p_j = \frac{1/j}{\sum_{j=1}^{k} 1/j}, \quad j = 1, \ldots, k \quad (2)$$

where j is the rank of the neighbors in ascending order. The weight function assigns weights to each of the k-nearest neighbors. The neighbor with the shortest distance gets a highest weight, while the neighbor with smallest distance gets the least weight.

Further to Euclidean distance as the measure of similarity, we also employed the cosine similarity as another affinity function (Dong et al., 2006):

$$-\cos(\angle(x,y)) = -\frac{\sum_{i=1}^{5} x_i y_i}{\sqrt{\sum_{i=1}^{5} x_i^2 \sum_{i=1}^{5} y_i^2}} \quad (3)$$

This function calculates similarity of two vectors of data (x and y), containing the crop characters, thus as the two data segments become more similar, their cosine similarity approaches 1.0 and their distances approaches 0.0. In this equation, i is the number of crop characters considered in the study, which in our case is $i = 5$. Therefore, the cosine measure was used as the distance, using the following equation:

$$D(x,y) = 1 - (\text{cosine function}) \quad (4)$$

In contrast to the Euclidean approach, the cosine similarity ignores the magnitude difference between the two vectors.

2.3. Approach implementation

A special program was developed to implement the k-NN approach and embed the required databases (Fig. 1). To run the program, the user should be familiar with the overall structure of DSSAT (Hoogenboom et al., 2004b), the functional approach of the crop simulation models (Jones et al., 1998, 2003) and FileA, which is part of the experimental measurement files of DSSAT (Table 2). The program allows selecting any treatment level directly from any experimental details FileX associated with the experiments defined within DSSAT. The program is flexible to find the required cultivar coefficients, even if the observations are limited, especially for $\text{LAI}_{\max}$.

2.4. Databases construction

Two databases, including the various combinations of cultivar coefficients and simulated crop characters, were constructed. The cultivar coefficients were obtained by using a combination of various values of the CSM-CERES-Maize cultivar coefficients as shown in Table 3. All possible combinations of the coefficient values (Table 3) were used to generate 9263 hypothetical cultivars. Three of the simulation experiments were performed in Gainesville, Florida; Ames, Iowa; and Spain within the DSSAT pool of experiments and two separate experiments including, central Punjab, Pakistan, and in Piracicaba, SP, Brazil were selected. All experiments were modified to provide a potential production environment. The simulation experiment, FileX, of all above locations were run for all 27,789 cultivar coefficients and the simulated crop characteristics were used to construct the crop character database. When different sets of cultivar coefficients resulted in the same model output, including yield, biomass, etc., or if the differences were quite small, then just one set of coefficients together with associated crop characteristics was included in the database and the remainder of the coefficients and associated crop characters were removed from the associated databases.

To evaluate our approach, we used 29 different maize cultivars as reported from field experiments that were available within the DSSAT cultivar database for maize. The model was run for these 29

Table 5
Evaluation measures of crop characteristics obtained with the original reported cultivar coefficients and with those obtained with the *best match* coefficients for Florida (FL), Iowa (IA) and Spain experiments.

Anthesis (planting to anthesis, days after planting)			
Site	r^2	RMSD	RMSDr
Florida	0.98	2.4	0.03
Iowa	0.93	4.1	0.15
Spain	0.92	6.1	0.09
Maturity (planting to physiological maturity, days after planting)			
Florida	0.97	3.7	0.03
Iowa	0.98	3.3	0.03
Spain	0.96	8.8	0.07
Grain yield (t ha^{-1})			
Florida	0.86	1.1	0.15
Iowa	0.86	2.1	0.22
Spain	0.57	4.5	0.31
Aboveground biomass (t ha^{-1})			
Florida	0.91	1.1	0.07
Iowa	0.85	2.7	0.16
Spain	0.69	7.4	0.27
$\text{LAI}_{\max}$ (m^2 m^{-2})			
Florida	0.84	0.6	0.15
Iowa	0.83	0.6	0.19
Spain	0.76	1.0	0.20

cultivars using their reported coefficients for all five sites. The simulated crop characteristics, including anthesis and maturity dates, grain yield, and aboveground biomass were then assumed to be representative as observed. To verify the new coefficients as *best match* in comparison with the coefficients reported originally within DSSAT, the same simulation experiments at each site were run again, but with the new coefficients (*best match*) to provide the crop characters representing the simulated data for the same 29 genotypes. The crop characters simulated with the *best match* were then compared with the crop characters simulated with the original coefficients of these same 29 cultivars for evaluation of our approach. In order to determine if further adjustment were needed of the cultivar coefficients obtained with the *k*-NN method, we evaluated the performance of four cultivars which were not used in development and evaluation of this tool. This included McCurdy 84aa, employed in an experiment that was conducted in Gainesville, Florida in 1982 under high input of water and nitrogen, and PIO × 304C and H610 (UH) which were employed in an experiment conducted in Waipoi, Hawaii in 1983, and AG9010, which was employed in an experiment (Soler et al., 2007a,b) conducted at the "Escola Superior de Agricultura Luiz de Queiroz" of the University of São Paulo, in Piracicaba (−22.7° latitude, −47.4° longitude, 580m elevation above sea level), São Paulo State, Brazil in 2001 and 2002. These data sets are included with the distribution version of DSSAT (Hoogenboom et al., 2004a,b).

The root mean square deviation (RMSD), coefficient of determination (r^2), and relative root mean square deviation (RMSDr) were used for final harvest data and the index of agreement (d), and normalized RMSE were used for time series data to evaluate the accuracy of the simulated crop characteristics based on the new coefficients (*best match*) with the original coefficients reported within the DSSAT maize genotype file. The relative RMSD, denoted as RMSDr, was calculated to be able to compare the RMSD among different characters. RMSD and RMSDr and the normalized root mean square error (RMSE) expressed in percent according to Loague and Green (1991), were calculated as

$$\mathrm{RMSD} = \left(\frac{\sum_{i=1}^{n}(x_i - y_i)^2}{n}\right)^{0.5} \tag{5}$$

$$\mathrm{RMSDr} = \frac{\mathrm{RMSD}}{Y\text{-bar}} \tag{6}$$

$$\mathrm{RMSE} = \frac{\sqrt{\sum_{i-1}^{M}(X_i - O_i)^2}}{n} \times \frac{100}{M} \tag{7}$$

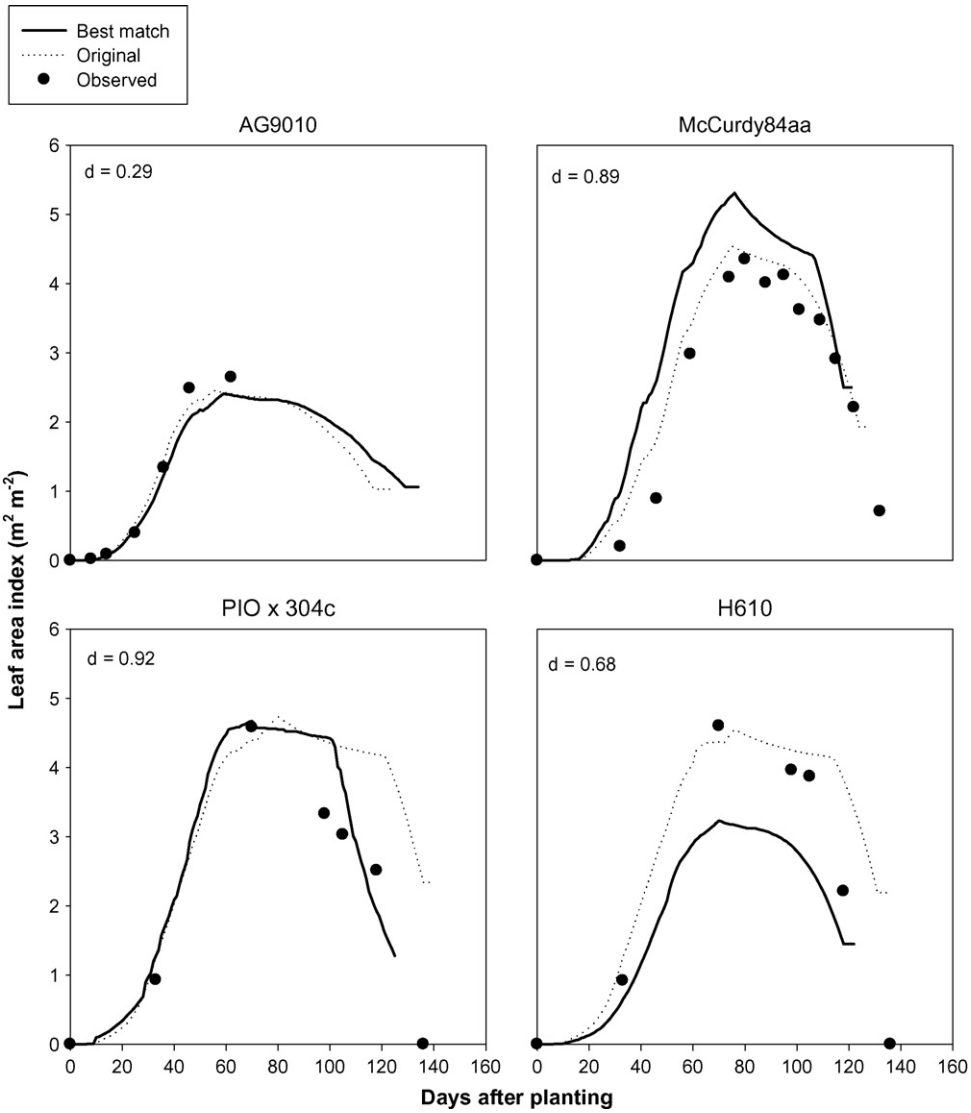

Fig. 3. Comparison of observed anthesis, maturity, LAI, aboveground biomass, and grain weight for four different cultivars against simulated using original and *best match* cultivar coefficients.

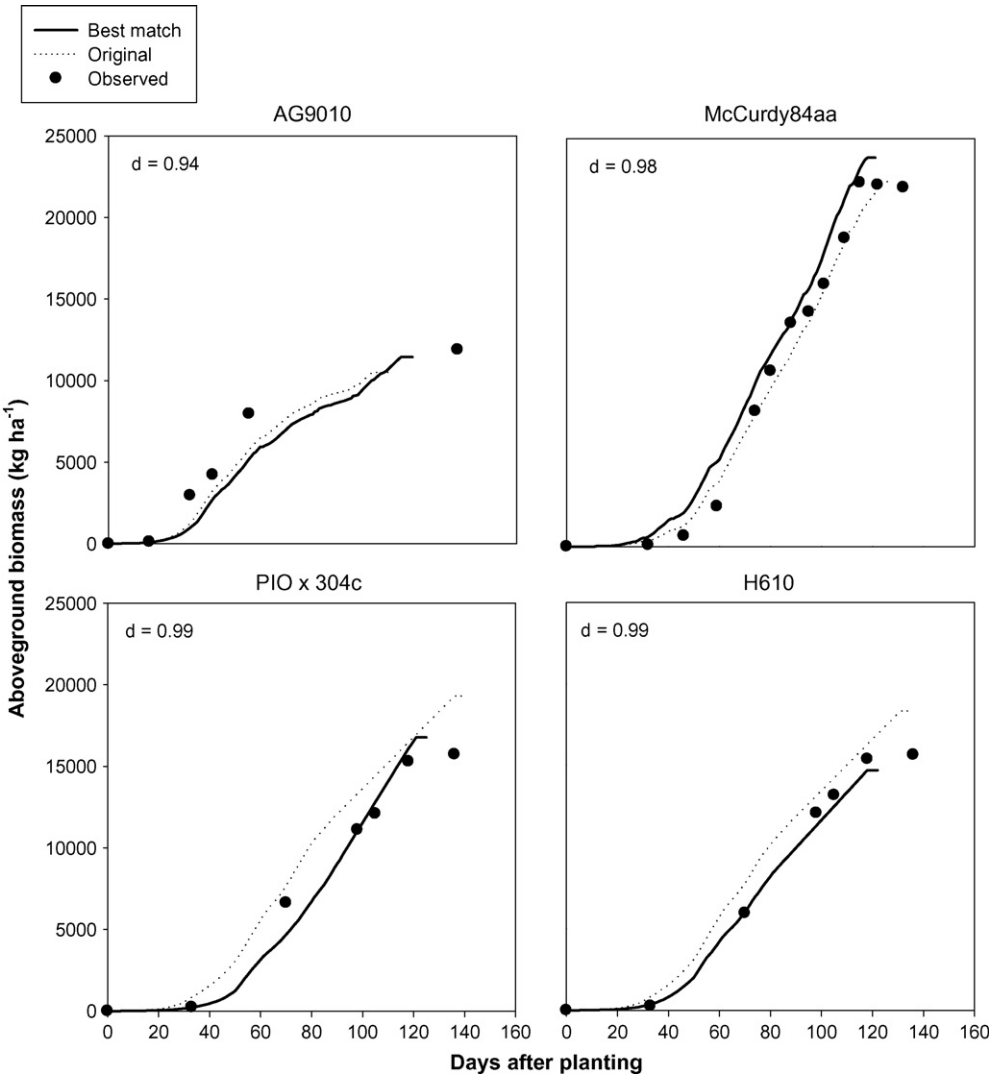

Fig. 3. (*Continued*).

In this approach, n sets simulated based on *best match* (x) and simulated based on original data (y) values were compared on the basis of the root mean squared deviation as the measure of the difference between the two. *Y*-bar is the average of simulated values based on the original data. For RMSE, M is the mean of the observed variable. Normalized RMSE gives a measure (%) of the relative difference of simulated versus observed data. The simulation is considered excellent with a normalized RMSE less than 10%, good if the normalized RMSE is greater than 10 and less than 20%, fair if the normalized RMSE is greater than 20% and less than 30%, and poor if the normalized RMSE is greater than 30% (Jamieson et al., 1991).

The index of agreement (d) was calculated as

$$d = 1 - \left[\frac{\sum_{i=1}^{n}(x_i - y_i)^2}{\sum_{i=1}^{n}(|x_i - \bar{y}| + |y_i - \bar{y}|)^2}\right], \quad 0 \le d \le 1 \tag{8}$$

where n, x_i and y_i are as previously defined and $\bar{y}$ is the average of the observed values. The index of agreement is a descriptive measure of the average relative error. A d value of 1 indicates perfect agreement between model simulations and observations.

3. Results and discussion

The goal of this study was to determine the cultivar coefficients based on the observed crop characteristics. Our initial study showed that the performance of the Euclidean distance approach was better than the Cosine approach. Therefore, all calculated values as *best match* are based on the Euclidean distance method. Both the original reported coefficients of the 29 maize cultivars used in this study and the *best match* obtained based on the *k*-NN approach are shown in Table 4. Although the *best match* coefficients for all three studies sites are in reasonable agreement to the original reported cultivar coefficients, it was not our goal to compare them, but mainly to evaluate their final impact on the simulated crop characteristics. The differences between the values of the original cultivar coefficients versus those obtained as the *best match* from the *k*-NN approach is partially due to the counterbalance effects of various coefficients on each other on the simulated crop characteristics. The CSM-CERES-Maize model simulation of the 29 maize cultivars at three different sites using the simulation results based on *best match* coefficients and based on the original reported coefficients crop characters are shown in Fig. 2. When comparing the two developmental stages, the RMSD values showed a slightly better simulation of the maturity date

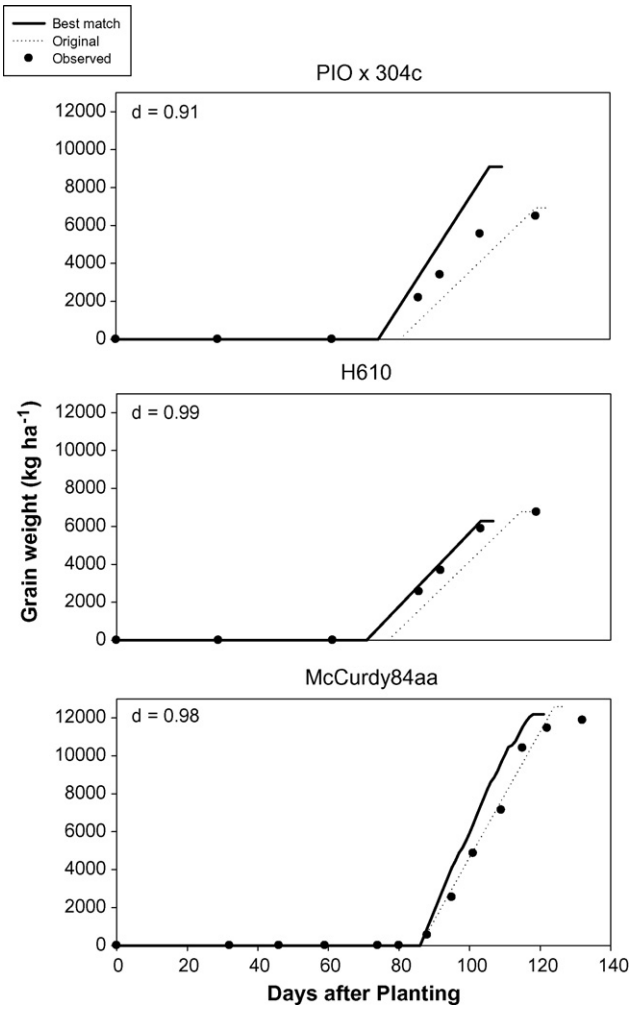

Fig. 3. (*Continued*).

compared to the anthesis date (Table 5). The comparison of the *best match* based simulated aboveground biomass with those simulated based on the original coefficients illustrate an acceptable performance of this approach (Table 5). With about 7% RSMDr obtained for the FL location, we are certain that with a proper set of input data for the crop model, the estimated values of the coefficients are able to simulate a similar final aboveground biomass as when the original coefficients were used for all cultivars. The lowest aboveground biomass (12.1 t ha^{-1}) was obtained for the IA location and the highest aboveground biomass (39.2 t ha^{-1}) for the Spain location. Similar to the original data, the lowest simulated aboveground biomass (8.9 t ha^{-1}) was obtained for the IA location while the highest aboveground biomass (39.2 t ha^{-1}) was obtained for the Spain location. The same ranking was obtained for maximum grain yield when comparing the original and simulated data. The highest original based simulation (21.6 t ha^{-1}) and *best match* based simulation (23.4 t ha^{-1}) grain yield was obtained for the Florida location. However, the lowest original grain yield (3.1 t ha^{-1}) was obtained for the Florida location, while the lowest *best match* based simulated grain yield (1.8 t ha^{-1}) was obtained for the Iowa location. Across all three locations, the cultivar that resulted in the highest grain yield was the same using both the original and *best match* coefficients.

The highest RMSDr was obtained for aboveground biomass and the lowest was RMSDr obtained for anthesis (Table 5). The maximum LAI across all three sites, using the original coefficients, ranged from 1.59 to 6.24 and the RMSDr for maximum LAI was 19%. The coefficient of determination (r^2), RMSDr and RMSD confirmed that the *k*-NN approach provided reliable estimates for the maize cultivar coefficients (Table 5). The highest r^2 (0.98) between the original and *best match* simulated data was obtained for anthesis in FL and maturity in IA, while the lowest r^2 (0.57) was obtained for grain yield in Spain. A Spearman's rank correlation test (Neave and Worthington, 1988) was used to investigate whether the rankings of simulated biomass, grain yield, anthesis, maturity, and LAI_{max} of all study cultivars based on the original coefficients were correlated with rankings of the simulated variables obtained by the *best match* coefficients. A significant goodness of fit was found for all variables with the highest r^2 of 0.96 for anthesis in FL and lowest of 0.49 for LAI_{max} in Spain. The correlation ranks and the direction consistency among the three different study sites suggest that the *k*-NN approach is able to offset any approaches in which their calculated coefficients are not consistent across different sites. The positive rank correlations also indicate that the simulated crop characters of all the study cultivars based on the original coefficients had the same order when compared to those simulated based on the *best match* coefficients.

The CSM-CERES-Maize model was also run for four cultivars which were not used in development and initial evaluation. The environmental data for two of these experiments were used for development of the databases, while the other location represented an independent data set. The model was run with the original values of the four cultivars coefficients, followed by the values of the coefficients obtained from the *k*-NN approach as *best match* (Fig. 3a–c). The results (Table 6) of three of cultivars showed that anthesis either based on original coefficients or *best match* were quite close to the observed data. Observed data of anthesis, maturity and LAI of the other cultivar were not available. The maximum difference between the original coefficients results and observed data were 3 days, while for best match coefficients the maximum difference was 4 days. For maturity, again the original coefficients showed a maximum difference of 4 days. However, for *best match* coefficients the minimum was 5 days and the maximum was 11 days.

The times series comparison of the original and *best match* aboveground biomass, grain yield and leaf area index (Fig. 3a) for all four cultivars showed a good similarity for all three variables. For LAI, *d* for the simulation based on the *best match* coefficients for two cultivars was either equal or smaller than for the LAI using the original coefficients. The normalized RMSE for one cultivar was lower (Table 6) using the *best match* coefficients compared to the original coefficients values. For aboveground biomass, the results based on the *best match* coefficients had a higher agreement with the observed data for two cultivars compared to when the original coefficients were used to run the crop simulation model (Fig. 3b, Table 6). For grain yield, one cultivar had better results compared to the original values. Our results showed that the *best match* coefficients values obtained with this approach would be definitely much closer to the real values of the cultivar coefficients than any other initial guess. The index of agreement and normalized RMSE values indicated a level of accuracy that would be sufficient for the users of the model who only have access to one year of experimental data and require the best possible initial estimate for the cultivar coefficients.

Crop simulation models such as the CSM-CERES-Maize model use cultivar coefficients to simulate various growth and development aspects of different maize plant cultivars in response to environment and management factors. Using incorrect cultivar coefficients could lead to erroneous crop model predictions that are either too low or too high. For instance, an incorrect response of the plant to temperature and photoperiod due to incorrect cultivar

Table 6
Comparison of crop characters of three cultivars of an independent experiment obtained as observed, simulated based on original coefficients, and simulated based on best match found coefficients.

Cultivar	Anthesis (DAP)			Maturity (DAP)			LAI_{max} (m^2 m^{-2})			Aboveground biomass (t ha^{-1})			Grain yield (t ha^{-1})		
	Obs	Original	Best match	Obs	Original	Best match	Obs	Original	Best match	Obs	Original	Best match	Obs	Original	Best match
McCurdy	75	75	79	128	126	123	4.09	4.42	5.52	22.0	21.2	24.4	11.9	11.8	11.9
PIO × 304C	78	81	75	136	140	125	4.58	4.73	4.67	15.7	19.3	16.7	6.5	6.9	9.1
H610	74	77	71	133	135	122	4.60	4.55	3.23	15.6	18.3	14.7	6.8	6.8	6.3

RMSE

LAI (m^2 m^{-2})	Original	Best match
McCurdy	19.0	28.9
PIO × 304C	86.3	36.1
H610	37.4	46.7
Aboveground biomass (t ha^{-1})		
McCurdy	8.2	12.3
PIO × 304C	22.1	10.6
H610	15.8	6.6
Grain yield (t ha^{-1})		
McCurdy	10.0	15.9
PIO × 304C	35.8	64.5
H610	25.4	11.3

coefficients would result in an incorrect vegetative and reproductive development durations (Bannayan et al., 2004). Incorrect phenology results in incorrect canopy growth, which subsequently could result in an over or under estimation of final yield.

The reasonable agreement we found between the crop characters that were simulated based on the reported original cultivar coefficients and based on the *best match* coefficients, showed the reliability of *k*-NN approach for determining the coefficients of any maize cultivar with the least available crop data. We believe that the *best match* coefficients provided by *k*-NN approach would best describe any new cultivar within the CSM-CERES-Maize model. The analog principle used in this study is a potential improvement in greater insight, but it should not be treated as firm evidence, as each cultivar is different. However, to cope with this diversity of the individual cultivars requires an approach that is able to identify the similarity of classes of attributes, such as the *k*-NN approach. This approach verification requires more independent dataset from various other locations and it would be more helpful if one can obtain similar accuracy while less observed crop characters would be available.

Acknowledgements

This work was conducted under the auspices of the Southeast Climate Consortium (SECC; www.SEClimate.org) and supported by a partnership with the United States Department of Agriculture-Risk Management Agency (USDA-RMA), by grants from the US National Oceanic and Atmospheric Administration-Climate Program Office (NOAA-CPO) and USDA-Cooperative State Research, Education and Extension Services (USDA-CSREES) and by State and Federal funds allocated to Georgia Agricultural Experiment Station Hatch project GEO01654.

References

Azam-Ali, S.N., Aguilar-Manjarrez, J., Bannayan, M., 2001. A Global Mapping System for Bambara Groundnut (*Vigna subterranea* L. Verdc) Production. FAO Agric. Information Management Series.

Bannayan, M., Hoogenboom, G., 2008a. Daily weather sequence prediction realization using the non-parametric nearest-neighbor re-sampling technique. Int. J. Climatol. 28 (10), 1357–1368.

Bannayan, M., Hoogenboom, G., 2008b. Weather analogue: a tool for real-time prediction of daily weather data realizations based on a modified k-nearest neighbor approach. Environ. Modell. Softw. 3, 703–713.

Bannayan, M., Kobayashi, K., Marashi, H., Hoogenboom, G., 2007. Gene-based modeling for rice: an opportunity to enhance the simulation of rice growth and development? J. Theor. Biol. 249, 593–605.

Bannayan, M., Hoogenboom, G., Crout, N.M.J., 2004. Photothermal impact on maize performance: a simulation approach. Ecol. Model. 180, 277–290.

Bannayan, M., Kobayashi, K., Kim, H.Y., Liffering, M., Okada, M., Miura, S., 2005. Modeling the interactive effects of atmospheric CO_2 and N on rice growth and yield. Field Crop Res. 93, 237–251.

Banterng, P., Patanothai, A., Pannangpetch, K., Jogloy, S., Hoogenboom, G., 2004. Determination of genetic coefficients of peanut lines for breeding applications. Eur. J. Agron. 21 (3), 297–310.

Banterng, P., Patanothai, A., Pannangpetch, K., Jogloy, S., Hoogenboom, G., 2006. Yield stability evaluation of peanut lines: a comparison of an experimental versus a simulation approach. Field Crops Res. 96 (1), 168–175.

Boote, K.J., Jones, J.W., Batchleor, W.D., Nafziger, E.D., Myers, O., 2003. Genetic coefficients in the cropgro-soybean model: links to field performance and genomics. Agron. J. 95, 32–51.

Boote, K.J., Jones, J.W., Hoogenboom, G., Pickering, N.B., 1998. The CROPGRO model for grain legumes. In: Tsuji, G.Y., et al. (Eds.), Understanding Options for Agricultural Production. Kluwer Academic Publisher, Dordrecht, The Netherlands, pp. 99–128.

Brandsma, T., Buishand, T.A., 1998. Simulation of extreme precipitation in the Rhine basin by nearest neighbor resampling. Hydro. Earth Syst. Sci. 2, 195–209.

Brooking, I.R., 1996. Temperature response of vernalization in wheat: a developmental analysis. Ann. Bot. 78, 507–512.

Buishand, T.A., Brandsma, T., 2001. Multisite simulation of daily precipitation and temperature in the Rhine basin by nearest-neighbor resampling. Water Resour. Res. 37 (11), 2761–2776.

Chi, M., Bruzzone, L., 2005. An ensemble-driven k-NN approach to ill-posed classification problems. Pattern Recog. Lett. 27, 301–307.

Chiu, N.H., Huang, S.J., 2007. The adjusted analogy-based software effort estimation based on similarity distances. J. Syst. Software 80, 628–640.

Davis, G.A., Nihan, N.L., 1991. Nonparametric regression and short-term freeway traffic forecasting. J. Transp. Eng.-ASCE. 117 (2), 178–188.

Dong, Y., Sun, Z., Jia, H., 2006. A cosine similarity-based negative selection algorithm for time series novelty detection. Mech. Syst. Signal Pr. 20, 1461–1472.

Fowler, D.B., Limin, A.E., Wang, S.Y., Ward, R.W., 1996. Relationship between low-temperature tolerance and vernalization response in wheat and rye. Can. J. Plant Sci. 76, 37–42.

Gangopadhyay, S., Clark, M., Rajagopalan, B., 2005. Statistical downscaling using *K*-nearest neighbors. Water Resour. Res. 41, 1–23.

Godwin, D.C., Jones, C.A., 1991. Nitrogen dynamics in soil-plant system. In: Hanks, R.J., Ritchie, J.T., (Eds.), Modelling Plant and Soil System, Monograph No. 31. American Society of Agronomy, Madison, WI, pp. 287–321.

Goswami, U., 1995. Phonological development and reading by analogy: what is analogy, and what is it not? J. Res. Read. 18 (2), 139–145.

Gungula, D.T., Kling, J.G., Togun, A.O., 2003. CERES-Maize predictions of maize phenology under nitrogen-stressed conditions in Nigeria. Agron. J. 95, 892–899.

Hammer, G.L., Kropff, M.J., Sinclair, T.R., Porter, J.R., 2002. Future contributions of crop modeling-from heuristics and supporting decision making to understanding genetic regulation and aiding crop improvement. Eur. J. Agron. 18, 15–31.

Hoogenboom, G., White, J.W., Jones, J.W., Boote, K.J., 1994a. BEANGRO: a process-oriented dry bean model with a versatile user interface. Agron. J. 86, 82–190.

Hoogenboom, G., Jones, J.W., Wilkens, P.W., Batchelor, W.D., Bowen, W.T., Hunt, L.A., Pickering, N.B., Singh, U., Godwin, D.C., Baer, B., Boote, K.J., Ritchie, J.T., White, J.W., 1994b. Crop models.. In: Tsuji, G.Y., Uehara, G., Balas, S. (Eds.), DSSAT, version 3, vol. 2. University of Hawaii, Honolulu, Hawaii, pp. 95–244.

. Hoogenboom, G., Wilkens, P.W., Tsuji, G.Y. (Eds.), 1999. DSSAT, Version 3, vol. 4. University of Hawaii, Honolulu, Hawaii.

Hoogenboom, G., White, J.W., Messina, C.D., 2004a. From genome to crop: integration through simulation modeling. Field Crops Res. 90 (1), 145–163.

Hoogenboom, G., Jones, J.W., Wilkens, P.W., Porter, P.W., Batchelor, C.H., Hunt, L.A., Boote, K.J., Singh, U., Uryasev, O., Bowen, W.T., Gijsman, A.J., du Toit, A., White, J.W., Tsuji, G.Y., 2004b. Decision Support System for Agrotechnology Transfer Version 4. 0 [CD-ROM] University of Hawaii, Honolulu, HI.

Hunt, L.A., Pararajasingham, S., Jones, J.W., Hoogenboom, G., Imamura, D.T., Ogoshi, R.M., 1993. GENCALC—software to facilitate the use of crop models for analyzing field experiments. Agron. J. 85 (5), 1090–1094.

Hunt, L.A., Pararajasingham, S., 1994. Genotype coefficient calculator. In: Tsuji, G.Y., Uehara, G., Baka, S. (Eds.), DSSAT v3, vols. 3–4. University of Hawaii, Honolulu, pp. 203–233.

Jagtap, S.S., Abamu, F.J., 2003. Matching improved maize production technologies to the resource base of farmers in a moist savanna. Agric. Syst. 76, 1067–1084.

Jagtap, S.S., Lall, U., Jones, J.W., Gismn, A.J., Ritchie, J.T., 2004. Dynamic nearest neighbor method for estimating soil water parameters. Trans. ASAE 47 (5), 1437–1444.

Jamieson, P.D., Porter, J.R., Wilson, D.R., 1991. A test of the computer simulation model ARC-WHEAT1 on wheat crops grown in New Zealand. Field Crops Res. 27, 337–350.

Jones, C.A., Kiniry, J.R., 1986. CERES-Maize. A Simulation Model of Maize Growth and Development. Texas A&M University Press, College Station, TX, 194 pp.

Jones, J.W., Tsuji, G.Y., Hoogenboom, G., Hunt, L.A., Thornton, P.K., Wilkens, P.W., Imamura, D.T., Bowen, W.T., Singh, U., 1998. Decision support system for agrotechnology transfer; DSSAT v3. In: Tsuji, G.Y., Hoogenboom, G., Thornton, P.K. (Eds.), Understanding Options for Agricultural Production. Systems Approaches for Sustainable Agricultural Development. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 157–177.

Jones, J.W., Hoogenboom, G., Porter, C.H., Boote, K.J., Batchelor, W.D., Hunt, L.A., Wilkens, P.W., Singh, U., Gijsman, A.J., Ritchie, J.T., 2003. DSSAT cropping system model. Eur. J. Agron. 18, 235–265.

Karlsson, M., Yakowitz, S., 1987. Nearest-neighbor methods for nonparametric rainfall–runoff forecasting. Water Resour. Res. 23 (7), 1300–1308.

Kropff, M.J., Cassman, K.G., Van Laar, H.H., 1994. Quantitative understanding of the irrigated rice ecosystems and yield potential. In: Virmani, S.S. (Ed.), Hybrid Rice Technology: New Developments and Future Prospects. IRRI, Los Banos, Philippines, pp. 97–114.

LeMay, V., Hailemariam, T., 2005. Comparison of nearest neighbor methods for estimating basal area and stems per hectare using aerial auxiliary variables. Forest Sci. 51 (2), 109–119.

Loague, K., Green, R.E., 1991. Statistical and graphical methods for evaluating solute transport models: overview and application. J. Contam. Hydrol. 7, 51–73.

Loomis, R.S., Rabbinge, R., Ng, E., 1979. Explanatory models in crop physiology. Ann. Rev. Plant Physiol. 30, 339–367.

Mavromatis, T., Hansen, J.W., 2001. Interannual variability characteristics and simulated crop response of four stochastic weather generators. Agric. For. Meteorol. 109, 283–296.

Neave, H.R., Worthington, P.L., 1988. Distribution Free Tests. Unwin Hyman Publisher, 430 pp.

Nemes, A., Rawls, W.J., Pachepsky, Y.A., 2006. Use of the nonparametric nearest neighbor approach to estimate soil hydraulic properties. Soil Sci. Soc. Am. J. 70, 327–336.

Rajagopalan, B., Lall, U., 1999. A *k*-nearest neighbor simulator for daily precipitation and other variables. Water Resour. Res. 35 (10), 3089–3101.

Ritchie, J.T., Singh, U., Godwin, D.C., Bowen, W.T., 1998. Cereal growth, development and yield. In: Tsuji, G.Y., Hoogenboom, G., Thornton, P.K. (Eds.), Understanding Options for Agricultural Production. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 79–98.

Soler, C.M.T., Hoogenboom, G., Sentelhas, P.C., Duarte, A.P., 2007a. Impact of water stress on maize grown off-season in a subtropical environment. J. Agron. Crop Sci. 193 (4), 247–261.

Soler, C.M.T., Sentelhas, P.C., Hoogenboom, G., 2007b. Application of the CSM-CERES-Maize model for planting date evaluation and yield forecasting for maize grown off-season in a subtropical environment. Eur. J. Agron. 27 (2), 165–177.

Suriharn, B., Patanothai, A., Pannangpetch, K., Jogloy, S., Hoogenboom, G., 2007. Determination of cultivar coefficients of peanut lines for breeding applications of the CSM-CROPGRO-Peanut model. Crop Sci. 47, 607–619.

Todini, E., 2000. Real-time flood forecasting: operational experience and recent advances. In: Marsalek, J., et al. (Eds.), Flood Issues in Contemporary Water Management. Kluwer Academic Publisher, The Netherlands, pp. 261–270.

Tsuji, G.Y., Hoogenboom, G., Thornton, P.K. (Eds.), 1998. Understanding Options for Agricultural Production. Systems Approaches for Sustainable Agricultural Development. Kluwer Academic Publishers, Dordrecht, The Netherlands, 400 pp.

White, J.W., Hoogenboom, G., 1996. Integrating effects of genes for physiological traits into crop growth models. Agron. J. 88, 416–422.

Wu, W., Xing, E.P., Myers, C., Mian, I.S., Bissell, M.J., 2005. Evaluation of normalization methods for cDNA microarray data by k-NN classification. Bioinformatics 6, 191.

Yakowitz, S., 1987. Nearest neighbor method for time series analysis. J. Time Ser. Anal. 8 (2), 235–247.

Yan, W., Hunt, L.A., 1999a. An equation for modelling the temperature response of plants using only the cardinal temperatures. Ann. Bot. (London) 84, 607–614.

Yan, W., Hunt, L.A., 1999b. Reanalysis of vernalization data of wheat and carrot. Ann. Bot. (London) 84, 615–619.

Xue, Q., Weiss, A., Baenziger, P.S., 2004. Predicting phenological development in winter wheat. Clim. Res. 25, 243–252.