Published online February 6, 2007

Regression-Based Evaluation of Ecophysiological Models

Jeffrey W. White,* Kenneth J. Boote, Gerrit Hoogenboom, and Peter G. Jones

ABSTRACT

Ecophysiological models are increasingly used as research and decision support tools in agriculture, but it is often difficult to assess how suitable a model is for a particular application. Model evaluations usually involve bivariate linear regression between observed and simulated values, which assumes statistical independence among observed values. However, observed data often have dependencies if they originate from series of experiments or involve experiments using nested designs (e.g., with split plots). By representing experiments, cultivars, or other variables as factors, linear regression models can specify expected dependencies, permitting analyses that are statistically more rigorous and provide more insights into model performance. This study evaluated the Cropping System Model (CSM)-CROPGRO-Soybean model using regressions that included environment and cultivars as factors as well as continuous variables such as temperature or daylength. When applied to 28 data sets for soybean [*Glycine max* (L.) Merr.], representing 113 treatment combinations, the regressions showed that the model simulated days to anthesis and grain yield well for a wide range of environments. Differences among environments represented a larger portion of unexplained variation than did differences among cultivars. Further improvements thus might be sought in modeling crop response to environment rather than in representing cultivar differences, or alternatively, in characterizing soil profiles or daily weather rather than cultivars. A submodel for photosynthesis that scaled leaf-level values to canopy simulated grain yield more accurately than a simpler submodel. Multiple regressions provided much more information on model performance than simple bivariate comparisons.

Ecophysiological models of crop species quantify and integrate hypotheses about how crops respond to the environment and specifically to different management scenarios and environmental conditions. These simulation models can provide explicit predictions of crop production and impacts on natural resources. They also can guide research on more fundamental questions such as how genetic alterations to partitioning affect crop growth or whether increased atmospheric CO_2 concentration will improve crop water and N use efficiency. Applications of crop simulation models include research on global warming (e.g., Mearns et al., 1999; Alexandrov and Hoogenboom, 2000; Jones and Thornton, 2003), crop response to sowing dates (Egli and Bruening, 1992; Acosta-Gallegos and White, 1995) and irrigation (Hood et al., 1987), characterizations of production experiments (Chapman et al., 2000), regional targeting of technologies (Hartkamp et al., 2004), plant breeding (White, 1998; Mavromatis et al., 2001, 2002; Banterng et al., 2004) and yield forecasting (Bannayan et al., 2003).

Increasingly, crop models are applied outside of the range of conditions for which they were developed and tested, especially in global change research, and new users may have little prior experience with underlying assumptions, features or limitations of the models. Model evaluation procedures can inform users of the likely accuracy of a given model as well as guide further experimental work and model revisions. Basic evaluations usually involve comparisons of observed vs. simulated data for variables like days to anthesis, grain yield, harvest index, and unit grain weight (e.g., Hoogenboom et al., 1997; Sau et al., 1999). The correlations, slope, intercept, or allied indices (e.g., Willmott's d-statistic; Willmott, 1982) are underlain by bivariate linear regressions that assume that observations are statistically independent of one another (Mayer et al., 1994; Kleinbaum and Kupper, 1978). When considering sets of observed data assembled from multiple locations or experiments with any type of nested (hierarchical) treatment structure, the assumption of independence of observed values is likely incorrect. Linear regression models, however, can be extended to incorporate information on treatment structures, especially those related to locations and cultivars. Such extensions are occasionally found in crop modeling literature (e.g., Hoogenboom et al., 1997), but more often, researchers either ignore lack of dependence or resort to subdividing data sets, which sacrifices statistical power and makes it harder to interpret the overall results.

Residuals from regressions of observed vs. simulated data should always be inspected (Kleinbaum and Kupper, 1978) for underlying patterns. Besides problems of independence (e.g., evidenced by residuals clustering by experiments or treatments), residuals can be analyzed for relations with factors not considered in the simulation model (e.g., pest damage or lodging) or associations with specific environmental characteristics, which might suggest problems with input data or the simulation model (e.g., du Toit and du Toit, 2003). Such relations also can be quantified with multiple regressions, providing a further example of the utility of regression analysis in model evaluation.

One may ask why validation procedures seldom consider independence among observations. One answer is that researchers have underestimated the biases that a lack of independence can introduce in evaluating simulations. A second reason is that when lack of independence manifests itself, the variation is viewed as a problem rather than as a pattern that can be analyzed, potentially providing further insights into model behavior. A third

J.W. White, USDA-ARS, Arid Land Agricultural Research Center, 21881 N Cardon Ln., Maricopa, AZ 85239; K.J. Boote, Agronomy Dep., P.O. Box 110500, Gainesville, FL 32611; G. Hoogenboom, Dep. of Biological and Agricultural Engineering, Univ. of Georgia, Griffin, GA 30223-1797; and P.G. Jones, CIAT, Apartado Aéreo 6713, Cali, Colombia. Received 3 Apr. 2006. *Corresponding author (jwhite@uswcl.ars.ag.gov).

Published in Agron. J. 99:419–427 (2007).
Modeling
doi:10.2134/agronj2006.0100

Abbreviations: CSM, Cropping System Model; DF, degrees of freedom; DSSAT, Decision Support System for Agrotechnology Transfer; NS, not significant.

reason concerns whether regression and analysis of variance (ANOVA) are viewed as discrete forms of analysis. The dominant view is that regression quantifies relations among vectors of numeric data, while ANOVA compares treatment means by partitioning variances. The alternate perspective is that linear regression models can include combinations of continuous and discrete variables, while the ANOVA describes how variation is explained by the different independent variables in the linear model. These two perspectives are easily recognized in different books on statistical analysis (e.g., Gomez and Gomez, 1984 vs. Kleinbaum and Kupper, 1978) and in statistical packages. For the latter, one difference is in how data are represented. In the SAS package (SAS Institute, Cary, NC), all data exist as variates, and it is only in specifying a linear model that variables having discrete levels are identified with a Class statement. In contrast, GenStat (VSN International Ltd., Hemel Hempstead, Herts., UK) and the R Project (www.r-project.org) recognize two main data classes, vectors and factors. Factors have discrete levels, and may or may not have level identifiers. When a factor is used in a linear regression model, the software creates a set of contrasts corresponding to the factor levels. Thus, the two perspectives result in numerically identical results regardless of whether vectors or factors are considered.

Our objective is to investigate how linear regressions can be used to test the responses of models to environments, cultivars, or other variables. Besides comparisons of observed and simulated data, we also illustrate use of linear models to assess the influence of alternate explanatory variables and to compare different models. We emphasize that the procedures discussed are meant to stimulate critical thinking on how a given model performs. No procedure can provide conclusive proof that a model is valid.

MATERIALS AND METHODS

Evaluation Data Sets

Twenty-eight sets of data from field experiments were obtained from the Decision Support Systems for Agrotechnology (DSSAT) Version 4 software package (Hoogenboom et al., 2004). Several of the data sets can also be obtained from the ICASA (International Consortium for Agricultural Systems Applications) Data Exchange (Bostick et al., 2004; www.ICASA.net). The experiments included 17 cultivars (Table 1) grown under various management practices in 11 sites (Table 2), providing 113 treatment combinations. Figure 1 summarizes temperature, solar radiation, daylength, and rainfall conditions during the experiments. Each data set included information on the soil profile, daily weather conditions during the growing season, and management practices for the experiment in question. Observed data considered were days to anthesis, grain yield (dry grain basis), and total aboveground biomass at harvest maturity (dry weight basis). Most of the data sets were used in the development and calibration of the CSM-CROPGRO-Soybean model (Jones et al., 2003), so the observed data do not represent a completely independent evaluation data set. However, regression is equally applicable to calibration and evaluation data.

Simulations Using CSM-CROPGRO-Soybean

All simulations were conducted with the CSM-CROPGRO-Soybean model Version 4.0.2.0 (Jones et al., 2003; Hoogenboom et al., 2004), which was developed from SOYGRO (Jones et al., 1989) and CROPGRO (Hoogenboom et al., 1992; Boote et al., 1998a). Photosynthesis and phenology are described in separate routines that use hourly time steps. Carbohydrate partitioning, growth, and the water and N balances are simulated with daily steps. The hedgerow photosynthesis submodel, which is the default option and predicts leaf photosynthesis for sunlit and shaded leaves in a canopy (Boote and Pickering, 1994), was used for most of the analyses, but a simpler canopy submodel (Boote et al., 1998b) was also tested to illustrate how models may be compared. The CSM-CROPGRO-Soybean simulates development by integrating phase-specific developmental rates over physiological time, which includes phase-specific temperature and photoperiod effects.

Simulations for individual experiments assumed field conditions and management as reported in the input files that specified experimental details for each experiment. Cultivar-specific parameters were used as provided with the soybean model in DSSAT V4. Full water and N balances, including symbiotic N fixation, were simulated (Godwin and Singh, 1998; Ritchie, 1998). No effects of pests or weeds were considered, and it was assumed that the experiments were otherwise well managed.

Table 1. Characteristics of the various soybean cultivars that were simulated with the CSM-CROPGRO-Soybean model.

Cultivar	Maturity group	Stem termination	Seed weight	Number of occurrences†	Geographic origin or source company
			mg seed^{-1}		
Altona	00	indeterminate	170	2	Canada
Bragg	7	determinate	150	24	FL
Centennial	6	determinate	90	14	MS
Chandor	0	determinate	180	4	France
Cobb	8	determinate	140	23	FL
Coker 6847	7	determinate	100	2	Northrup King
Elgin-87	2	indeterminate	150	2	IA
Evans	0	indeterminate	160	2	MN
Forrest	5	determinate	120	14	MS
Labrador	00	indeterminate	150	14	France
Leflore	6	determinate	160	1	MS
Major	000	indeterminate	NA‡	14	France
Maple Arrow	00	indeterminate	170	2	Canada
McCall	00	indeterminate	160	2	MN
Ransom	7	determinate	140	14	NC
Wayne	3	indeterminate	160	10	IL
Williams-82	3	indeterminate	150	10	IL

† The total number of times a given cultivar appears in the field experiments.
‡ NA, value for seed weight not available.

Table 2. Summary of locations used as sources of data for the CSM-CROPGRO-Soybean model.

Location	Latitude	Longitude	Elevation	Total number of experiments	Total number of treatments	Treatments
			m			
Ames, IA	42.00	−93.77	320	2	4	cultivars
Castana, IA	42.07	−95.91	360	1	4	row spacing, irrigation
Clayton, NC†	35.65	−78.46	100	1	18	cultivars × planting dates
Elora, ON, Canada†	43.39	−80.25	380	2	6	cultivars
Gainesville, FL	29.63	−82.37	24	8	16	irrigation, row spacing
Greenwood, FL	30.79	−85.25	35	2	3	irrigation
Lugo, Spain‡	43.00	6.20	480	7	32	cultivars, planting dates, irrigation
Morris, MN	45.57	95.97	340	1	2	irrigation
Quincy, FL	30.60	−86.40	70	1	2	defoliation
Rocky Mount, NC†	35.66	−77.89	50	1	24	cultivars × planting dates
Wooster, OH	40.78	−81.93	310	2	2	cultivars

† Data not used for calibration.
‡ Data used for calibration of model but with the assumption of a 2°C downward shift in the base temperature for photosynthesis.

Regression Analysis

Linear regression models with ordinary least squares estimation were used to test the effects of independent variables. All regressions were conducted using the GLM procedure of SAS Version 8 (SAS Institute, Cary, NC, USA). For qualitative variables such as location and cultivar, the CLASS statement was used, which is equivalent to creating a set of dummy variables or orthogonal contrasts.

Each linear model was related to a specific proposition concerning performance of the simulation model, such as whether the simulations effectively accounted for variability due to location or genotype. In most cases, the proposition involved a comparison of sequential effects of independent variables, so the relevant tests were based on sequential changes in sums of squares using the residual sums of squares as the error term, remembering that it was the null hypothesis (H_0) that was being tested. This stated that there was no effect other than random variation for the factor or variates being tested. This perspective sometimes creates confusion. With large datasets, quite small effects can be significantly different from zero. Thus, it is much more important to consider the portion of the variation accounted for by the effect than the actual significance level. Unfortunately, there is no simple test of probability for assessing relative magnitudes of variances.

Examples of the linear models used are outlined below. The simplest case is the traditional comparison of vectors of an observed (y) vs. (x) simulated response, using a bivariate linear model,

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad [1]$$

where β_0 is the intercept, β_1 is the slope, and ε is the error or residual component of the model. An effect of location, year, cultivar, or other factor is introduced using a variable z,

$$y = \beta_0 + \beta_1 x + \beta_2 z + \varepsilon \qquad [2]$$

Since z has discrete levels, the statistical software creates contrasts to represent the factor levels. If the factor has n levels, the degrees of freedom (DF) associated with the effect of z will be $n - 1$. Equation [2] is equivalent to fitting n lines with constant slope but different intercepts. The interaction of x and z is easily incorporated,

$$y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 zx + \varepsilon \qquad [3]$$

and is interpretable as a test for differences in slopes. The equation can include additional variates or factors,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 z + \varepsilon \qquad [4]$$

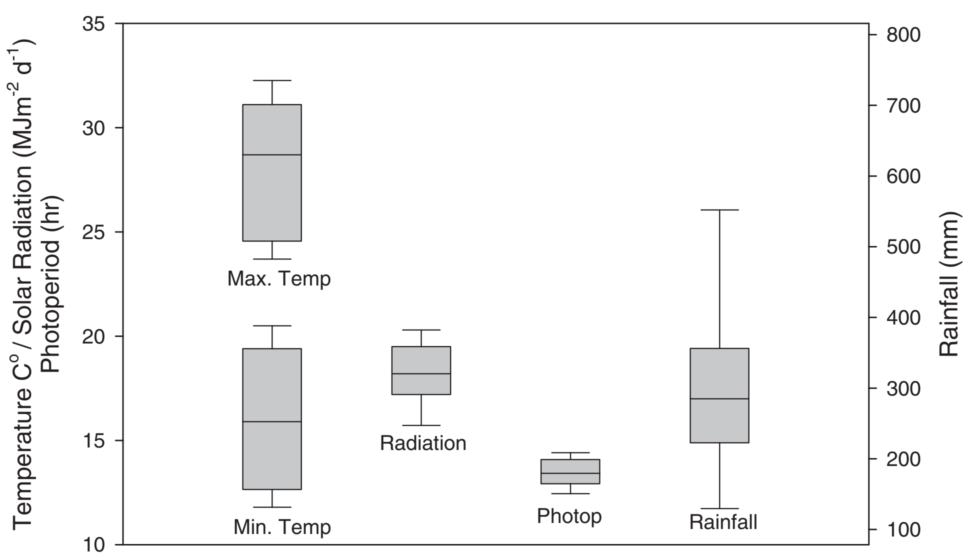

Fig. 1. Box-and-whisker plots for mean temperatures, daylength, solar radiation and total season precipitation for the 28 experiments from 11 locations considered in the simulations with the CSM-CROPGRO-Soybean model and the regressions. The lower and upper limits of the boxes indicate the first and third quartiles, and the T-bars indicate the minimum and maximum values.

where x_2 is a second vector. Depending on the proposition being examined, a factor can be introduced before a vector. We emphasize that these are merely examples of equations that might be used. More extensive discussions are found in texts on regression analyses (e.g., Kleinbaum and Kupper, 1978; Maindonald and Braun, 2003).

For the bivariate regressions (Eq. [1]), various validation statistics are provided to facilitate interpretation. These include the mean absolute difference, root mean square error, slope, intercept, the square of the correlation coefficient (r^2), and Willmott's Index of Agreement or d-statistic (Willmott, 1982). The d-statistic is calculated as,

$$d = 1 - \Sigma(y_i - x_i)^2/\Sigma(|x_i - \overline{Y}| + |y_i - \overline{Y}|)^2 \quad [5]$$

where $\overline{Y}$ is the mean of the observed values. Similar to r^2, it is scaled from 0 to 1 but unlike r^2, the d-statistic is influenced by the bias between observed and simulated values and by the variation in the observed values.

RESULTS AND DISCUSSION

Proposition 1: The Simulation Explains a Large Portion of the Observed Variation

The regressions for observed vs. simulated days to anthesis and grain yield (Eq. [1]) indicated that the simulation explains a large portion of the observed variation for both traits (Tables 3 and 4). The slopes had values near one, and the intercepts were small (Table 3) relative to the mean values of 49 d to anthesis and 2680 kg ha^{-1} grain yield. Proposition 1 seems reasonable, but inspection of plots of observed vs. simulated values (Fig. 2) indicated probable differences in how well responses were simulated in different locations. For example, simulated days to anthesis for Lugo was consistently later than observed values, while simulated values for Rocky Mount were often earlier than observed data. Thus, it would seem that the proposition is correct, but the caveat is required that there was evidence of lack of independence among observations.

Proposition 2: Once an Effect of the Simulation is Allowed For, the Effect of Environments is Small

One likely source of dependencies is due to differences among environments, assumed here to include combined effects of locations and growing seasons. Explicitly introducing an effect of environment in the linear regression can test whether there is an environment-related bias in the simulations. This is equivalent to saying that results from within an environment are more similar (less independent) than observations from different environments. Using Eq. [3], the bivariate model was extended by adding effects of environment and of the interaction of simulations and environments.

For days to anthesis, the mean effects of environment and the interaction were significant, representing about 9% of the model sums of squares (Table 4). The simulation $\times$ environment effect was of similar magnitude for grain yield, but due to the larger error term, the interaction was nonsignificant. Taken together, the two regressions suggest that, while a large portion of variation (92% for anthesis and 83% for grain yield) was accounted for by the simulations, there are differences among the environments that were not simulated by the model. Proposition 2 appears reasonable, but depending on needs of future research or applications, the results support examining the bias across environments. With the data at hand, however, no information is provided as to whether the bias is due to problems with input data (e.g., errors in soil profile descriptions or daily weather data), representation of processes in the simulation model, or errors in the observed data.

Proposition 3: Once Effects of the Simulation and of Environments are Accounted for, the Effect of Genotype is Small

The appropriate test for effects of genotypes differs from that of environments because genotypes are usually chosen within environments. This implies a hierarchical structure, so the appropriate regression model describes the effect of cultivars as being nested within environments (Table 4). No cultivar effect was found for days to anthesis, and the cultivar effect for grain yield represented only 5% of total model sums of squares ($P < 0.001$). Proposition 3 is therefore supported. Comparing this result to that for effects of environments (Proposition 2), the two propositions jointly suggest that more attention should be paid to improving simulation of differences among environments than of cultivar differences.

Proposition 4: Once Effects of the Simulation and of Environments are Accounted for, the Residual Effect of Temperature and Daylength are Negligible

In comparing observed vs. simulated data, the question arises as to whether the unexplained (residual) variation is associated with other explanatory variables. For crop models, one might seek relations with environmental variables like temperature or daylength, stress parameters such as indicators of water or N deficits, or other parameters produced from the simulations. Trends with residuals can evidence problems in the environmental data used as inputs, responses that are being simulated, or the observed data. The usual approach for

Table 3. Basic statistics for comparisons of observed and simulated values of days to anthesis, grain yield, and aboveground dry biomass at maturity. Data are for experiments at 11 locations. Slope, intercept, and r^2 are for linear regressions of observed vs. simulated values.

Variable name	Mean		Mean absolute difference	Root mean square error	Slope	Intercept	r^2	Willmott's d-statistic	Number of observations
	Observed	Simulated							
Days to anthesis	49	49	2	3.445	0.94	2.5	0.90	0.97	108
Grain yield (kg ha^{-1})	2626	2536	329	414.288	0.94	240.0	0.80	0.94	113
Aboveground biomass (kg ha^{-1})	5538	4858	871	1142.872	0.99	711.3	0.69	0.86	70

Table 4. Analyses of variance for linear regressions of days to anthesis, grain yield (kg ha^{-1}) and aboveground biomass at maturity (kg ha^{-1}). Independent variables include simulation outputs, environment (combinations of locations and growing seasons) and cultivars (nested within environments). Effects are tested sequentially, so the sums of squares for successive variables reflect changes in sums of squares attributable to the effect of each additional variable.

Dependent variable	Source	df	Sums of square	Mean squares	F	Probability	R^2
Proposition 1: The simulation explains a large portion of the observed variation							
Anthesis	Simulation	1	10 420	10 420.4	906.5	<0.001	0.90
	Error	106	1 219	11.5			
Yield	Simulation	1	71 664 500	71 664 500.0	437.6	<0.001	0.80
	Error	111	18 179 475	163 779.1			
Proposition 2: Once an effect of the simulation is allowed for, the effect of individual environments is small							
Anthesis	Simulation	1	10 420	10 420.4	3 218.6	<0.001	0.98
	Environment	23	786	34.2	10.6	<0.001	
	Simulation × environment	12	203	16.9	5.2	<0.001	
	Error	71	230	3.2			
Yield	Simulation	1	71 664 500	71 664 500.0	754.3	<0.001	0.93
	Environment	27	9 901 759	366 731.8	3.9	<0.001	
	Simulation × environment	21	2 292 577	109 170.3	1.1	0.326	
	Error	63	5 985 139	95 002.2			
Proposition 3: Once effects of the simulation and of environments are accounted for, the effect of genotype is small							
Anthesis	Simulation	1	10 420	10 420.4	1 905.5	<0.001	0.97
	Environment	23	786	34.2	6.2	<0.001	
	Cultivar (environment)	18	77	4.3	0.8	0.708	
	Error	65	355	5.5			
Yield	Simulation	1	71 664 500	71 664 500.0	1 206.1	<0.001	0.96
	Environment	27	9 901 759	366 731.8	6.2	<0.001	
	Cultivar (environment)	18	4 356 135	242 007.5	4.1	<0.001	
	Error	66	3 921 581	59 417.9			
Proposition 4: Once effects of the simulation and of environments are accounted for, the residual effect of temperature and daylength within environments is negligible							
Anthesis	Simulation	1	10 420	10 420.4	2 652.6	<0.001	0.97
	Environment	23	786	34.2	8.7	<0.001	
	Temperature	1	35	35.5	9.0	0.004	
	Daylength	1	54	54.5	13.9	<0.001	
	Temperature × daylength	1	29	28.6	7.3	0.008	
	Error	80	314	3.9			
Yield	Simulation	1	71 664 500	71 664 500.0	809.0	<0.001	0.92
	Environment	27	9 901 759	366 731.8	4.1	<0.001	
	Temperature	1	819 086	819 086.2	9.2	0.003	
	Daylength	1	369	368.8	0.0	0.949	
	Temperature × daylength	1	283 330	283 330.1	3.2	0.077	
	Error	81	7 174 930	88 579.4			
Proposition 5: Effectiveness of the simulation in explaining variation in grain yield is mainly because of accurate simulation of time to anthesis							
	Simulated anthesis date	1	26 040 421	26 040 421.5	157.6	<0.001	0.80
	Simulated grain yield	1	45 629 409	45 629 408.6	276.2	<0.001	
	Error	110	18 174 145	165 219.5			
Proposition 6: The version of the simulation model that includes leaf-level photosynthesis explains more variation in growth and yield than the canopy version							
Yield	Canopy submodel	1	64 996 052	64 996 051.7	400.8	<0.001	0.80
	Leaf submodel	1	7 010 817	7 010 816.6	43.2	<0.001	
	Error	110	17 837 106	162 155.5			
Biomass	Canopy submodel	1	129 057 529	129 057 529.1	147.2	<0.001	0.69
	Leaf submodel	1	1 616 080	1 616 079.6	1.8	0.179	
	Error	67	58 756 732	876 966.2			
Proposition 7: The simulation model performs similarly for high- and low-yielding situations							
Anthesis	Yield level	1	2 693	2 692.9	319.1	<0.001	0.92
	Simulation	1	7 845	7 844.9	929.5	<0.001	
	Yield level × simulation	1	223	223.4	26.5	<0.001	
	Error	104	878	8.4			
Yield	Yield level	1	58 590 335	58 590 334.8	434.5	<0.001	0.84
	Simulation	1	16 545 703	16 545 702.6	122.7	<0.001	
	Yield level × simulation	1	9 658	9 658.1	0.1	0.789	
	Error	109	14 698 279	134 846.6			

analyzing variation in residuals is to regress observed data on simulated values and to analyze patterns among the residuals as a second step (e.g., du Toit and du Toit, 2003). Examining sequential changes in model sums of squares provides a statistically equivalent procedure, but allows effects of the additional explanatory variables to be compared with the initial bivariate regression and simplifies analyses where interactions are considered.

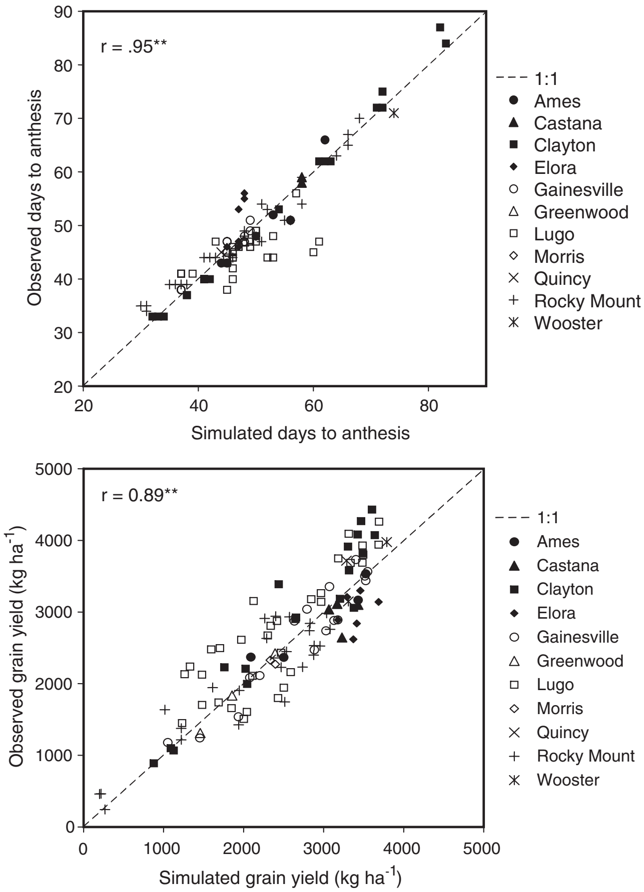

Fig. 2. Comparison of observed vs. simulated values for days to anthesis and final grain yield for the CSM-CROPGRO-Soybean model in 28 experiments from 11 locations.

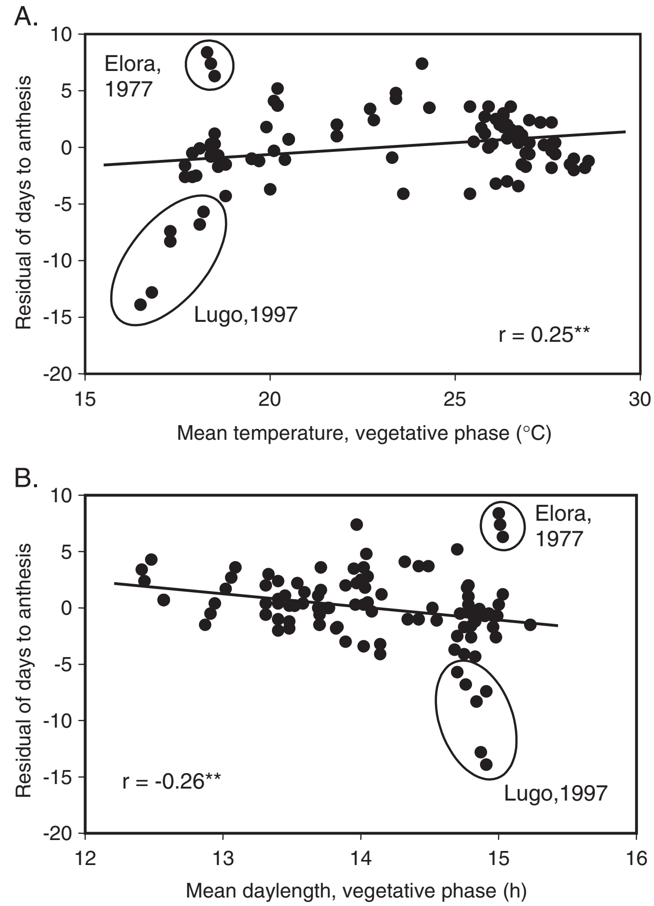

Fig. 3. Comparison of residuals from regressions of observed vs. simulated values for days to anthesis for the CSM-CROPGRO-Soybean model in 28 experiments from 11 locations. A, Relation with mean temperature during vegetative phase; B, Relation with mean daylength during vegetative phase.

Average temperature and daylengths were calculated for each treatment, considering vegetative and reproductive phases separately. Mean temperature, mean daylength, and their interaction were then tested as additional explanatory variables in the linear models. As examples, we describe results for temperature, daylength, and their interaction during the vegetative phase (Table 4, Fig. 3, and Fig. 4). All three variables affected days to anthesis (Table 4), but their combined effect accounted for less than 2% of sums of squares. Similarly, temperature and temperature × daylength effects were found for grain yield, but their combined effect was also less than 2% of sums of squares. Proposition 4 is supported. However, data for Elora (1977) and Lugo (1997) in the graphs of the residuals for days to anthesis vs. temperature and daylength (Fig. 3) suggest clustering of residuals by environment that merits further investigation. Graphs comparing residuals for grain yield with various factors did not evidence strong clustering (Fig. 4).

Proposition 5: Effectiveness of the Simulation in Explaining Variation in Grain Yield is Mainly Because of Accurate Simulation of Time to Anthesis

In many production situations, grain yield increases with length of the growing season unless stresses constrain late-season growth. Thus, a simple, albeit pessimistic, proposition is that the variation in simulated yield is mainly attributable to accurate simulation of phenology. This was tested by first regressing observed grain yield against simulated days to anthesis, and then determining whether simulated grain yield accounted for a large portion of the remaining variation in observed yield (Table 4). Anthesis date only accounted for 36% of the model sum of squares, so a large portion of the variation in yield that was accounted for by the simulation model was independent of simulated anthesis date. Proposition 5 is therefore rejected.

Proposition 6: The Version of the Simulation Model that Includes Leaf-Level Photosynthesis Explains More Variation in Growth and Yield than the Canopy Version

The more mechanistic leaf photosynthesis submodel, available as an option in CSM-CROPGRO, was expected to improve simulations of grain yield and biomass at maturity, especially if different row-spacings were used or water deficits were severe enough to limit canopy development. For grain yield, the comparison of the simulations from the two submodels revealed that the leaf

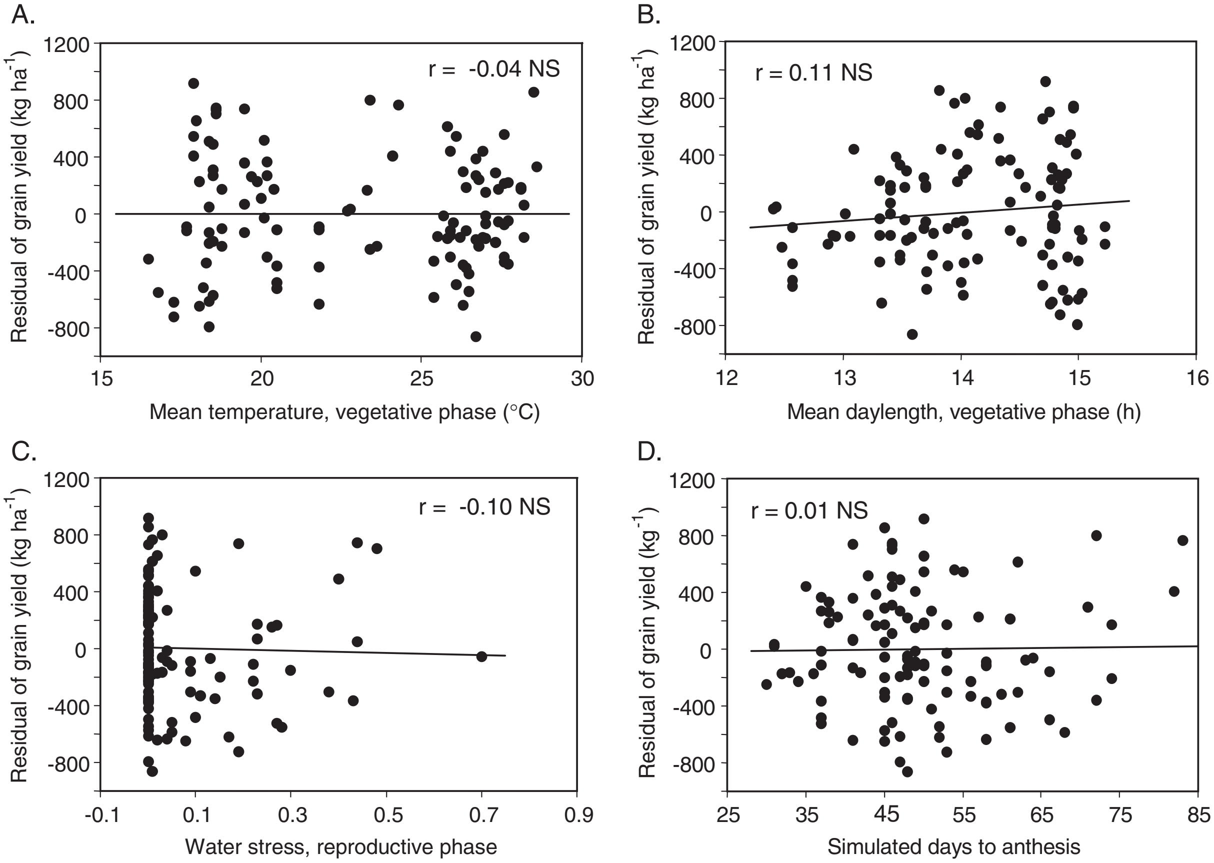

Fig. 4. Comparison of residuals from regressions of observed vs. simulated values for days to anthesis for the CSM-CROPGRO-Soybean model in 28 experiments from 11 locations. A, relation with mean temperature during vegetative phase; B, relation with mean daylength during vegetative phase; C, relation with mean water stress indicator during reproductive phase; D, Relation with simulated days to anthesis.

submodel accounted for an additional 8% of sum of squares (Table 4; $P < 0.001$). For biomass at maturity, however, the leaf submodel explained less than 1% of the variation (Table 4; not significant, NS), indicating no improvement in biomass prediction. Because biomass was often not measured or not reported, only 78 values of biomass at maturity were available vs. 113 for grain yield. Thus, a partial explanation for the contrasting results for grain yield and biomass may be that the biomass data represented too small a subset of the experiments with yield data. An additional explanation is that biomass at maturity may have a higher measurement error than yield due to smaller sample (plot) sizes, uncertainty over inclusion of senesced tissue, differences in drying regimes, or other familiar problems of field research. The role of measurement error could be assessed if observed data were provided with respective standard errors or as data from individual replicates. Proposition 6 is supported for grain yield, but not for final biomass. Since there is a benefit in simulated grain yield and no apparent drawback for biomass, use of the leaf-level submodel appears justified.

Proposition 7: The Simulation Model Performs Similarly for High- and Low-Yielding Situations

Simulation models often are developed and calibrated using data from environmental conditions or management practices that vary more than those expected for commercial production. This raises the concern that while performance of a model may appear satisfactory when tested under this wide range, model performance may differ for high- or low-yielding situations. By introducing a variable characterizing observations as coming from high-yield (here defined as observed grain yield $\geq$2500 kg ha^{-1}) or low-yield (observed grain yield $<$2500 kg ha^{-1}) situations, regressions can test for an interaction of yield level with the simulations.

For days to anthesis, the direct effect of yield level (Table 4) simply confirmed that a long vegetative phase is associated with higher grain yields, as noted in Proposition 5. For grain yield, the large effect of yield level simply is a consequence of the data being partitioned into the two yield levels. The key issues in assessing Proposition 7 are whether the simulation still explains a large portion of variation, indicating that the model predicts variation within each group, and whether the interaction of yield level with simulation is large enough to suggest a model bias related to yield level.

Simulation per se still explained an important portion of variation for days to anthesis and grain yield (Table 4). Days to anthesis did show an interaction with yield level ($P < 0.001$), but the portion of variation attributable to the interaction was less than 3% of that attributable to the simulation (Table 4). Overall, the results support Proposition 7, suggesting that CSM-CROPGRO will perform similarly for both high- and low-yield situations. This analysis could be refined by examining yield variation in relation to specific factors such as water or fertilizer regimes, weather conditions, or unidentified stresses. If quantitative factors were used (e.g., stored soil moisture at time of planting), the linear regressions would be equivalent to the analysis of residual variation demonstrated in Proposition 4.

Further Considerations on Use of Regression Models

Extension of bivariate regressions to include factors and additional variates can provide valuable insights on model performance. However, the approach should not be applied mechanically without understanding potential limitations. Foremost is that a strong effect in a regression model does not automatically indicate a causal relation. Care is also needed in assuring that the correct variables are considered, that possible nesting is accounted for, and that the variables are introduced in the most meaningful order.

As illustrated in Proposition 6, multiple regressions also can bring more rigor to comparisons of models, including model options or versions. The approach offers a partial solution to the problem of determining whether a new model truly improves prediction. This issue is often raised in discussions of parsimony and the merits of simple vs. complex models (Monteith, 1996; Gauch, 2003). We emphasize, however, that regression tests can be abused by using them in an exploratory fashion to find the combinations of model parameters or equations that result in an apparently significant improvement. The preferred approach is to use other criteria for goodness of fit such as root mean square error to identify the best new model and only use the regression-based test with an independent set of validation data.

The strength of any conclusions supported by regressions will increase with the quality, quantity, and representativeness of the observed data. Data quality relates to basic issues such as experimental design, plot size, sampling strategy, and protocols for processing samples, but for simulation models, it also involves the thoroughness with which growing conditions and crop management are described and can be accurately quantified.

Determining whether a model explains essentially all of the variation that is not readily attributable to error requires having an independent estimate of error. Such an estimate is provided by the residual error from the individual trials once all replicate and treatment effects have been removed. If the regression residual variance is not significantly different from the mean experimental variance (testable with a simple *F* test), then one can conclude that the simulation model explained the variance to a reasonable level with no effects left to explain. The null hypotheses, H_0, thus becomes "there are no factors or variates that would explain the regression residual". Rejecting H_0 implies that there still is residual variation that should be explainable.

Regression analyses are sensitive to the number of data points considered, and due to the lack of orthogonality and balance among treatments, an especially large number of observations are desirable. We suggest that the 108 observations available for days to anthesis in this soybean data set approached the viable lower-limit for multiple regressions of this type, but a more rigorous analysis of sample size is justified.

Various procedures can improve the robustness of linear regression analysis. The R Project package has tools to facilitate graphical inspection of residuals and detection of values that have a large influence (leverage) on the regression (Maindonald and Braun, 2003). If experimental error is thought to vary substantially for different experiments or main treatments within an experiment, weighted-least squares analysis may be used (Kleinbaum and Kupper, 1978). Transformations can be used to stabilize the variance of the observed variable.

The question of whether the available data represent the application domain of interest is often neglected in evaluating models. Ideally, the selection of locations and treatments should match the application domain. The tests of Proposition 7 on model performance by yield level provided a simple example of how simulations can be analyzed according to specific characteristics. The same approach could be applied for specific management conditions (e.g., atmospheric CO_2 concentrations), cultivar types (e.g., determinate vs. indeterminate), or even genetic loci (e.g., the *E* loci that affect flowering time in soybean). Another interesting opportunity is to strengthen evaluations by conducting special experiments to link data sets through common set of cultivars or other treatments. This might involve assembling data from previous multilocation trials or conducting specially designed model evaluation trials.

CONCLUSIONS

The use of multiple regressions allowed for the testing of diverse hypotheses that relate to model performance. Explicitly testing for effects of locations, cultivars, management, or specific environmental factors can help researchers understand the relative strengths and weaknesses of a model or different models. Similarly, the approach is useful for investigating whether changes in a model improve the predictive power of the model relative to previous versions.

However, performing these analyses requires having access to large numbers of treatments $\times$ environment combinations. Soybean data on experimental conditions and observed crop responses were available in the DSSAT Version 4.0 format (Hoogenboom et al., 2004), which partially implements standards developed by the ICASA (Hunt et al., 2001). Use of such standards should help promote wider and more effective sharing of data. A sample program using the standard DSSAT output files as data sources for the statistical analysis program SAS Version 8 (SAS Institute, Cary, NC, USA) is available from the first author.

ACKNOWLEDGMENTS

The authors thank the various research groups that made their soybean data sets available through the DSSAT4 software system.

REFERENCES

Acosta-Gallegos, J., and J.W. White. 1995. Phenological plasticity as an adaptation by common bean to rainfed environments. Crop Sci. 35:199–204.

Alexandrov, V.A., and G. Hoogenboom. 2000. Vulnerability and adaptation assessments of agricultural crops under climate change in the Southeastern USA. Theor. Appl. Climatol. 67:45–63.

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

Bannayan, M., N.M.J. Crout, and G. Hoogenboom. 2003. Applying the CERES-Wheat model for real-time forecasting of winter wheat. Agron. J. 95:114–125.

Boote, K.J., J.W. Jones, and G. Hoogenboom. 1998a. Simulation of crop growth: CROPGRO Model. p. 651–692. *In* R.M. Peart and R.B. Curry (ed.) Agricultural systems modeling and simulation. Marcel Dekker, New York.

Boote, K.J., J.W. Jones, G. Hoogenboom, and N.B. Pickering. 1998b. The CROPGRO model for grain legumes. p. 99–128. *In* G.Y. Tsuji et al. (ed.) Understanding options for agricultural production. Kluwer Academic Publ., Dordrecht, the Netherlands.

Boote, K.J., and N.B. Pickering. 1994. Modeling photosynthesis of row crop canopies. HortScience 29:1423–1434.

Bostick, W.M., J. Koo, V.K. Walen, J.W. Jones, and G. Hoogenboom. 2004. A web-based data exchange system for crop model applications. Agron. J. 96:853–856.

Chapman, S.C., G.L. Hammer, D.G. Butler, and M. Cooper. 2000. Genotype by environment interactions affecting grain sorghum: III. Temporal sequences and spatial patterns in the target population of environments. Aust. J. Agric. Res. 51:223–233.

du Toit, A.S., and D.L. du Toit. 2003. Short description of the model statistical package and weather analogue program. p. 42–46. *In* J.W. White (ed.) Modeling temperature response in wheat and maize. CIMMYT, El Batán, Mexico.

Egli, D.B., and W. Bruening. 1992. Planting date and soybean yield: Evaluation of environmental effects with a crop simulation model: SOYGRO. Agric. For. Meteorol. 62:19–29.

Gauch, H.G., Jr. 2003. Scientific method in practice. Cambridge Univ. Press, Cambridge, UK.

Godwin, D.C., and U. Singh. 1998. Nitrogen balance and crop response to nitrogen in upland and lowland cropping systems. p. 55–77. *In* G.Y. Tsuji et al. (ed.) Understanding options for agricultural production. System approaches for sustainable agricultural development. Kluwer Academic Publ., Dordrecht, the Netherlands.

Gomez, K.A., and A.A. Gomez. 1984. Statistical procedures for agricultural research. 2nd ed. John Wiley & Sons, New York.

Hartkamp, A.D., J.W. White, W.A.H. Rossing, M.K. van Ittersum, E.J. Bakker, and R. Rabbinge. 2004. Regional application of a cropping systems simulation model: Crop residue retention in maize production systems of Jalisco, Mexico. Agric. Syst. 82: 117–138.

Hood, C.P., R.W. McClendon, and J.E. Hook. 1987. Computer analysis of soybean irrigation management strategies. Trans. ASAE 30:417–423.

Hoogenboom, G., J.W. Jones, and K.J. Boote. 1992. Modeling growth, development, and yield of grain legumes using SOYGRO, PNUTGRO and BEANGRO. A review. Trans. ASAE 35:2043–2056.

Hoogenboom, G., J.W. Jones, P.W. Wilkens, C.H. Porter, W.D. Batchelor, L.A. Hunt, K.J. Boote, U. Singh, O. Uryasev, W.T. Bowen, A. Gijsman, A. du Toit, J.W. White, and G.Y. Tsuji. 2004. Decision support system for agrotechnology transfer [CD-ROM]. Version 4.0. Univ. of Hawaii, Honolulu.

Hoogenboom, G., J.W. White, J. Acosta-Gallegos, R. Gaudiel, J.R. Myers, and M.J. Silbernagel. 1997. Evaluation of a crop simulation model that incorporates gene action. Agron. J. 89:613–620.

Hunt, L.A., J.W. White, and G. Hoogenboom. 2001. Agronomic data: Advances in documentation and protocols for exchange and use. Agric. Syst. 70:477–492.

Jones, J.W., K.J. Boote, S.S. Jagtap, G. Hoogenboom, and G.G. Wilkerson. 1989. SOYGRO V5.42 soybean crop growth simulation model: Users guide. Florida Agric. Exp. Stn. J. no. 8304. Agricultural Engineering Dep. and Agronomy Dep., Univ. of Florida, Gainesville.

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

Jones, P.G., and P.K. Thornton. 2003. The potential impacts of climate change on maize production in Africa and Latin America in 2055. Glob. Environ. Change 13:51–59.

Kleinbaum, D.G., and L. Kupper. 1978. Applied regression analysis and other multivariable methods. Wadsworth Publ. Co., Belmont, CA.

Maindonald, J., and J. Braun. 2003. Data analysis and graphics using R. Cambridge Univ. Press, New York.

Mayer, D.G., M.A. Stuart, and A.J. Swain. 1994. Regression of real-world data on model output: An appropriate overall test of validity. Agric. Syst. 45:93–104.

Mavromatis, T., K.J. Boote, J.W. Jones, A. Irmak, D. Shinde, and G. Hoogenboom. 2001. Developing genetic coefficients for crop simulation models with data from crop performance trials. Crop Sci. 4:41–51.

Mavromatis, T., K.J. Boote, J.W. Jones, G.G. Wilkerson, and G. Hoogenboom. 2002. Repeatability of model genetic coefficients derived from soybean performance trails across different states. Crop Sci. 42:76–89.

Mearns, L.O., T. Mavromatis, and E. Tsvetsinskaya. 1999. Comparative responses of EPIC and CERES crop models to high and low spatial resolution climate change scenarios. J. Geophys. Res. 104(D6):6623–6646.

Monteith, J.L. 1996. The quest for balance in crop modeling. Agron. J. 88:695–697.

Ritchie, J.T. 1998. Soil water balance and plant water stress. p. 45–58. *In* G.Y. Tsuji et al. (ed.) Understanding options for agricultural production. Kluwer Academic Publ., Dordrecht, the Netherlands.

Sau, F., K.J. Boote, and B. Ruiz-Nogueira. 1999. Evaluation and improvement of CROPGRO-soybean model for a cool environment in Galicia, northwest Spain. Field Crops Res. 61:273–291.

White, J.W. 1998. Modeling and crop improvement. p. 179–188. *In* G.Y. Tsuji et al. (ed.) Understanding options for agricultural production. Kluwer Academic Publ., Dordrecht, the Netherlands.

Willmott, C.J. 1982. Some comments on the evaluation of model performance. Bull. Am. Meteorol. Soc. 63:1309–1313.