EVALUATION OF CROP YIELD MODELS

Wendell W. Wilson
Supervisory Mathematical Statistician
Yield Evaluation Section, Yield Research Branch
Statistical Research Division, Statistical Reporting Service
U. S. Department of Agriculture

INTRODUCTION

The evaluation of crop yield models has been a very important element of the AgRISTARS Program[1/]. Its importance to the program arose because of the need to understand the characteristics and capabilities of existing models before embarking on extensive research programs to develop new or improved crop yield models. Many efforts have been made in model evaluation and associated areas in the AgRISTARS Program. The subset of those efforts discussed below involves evaluating models for their usefulness in providing information to support improved crop yield forecasts and estimates for large areas. That is, improving the forecasts made during the crop season and the estimates made at or soon after the season's end of the average yield over specified areas. Yield estimates are needed for these areas (often referred to as large areas to differentiate them from areas as small as plots, fields or counties) so they can be multiplied times crop area estimates to obtain estimates of the crop's total production.

There are many points of view from which to examine crop yield model evaluation. Basically, crop yield models are evaluated to first discriminate between models with some capabilities and those with little or no value and then, hopefully, to identify those of substantial value from those with some capabilities. Figuratively, model evaluation is designed to separate the wheat

1/ The AgRISTARS Program is a multi-agency research program to meet some current and new information needs of U. S. Department of Agriculture. AgRISTARS is an acronym for Agriculture and Resources Inventory Surveys Through Aerospace Remote Sensing.

from the chaff and then if one is fortunate to separate the high protein gluten from the seed coat. To provide some organization to the topic, first the eight evaluation criteria, or properties that preferred models would possess, will be reviewed. Then, in somewhat more detail, techniques for evaluating models for one of these criteria, yield indication reliability, will be discussed. After acquainting the reader with these techniques, methods of comparing the indication reliability of competing models will be discussed.

Then, in the second part of the paper, official yield forecasts will be discussed. These are the published forecasts that both those who prepare and release them, and those who use them and depend on their accuracy wish were better. The need to assess the impact of the use of selected models on the official forecasts will form the basis for describing a method of objectively simulating official forecasts from the set of quantifiable survey and model indications on which they are based. Then, an extension of the simulation method will be described, which assesses how the official forecasts may have changed had a selected model been used as either a supplement to the other indications or in lieu of one or more of them. Finally, the official forecasts and forecasts as they would be altered by the use of a selected model will be evaluated and compared much as would be done for individual models or competing models.

EVALUATION CRITERIA: PROPERTIES THAT PREFERRED MODELS WOULD POSSESS

Eight crop yield model evaluation criteria were established early in the AgRISTARS Program (Wilson, et al., 1980). They are: yield indication reliability, objectivity, consistency with scientific knowledge, adequacy, timeliness, minimum cost, simplicity and provision of accurate current indications of modeled yield reliability. The criteria can be thought of, in positive terms, as the properties that preferred models would possess. These positive

properties that one looks for in crop yield models, or for that matter, in any method which provides a yield indication, will now be briefly described.

Yield indication reliability is a measure of the degree to which users can rely on crop yield indications from a model or other method as a source in setting official yield forecasts and estimates, and in using them as a basis for policy determinations. Users will often have multiple sources of information. They need to know how much confidence they should have in each source.

Objectivity would be fully achieved for a crop yield model that requires no subjective judgments which involve adjusting the model form, parameter estimates or input variables. Subjectivity may have been involved in model development, but for the fixed model all parameters and input variables are "measurable," methods of estimation or derivation are fully documented, and the model is exactly repeatable under the same conditions. Even though model users may wish to apply subjective judgment in using various sources of information in arriving at a yield figure, it is still desirable that to the extent possible individual yield information sources be objective. Greater objectivity will allow users to more fully understand each model's or method's characteristics, limitations and capabilities.

The consistency with scientific knowledge that a model possesses can be examined at various levels of detail. The agreement or consistency of a crop yield model's form and parameter values with experimental data and scientific knowledge is an important criterion in model evaluation. The sensitivity of the modeled yield to important environmental factors is an important measure of model capability and acceptance. The omission of important factors or the predominance of a few inputs such that other important variables have a minimal impact can also be examined. Understanding when or under what conditions a

model might be inconsistent with known physical and biological responses is important.

Adequacy of crop yield models can be assessed in terms of the extent of geographic coverage of a crop, the level of detail provided and in the appropriateness of the model for intended future applications. A model with greater potential for adaptability in providing coverage of important producing areas may be considered more adequate. Limitations in coverage will often be related to the lack of or inaccurate measures of model required input variables, in some producing areas. Greater adequacy in terms of the level of detail could provide reliable yield information for smaller geographic subdivisions, or separate estimates for different crop production systems or different crop utilization categories. Appropriateness for future applications might involve the provision of yield indications for the same utilization categories, production systems, geographic areas or other strata that are used in estimating crop area.

Minimum cost is obviously a very desirable characteristic for successful crop yield models or other methods of obtaining yield indications. Cost of the operating system associated with a model or method is the primary consideration. Cost of operating models will be appraised for various types of activities. Some of these are: acquiring, formatting and using historical data bases to update model parameters; acquiring, editing and summarizing current year values in a timely manner for model execution, and activities associated with the need for frequent model updates, number and kind of variables, and the complexity of the model.

Simplicity is a desirable model characteristic. If two models were equal for the other seven criteria, then one would, of course, select the simpler model. Simplicity in crop yield model form and use is often associated with lower operating costs. However, a more important aspect of model simplicity can

be an enhanced ability of the user to understand the concepts, capabilities and limitations of the model. A thorough understanding allows the user to evaluate the model's indication in the light of other information and make valid judgments. A simpler model would generally have lower user training and experience requirements.

The availability, at the time of model use, of a model generated indication of the reliability of the model's yield point estimate is desirable if it provides any information on the actual reliability of that point estimate. This provision of accurate current indications of modeled yield reliability will be appraised for its availability and utility for each candidate yield model. The degree to which such an indication (when available) corresponds to subsequently determined actual performance will be assessed. The basic task is to ascertain the degree to which the user can depend upon a model's indication of reliability for guidance on the degree of confidence to be placed in that model's current yield indication or in the appropriate confidence interval around its point estimate.

TECHNIQUES FOR EVALUATING THE YIELD INDICATION RELIABILITY OF INDIVIDUAL MODELS

Examination of yield indication reliability over a period of years usually involves independent operation of models and the measurement of such things as: the mean square error, variance, bias, proportion of years beyond a critical error limit, worst and second to worst performance during the period, direction of change from mean and previous year yields, and the simple correlation coefficient between actual and model predicted yields. To obtain realistic tests of how a model's yield indication would perform in the future, it is necessary to simulate the performance of the fixed form of the model over a period of past

years for which the actual yield is known. In order to accomplish this simulation, a "bootstrap" technique is used (Wilson, et al., 1980). Years from an earlier base period are used to estimate the model parameters. A predicted yield, $\hat{Y}$, is generated for the following year based on input variable values for that year. Then, the first test year is added to the base period and the process is repeated for the second test year in the sequence. Continuing in this manner predicted yields are generated for each of the test years which are independent of data from the test year or any year following it. It should be noted that even though the data used to estimate the model parameters do not include the test year, this technique does not necessarily result in a predicted yield which is totally independent of the data from the test years if data from any of the test years were used to develop the model's form. Therefore, the bootstrap technique will not provide information about how a model will perform in the future if the model form is altered in any way. However, the bootstrap test procedure does provide a valid independent test of a model in its current form. Measures of yield reliability were developed for use in AgRISTARS by first applying them to a simple trend model (Sebaugh, 1981a). Subsequently, they have been applied in evaluating a wide variety of models for various crops and regions.

The values required to compute the measures of yield indication reliability are the predicted yield, $\hat{Y}$, the actual (reported) yield, Y, and the difference between them, $d = \hat{Y} - Y$, for each test year. From the d value, the mean square error (root and relative root mean square error), the variance (standard deviation and relative standard deviation), and the bias (its square and the relative bias) are calculated. Statistical formulas for these and other calculations are shown in the appendix and discussed in Wilson and Sebaugh, 1981. The root mean square error (RMSE) and the standard deviation (SD) indicate the accuracy and

the precision over the test period in the original units of yield measurement (bushels/acre, quintals/hectare, etc.). Accuracy is indicated by a small RMSE. A non-zero bias means the model is, on the average, overestimating the yield (positive bias) or underestimating the yield (negative bias). The SD is smaller than the RMSE when there is a non-zero bias and indicates what the RMSE would be if there were no bias. If the bias is near zero, the SD and RMSE will be close in value.

The relative difference expressed as a percentage, RD = 100 d/Y, is useful in computing the number of test years beyond some critical percentage error limit, such as 10 percent, when the model indication would be of little value. The worst and next to the worst performance are defined as the largest and next to largest absolute value of the relative difference. Another set of measures demonstrates the correspondence between the actual and predicted yields. It would be desirable for increases (decreases) in actual yield to be accompanied by increases (decreases) in predicted yields. Two measures are computed for the correspondence or lack of correspondence in the direction of yield changes. One looks at the proportion of agreement in the direction of yield changes from the previous year and the other the level of agreement in changes from the average yield over the three previous years. A base period of three is used since a longer period would reduce the number of observations over the limited number of test years and a shorter period would not be very different from comparison to a single previous year. Finally, the Pearson correlation coefficient, r, between the set of actual and predicted test period yields is computed. It is desirable for $r(-1 \leq r \leq 1)$ to be large and positive.

To illustrate the application of techniques for evaluating the yield indication reliability of an individual model, a table and figure are shown below. Table 1 presents yield indication reliability measures for a research

model (X_1) considered for estimating final yield of a spring planted small grain in a major producing state. Figure 1 shows the final reported yields and the independently predicted yields of model X_1.

Table 1. Yield Indication Reliability Measures for Research Model X_1 for a Spring Planted Small Grain in a Major Producing State, Recent Thirteen Year Independent Test Period

Measure of Yield Reliability	Unit	Model X_1 Value
Bias = B	Bu/A	-2.31
Relative Bias = RB	%	-5.7
Mean Square Error = MSE	$(Bu/A)^2$	32.42
Root Mean Square Error = RMSE	Bu/A	5.69
Relative Root Mean Square Error = RRMSE	%	14.1
Variance = VAR	$(Bu/A)^2$	27.10
Standard Deviation = SD	Bu/A	5.21
Relative Standard Deviation = RSD	%	13.7
Percent of Years $\|RD\| > 10\%$	%	54
Largest $\|RD\|$	%	30.0
Next Largest $\|RD\|$	%	-22.4
Percent of years direction of change in predicted yield agrees with actual yield changes from		
(1) the previous year	%	50
(2) the average of the previous three years	%	80
Pearson correlation coefficient between actual and predicted yields	-	0.64

METHODS OF COMPARING THE YIELD INDICATION RELIABILITY OF COMPETING MODELS

The performance of two competing models may be compared using the measures of yield indication reliability discussed above. However, it is also desirable to perform statistical tests comparing the reliability of competing models. Such statistical tests were developed and presented in an early AgRISTARS report by Sebaugh (1981b). A formal statistical test considers the variability of each

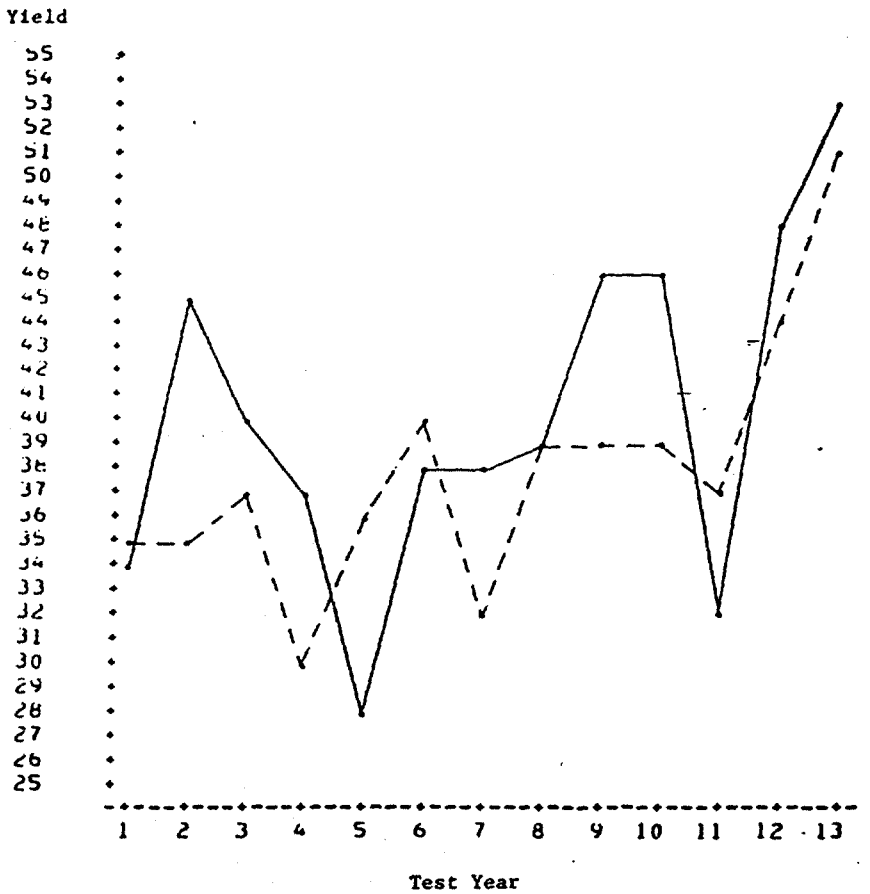

Figure 1. Time series plot of Model X_1 predictions (——) and final reported yields (——) for a spring planted small grain in a major producing state, for a recent thirteen year test period. Yield in bushels per acre.

model's performance over time and allows specification of an upper limit on the probability of incorrectly declaring one model more reliable than another. Because of the manner in which models are chosen for testing, it is challenging to construct a meaningful statistical test. Only yield models which have been presented in the literature or developed by known experts are considered. Therefore, great differences in the reliability of the models are not expected. A powerful statistical procedure is needed which is able to detect small, although important, differences in reliability. Also, the test should be able to function well with relatively small samples of data for each model, say ten years. The test should perform well when only two models are being compared.

Often only two models of a particular type are competitive and available for testing. When models of different types are to be compared, it is likely that the best models of each type will be compared.

It would appear that an F test could be useful in comparing the mean square errors of two models. However, if the mean square errors are based on ten years of data and the upper probability limit or significance level is set at $\alpha = .05$, then one model's mean square error must be four times larger than the others before the models can be declared different. This is an unreasonable requirement since models in the evaluation process will almost always be more competitive than this.

A more sensitive test can be constructed by considering that one model is considered more reliable than another if its predicted yields, $\hat{Y}$'s, are closer to the actual yields, Y's. No difference in the yield indication reliability of two models for a particular year implies the absolute value of the difference between their predicted yields and the actual yield is the same. The reliability of a model for that year is related to the amount of the discrepancy, not its direction. By defining $d_1 = \hat{Y}_1 - Y$, $d_2 = \hat{Y}_2 - Y$ and $D = |d_1| - |d_2|$, the models are equally reliable in a year for which D equals zero. If D is not equal to zero, one model is more reliable than the other for that year. In formal terms, this sets up the null hypothesis that there is no difference in yield indication reliability between the two models over the test years. To test that hypothesis, the D values from the test years can be used to compute a test statistic and a decision made whether or not to reject the hypothesis. Since the results for the models are paired for the same set of test years, paired-sample statistical tests are used.

Two types of paired-sample statistical tests are used: a parametric test using the student "t" test statistic and a nonparametric test using the Wilcoxon

signed rank test statistic. One reason for applying both tests is that they require different assumptions. The parametric t-test assumes the D values are normally distributed while the nonparametric test does not. If the d values are normally distributed, then the $|d|$ values would be folded normals rather than normally distributed. Although both models are folded at $|d| = 0$, their means may differ and the distribution of D has a possibility of not being normally distributed. The t-test is robust with respect to the normality assumption; however, this possible assumption violation is one reason for also conducting the nonparametric test.

The other reason for running both tests concerns the different conditions under which the null hypothesis is rejected by each test. For the parametric test, the basis for rejecting the no model differences hypothesis is a sufficiently large average D value as compared to the variability of the Ds. The hypothesis will be rejected and the model with the smaller $|d|$ values declared more reliable if t is large (either positive or negative). It would be possible for one model to have a smaller $|d|$ value for each of the test years, that is, consistently outperform the other model. However, such consistently good performance would result in rejection of the parametric null hypothesis only if the average D value were large enough relative to the variance of the Ds.

Using the nonparametric test, the null hypothesis will always be rejected if one model has smaller $|d|$ values for each of the test years. Therefore, even if the models are very competitive in terms of the $|d|$ values each year, but one model consistently outperforms the other model, the nonparametric test will still declare the consistent model to be more reliable. The hypothesis of equal model performance will only be rejected by the nonparametric test if one model has enough years with smaller $|d|$ values than the other model. The model with more smaller $|d|$ values is considered to be more reliable in its consistency of

performance. However, to reject the null hypothesis and declare one model clearly more reliable, consistency of performance is not a sufficient requirement. Consider the situation in which one model is more consistent in outperforming the other model but the largest D values occur when the less consistent model performs better. In the few years the less consistent model performs better, it performs much better. When such a contradiction exists, the null hypothesis will not be rejected and the consistent model will not be declared more reliable. The nonparametric null hypothesis will be rejected only if one model is more consistent in outperforming the other model and the largest differences between the models occur when the consistent model performs better.

To illustrate methods of comparing the yield indication reliability of competing models, another research model, model X_2, is compared to model X_1. Table 2 presents the yield indication reliability measures for this new model. The results for each measure on model X_2 can be compared to results for model X_1, in Table 1. Figure 2 shows (again) the final reported yields of a spring planted small grain in a major producing state along with the yield predictions for model X_1 and model X_2. The parametric test shows a highly significant difference ($P=.0011$) in favor of model X_2. Model X_2 is also found to be the model with greater reliability by the nonparametric test ($P<.0025$).

THE NEED TO ASSESS THE IMPACT OF SELECTED MODEL USE ON OFFICIAL YIELD FORECASTS

Evaluation of an individual crop yield model will be useful in identifying if there is any potential for using the model and may provide feedback to the model developer that is helpful in improving the model. Likewise, the comparison of models will help identify the best available models for further development or even actual use. However, eventually the most promising models must be appraised for the impact they might have if actually used in setting

official yield forecasts. Because several yield indications are often used in preparing the official forecasts, it is uncertain how useful a new model would be when used along with these indications. It is possible that use of a somewhat less reliable model, that is more independent or unique from the other indications, would be of more benefit than a more reliable model, which tends to duplicate some of the other indications. However, this possibility does not mean that just any model needs to be assessed for its impact on official forecasts. Promising models, no matter how capable they appear based on evaluation and comparison results, which are being seriously considered for use should be assessed for their likely impact on the official forecasts.

Table 2. Yield Indication Reliability Measures for Research Model X_2 for a Spring Planted Small Grain in a Major Producing State, Recent Thirteen Year Independent Test Period

Measures of Yield Reliability	Unit	Model X_2 Value
Bias = B	Bu/A	-0.01
Relative Bias = RB	%	-0.0
Mean Square Error = MSE	$(Bu/A)^2$	1.07
Root Mean Square Error = RMSE	Bu/A	1.03
Relative Root Mean Square Error = RRMSE	%	2.6
Variance = VAR	$(Bu/A)^2$	1.07
Standard Deviation = SD	Bu/A	1.03
Relative Standard Deviation = RSD	%	2.6
Percent of Years $\|RD\| > 10\%$	%	.0
Largest $\|RD\|$	%	4.6
Next Largest $\|RD\|$	%	4.2
Percent of years direction of change in predicted yield agrees with actual yield changes from		
(1) the previous year	%	92
(2) the average of the previous three years	%	100
Pearson correlation coefficient between actual and predicted yields	-	0.99

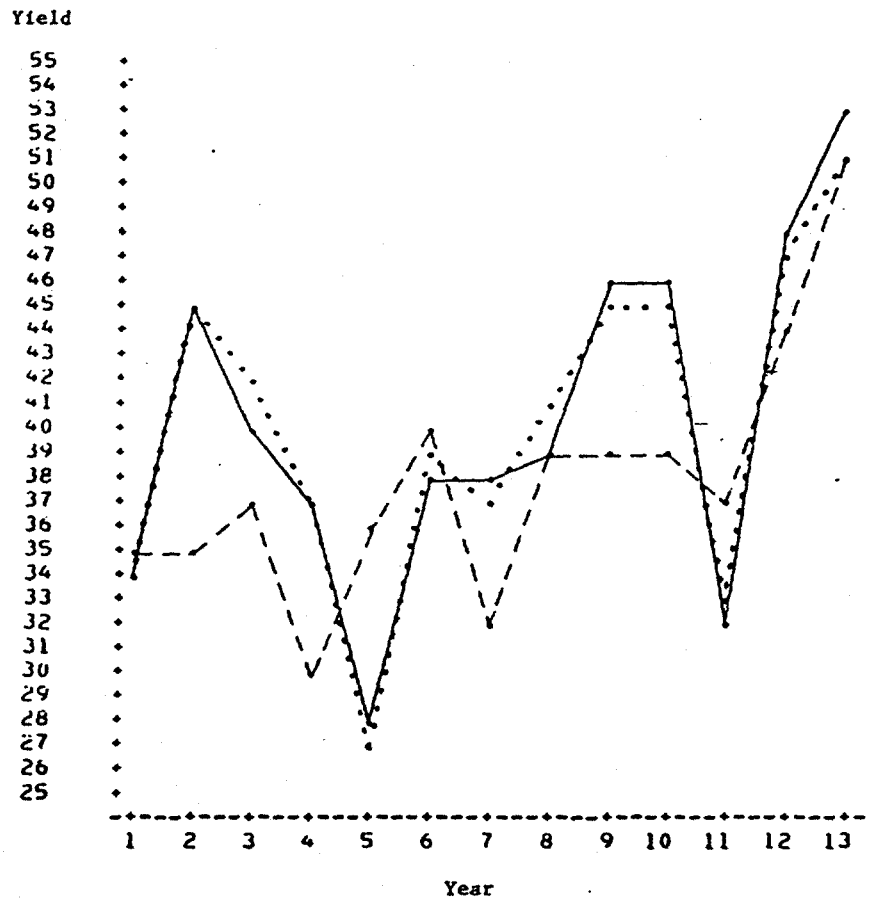

Figure 2. Time series plot of Model X_1 (----) and Model X_2 (••••) predictions and final reported yields (——) for a spring planted small grain in a major producing state, for a recent thirteen year test period. Yield is in bushels per acre.

Because newly developed models or methods of obtaining yield indications are unlikely to change the way in which quantifiable indications are considered in setting official yield forecasts, it should be possible to learn something by studying the current procedures. By understanding how the current set of indications is used in determining the official published forecasts, it should be possible to retrospectively determine how the use of new yield indication tools would have changed the official forecasts and whether their use would result in an improvement in accuracy or not.

A METHOD OF OBJECTIVELY SIMULATING OFFICIAL YIELD FORECASTS FROM THE QUANTIFIABLE SURVEY AND MODEL INDICATIONS ON WHICH THEY ARE BASED

It is possible, with some study, to acquire an intuitive understanding of how individual yield indications are examined, how some are adjusted for a pattern of consistent bias and how they are combined in arriving at the official published yield forecasts. However, to enable one to answer a series of "what if" questions it is desirable to first have a method which objectively mimics the entire current process. Such a method is described below.

Crop yield indications are often examined by plotting them on a time series chart along with the official final yield estimates. The plots usually cover a period of recent years for which the methods of obtaining the indications and general crop husbandry practices are assumed to have changed very little. Some of the indications may exhibit a consistent bias from the final yield, and for these it may be possible to make a current year adjustment based on the pattern in the previous years. For other indications, the bias is not very substantial or cannot be detected because the variability of the indication with respect to the final yield is large. Such a time series chart (see Figure 3) is shown for a wheat crop in a major producing U.S. state for the first official forecast in the 1984 crop year. Four currently used quantifiable survey average or model generated yield indications are plotted. They are: indication A - farmer reported crop condition, adjusted for trend; indication B - farmer reported mean locality yields; indication C - multivariate regression model which uses reported condition, monthly precipitation and trend; and indication D - forecasting method which utilizes sample plant measurement and count data relationships. Additional information on the types of yield indications used may be found in Scope and Methods of the Statistical Reporting Service, Miscellaneous Publication No. 1308 (1983).

In order to simulate the chart reading process whereby a value is determined for the forecast of the final yield for the current year, it is first necessary to examine how attempts to "read-out" any bias in the indications are employed. The algorithm presented here (developed by the author and Jeanne L. Sebaugh) is based on the information that would be looked at while reading the time series chart for an individual indication, for example indication D in Figure 3. The practitioner usually views the differences between the indication and the final yield to see if there is a consistent pattern which can be applied to the current year. If there are one or two unusual years which do not fit the overall pattern, they may be given less weight. Considering the overall pattern, perhaps with less consideration given to unusual years, the practitioner derives a "typical" difference value and uses it to obtain a chart-read value for the current year's indication.

The simplest objective chart reading algorithm would be to simply compute the past five years arithmetic mean of differences between the indication and final yield, for example indication D minus final yield (I_D-Y), and subtract that bias estimate from current year's indication. However, such a method would give equal weight to each year's difference. Therefore, the method selected gives additional weight to differences near the median difference and less to those further from it. The weights applied are the inverse of the distance of each difference from the median difference. To prevent the weight for any year from being excessively large, the minimum distance is restricted to 0.5 bushels. This is approximately the degree of rounding employed in the official forecasts and allows no weight to be greater than two. Table 3 shows the calculations for indication D for use in 1984. Of course, this algorithm may be applied to each indication, but would not be used unless such a chart reading improved the indications correspondence to the final yield, that is resulted in a smaller

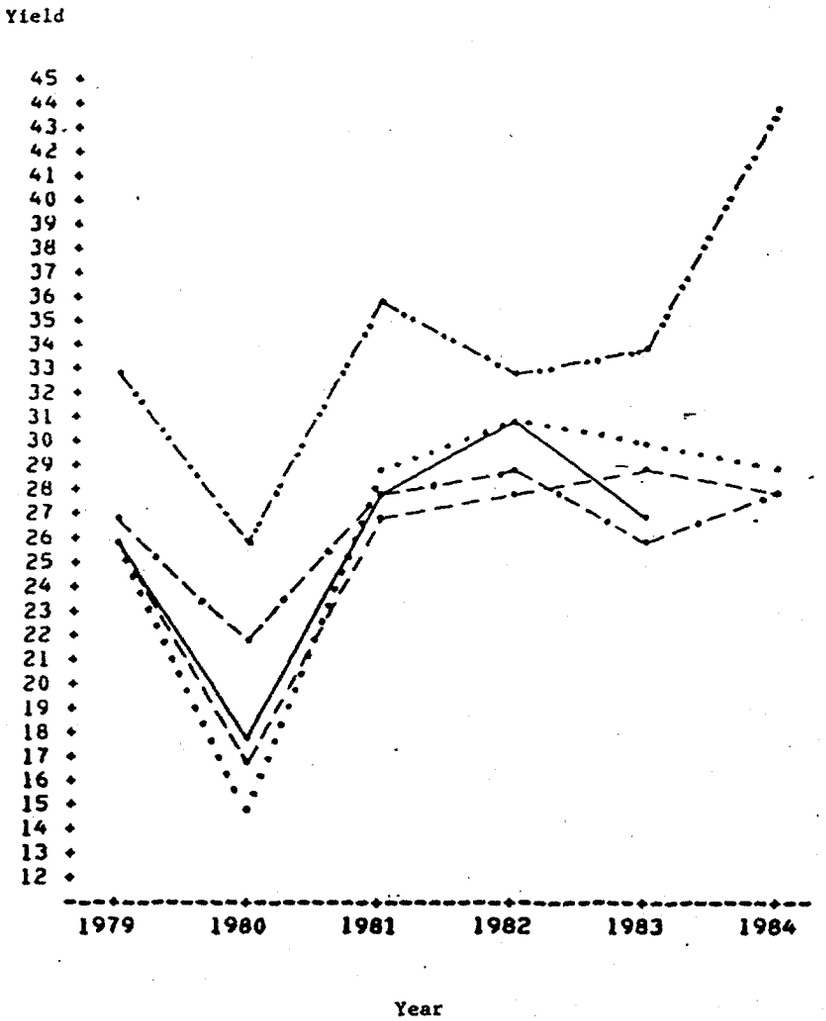

Figure 3. A wheat yield time series plot of four yield indications: A (----), B (• • •), C (• – –), D, (– • • –), 1979-1984, and the final official yield estimates (——), 1979-1983, for use in the first 1984 forecast. Indicated and actual yield is in bushels per acre.

mean square error for a period of years. Such an improvement will not result unless the bias squared contributes a substantial proportion to the mean square error. When the chart reading is effective, the chart read value can be used in lieu of the direct indication. When a chart reading fails to result in improvement, the direct indication should be used.

Once the various indications have been chart-read, another algorithm needs to be employed to objectively simulate the process of combining the indications, chart-read or not, in determining the official forecast. This algorithm should also be analogous to the thought process which practitioners use in arriving at

Table 3. Objective Simulated Chart Reading of Indication D for the First Forecast of Wheat Yield in 1984

Year	Final Yield	Yield Indication	Difference	|Difference-Median| 2/ (1/ |Difference-Median|)	Weight	Weight x Difference
1979	26.5	33.2	6.7 1/	0.5 (2.00)	.28	1.876
1980	18.5	25.6	7.1	0.5 (2.00)	.27	1.917
1981	28.0	35.6	7.6	0.9 (1.11)	.15	1.140
1982	31.0	33.0	2.0	4.7 (0.21)	.03	.060
1983	27.0	33.6	6.6	0.5 (2.00)	.27	1.782
				(7.32)	1.00	6.775 3/

1/ Median Difference
2/ Minimum |Difference-Median| = 0.5
3/ Adjustment to be subtracted from the 1984 Indication value of 43.7 in arriving at a chart-read value for that indication of 36.9 bushels per acre.

a forecast by considering the indications displayed on a time series chart. The method should mimic the practitioner's mental evaluation of how accurately in the past each indication has forecast the final official yield. It would be natural to combine the indications in such a way that greater emphasis is given to the more accurate indications.

Again, the method presented here was developed by Jeanne L. Sebaugh and the author. As a measure of the historic accuracy of an indication the computer algorithm uses the root mean square error (RMSE) between the indication and the final yield, computed for the five years previous to the forecast year. By representing the magnitude of an indications discrepancy from its target, the final yield, the RMSE indicates the reliability of the indication. A small RMSE indicates greater accuracy. Therefore, the inverse of the RMSE, 1/RMSE, is proportional to the accuracy of the indication. That is, the larger the inverse

is, the more accurate the indication has performed over the previous five years. The inverse of the root mean square error over that period is computed for either the chart-read or direct value for each quantifiable indication available for a particular forecast. The combined forecast indication or objectively simulated official forecast is then computed as the weighted average of the indications using the corresponding inverse of each indication's RMSE as the weight. Table 4 shows the calculations for the four currently used indications for the first forecast of wheat yield in 1984.

Table 4. Combined Forecast Indication or Objectively Simulated Official Forecast for the First Forecast of Wheat Yield in 1984

Indication	RMSE[1/]	1/RMSE[1/]	Weight[1/]	1984 Indication	Weight x Indication
A	1.86	.538	.296	28.2	8.347
B	2.31	.433	.238	29.5	7.021
C	1.92	.521	.286	28.3	8.094
D[2/]	3.05	.328	.180	36.9	6.642
		1.820	1.000		30.104[3/]

1/ The RMSE, 1/RMSE and the weight are computed for the previous five year period, 1979-1983. The RMSE is

$$\text{RMSE} = \left[\frac{1}{5} \sum_{1979}^{1983} (\text{First Forecast Indication} - \text{Final Yield})^2 \right]^{\frac{1}{2}}$$

2/ The chart-read indication is used only for indication D. The direct indication is used for the other indications.

3/ The result of the calculations, rounded to the nearest whole bushel (30 bushels per acre), is the combined first forecast indication for 1984 or, in general, it is one of the objectively simulated official yield forecasts.

Results of the two algorithm simulation programs have been applied to several different crops in different states. These results have been very encouraging. The simulated forecasts have generally agreed with the actual USDA

Crop Reporting Board forecasts and usually differ by less than two bushels per acre. Where they differ it is likely that some non-quantifiable information was utilized in determining the official forecasts, but as can be seen from the following table this may not have always resulted in improved accuracy. Table 5 presents some results on the correspondence between the simulations and the Crop Reporting Board published forecasts for selected crops and states. The correspondence is shown for various forecast dates over the 1980 through 1983 period in which some highly variable yields occurred. The final board estimate of the "true" yield is reported in the table.

EXTENSION OF THE SIMULATION METHOD IN ASSESSING THE IMPACT ON THE OFFICIAL FORECASTS OF USING A NEW MODEL OR ALTERING THE SET OF INDICATIONS IN OTHER WAYS

The method of simulating official crop yield forecasts, to the extent that it succeeds, opens up a multitude of possibilities for further analysis. It extends the capabilities of evaluating and comparing crop yield models and survey indications, to simulating what the probable impact would be on the official forecasts of using a promising new model, or indeed, altering the set of indications used in other ways. For the crop and state combinations shown in Table 5, the simulations are remarkably accurate. With some caution, it should be possible to simulate what might have happened (retrospectively) if the set of indications had been altered and from that analysis gain a good understanding of how an altered set of indications might affect the accuracy of Crop Reporting Board forecasts in the future.

Returning to the case of the first yield forecast of a wheat crop in a major producing U. S. state, let us examine changes that could be considered in altering the currently used set of four quantifiable indications. First, a promising new research model will be introduced for use along with the other

Table 5. Objectively Simulated Official Yield Forecasts and USDA Crop Reporting Board Actual Forecasts for Various Forecasts, and Final Crop Reporting Board Yield Estimates Bushels Per Acre, 1980-1983

State/ Crop	Year[2]	First Forecast[1] Simulated	First Forecast[1] Actual	Second Forecast Simulated	Second Forecast Actual	Third Forecast Simulated	Third Forecast Actual	Final Board Yield
Kansas	1980	33	32	35	34	34	34	35.0
Winter	1981	32	32	29	27	26	25	25.0
Wheat	1982	35	35	37	37	36	36	35.0
	1983	40	40	40	39	40	40	41.5
N. Dakota	1980	34	27	28	27	28	28	32.0
Barley	1981	49	47	49	47	48	48	48.0
	1982	48	47	50	49	52	52	53.0
	1983	52	53	50	50	44	44	45.5
N. Dakota	1980	18	18	18	18	18	19	19.0
Durum	1981	30	30	29	30	29	29	29.0
Wheat	1982	30	31	32	32	32	33	32.5
	1983	30	31	27	27	26	27	26.5
N. Dakota	1980	19	19	18	19	18	19	18.5
Spring	1981	29	28	28	28	28	28	28.0
Wheat	1982	29	29	31	32	32	31	31.0
(other than Durum)	1983	28	30	27	27	27	27	27.0

1/ Forecasts relate to the first of the month and are published around the tenth. They are issued monthly during the crop season with the following first forecast dates: winter wheat-May 1, barley-July 1, durum and other spring wheat-August 1.

2/ Simulation results for 1984 are not shown because in that year they were provided to the Statistical Reporting Service's State Offices in those states and may have been considered in determining the actual forecasts.

four indications. This will allow simulation of how the official forecasts over a period of years would be changed by simply adding the model to the set of yield indication tools. Then, various combinations of three of the currently used quantifiable indications will be considered to learn which of the

indications are most helpful in obtaining greater early forecast accuracy. These simulations allow one to understand how the official forecasts would change if the survey on which an indication is based were discontinued. Finally, the research model will be used along with the most promising combinations of the three current indications. This will allow investigation of the value of the research model when used along with the more useful current indications. An illustration of 1984 computations is shown in Table 6 for computing combined forecasts following the procedure outlined above with indication A selected for exclusion.

Table 6. Combined Forecasts Based on Alternative Model and Survey Indications for the First Forecast of Wheat Yield in 1984

Indication	RMSE[1/]	1/RMSE[1/]	Weight I[1/]	Weight II[1/]	Weight III[1/]
A	1.86	.538	.221	-	-
B	2.31	.433	.178	.338	.229
C	1.92	.521	.215	.406	.275
D[2/]	3.05	.328	.135	.256	.173
Research Model (RM)	1.64	.610	.251	-	.323
			1.000	1.000	1.000

Indication	1984 Indication	Weight I x Indication	Weight II x Indication	Weight III x Indication
A	28.2	6.232	-	-
B	29.5	5.251	9.971	6.776
C	28.3	6.084	11.490	7.782
D[2/]	36.9	4.982	9.446	6.384
RM	26.4	6.626	-	8.527
		29.175[3/]	30.907[3/]	29.469[3/]

1/ RMSE, 1/RMSE and the weights are computed for the previous five year period, 1979-1983.

2/ The chart-read indication is used only for indication D.

3/ The result of the calculations based on use of all five indications; indications B, C and D; and on B, C and D along with the research model are (rounded to whole bushels) 29, 31 and 29, respectively.

To assess the performance of various combinations of indications, it is desirable to repeat the simulation procedure over more than one year. The performance of the first official forecast, simulated forecast based on use of the current set of indications (simulated official forecast) and simulated forecasts based on various changes that could be considered in altering the set of currently used quantifiable indications are examined for a five year period. In Table 7, performance of the actual forecast and various simulated forecasts are reported in terms of the root mean square error, largest difference from the final yield and bias for the 1980 through 1983 and the 1980 through 1984 periods. In Figure 4, departures of selected forecasts from the final Crop Reporting Board yield estimate (forecast minus estimate) are shown for the five year test period. For each year this includes the first Board forecast and the simulated combined forecasts based on the (1) current indications, (2) current indications supplemented by the research model and (3) most promising set of three currently used indications.

From the figure it can be seen that the published and simulated forecasts based on the current indications are similar. They are identical in two years, differ by a bushel in two and by two bushels in one year (1983). As can be seen in Table 7, the simulated forecasts correspond more closely to the final estimates, both when 1984 is included and excluded, than the official first forecast. It can also be discerned that forecasts based on the current indications as supplemented by the research model result in a slight improvement in accuracy (as compared to the current indications alone) for the first four years, but that use of the model is somewhat detrimental to accuracy in 1984. The forecasts from the reduced set of indications, that were found to be the most promising over the five year period, (B, C, D) provide identical results to the full set of current indications for three years. But, this reduced set of indications

Table 7. Selected Yield Reliability Measures for the Official Yield Forecast and Simulated Forecasts Based on the Combined Use of the Current Indications and Various Combinations of Indications That Could be Considered in Altering the Set of Quantifiable Indications for the First Wheat Yield Forecast in a Major Producing U.S. State, 1980-1983 and 1980-1984

Forecast	Measures of Reliability (bushels per acre) 1980 - 1983			1980 - 1984		
	Root Mean Square Error	Largest Difference	Bias	Root Mean Square Error	Largest Difference	Bias
Actual Board Forecast	1.8	3	0.4	2.8	-5	-0.7
Simulated Board Forecast - Indications A,B,C,D	1.2	-2	0.1	2.1	-4	-0.7
Other Simulated Forecasts						
- Indications A,B,C,D,RM	1.0	-2	-0.4	2.4	-5	-1.3
- Indications A,B,C	1.4	±2	-0.1	2.6	-5	-1.1
- Indications A,B,D	1.5	±2	0.1	1.9	-3	-0.5
- Indications A,C,D	1.8	-3	-0.1	2.4	-4	-0.9
- Indications B,C,D	0.9	±1	0.4	1.6	-3	-0.3
- Indications A,B,D,RM	1.1	-2	-0.1	2.5	-5	-1.1
- Indications B,C,D,RM	1.1	-2	-0.1	2.5	-5	-1.1

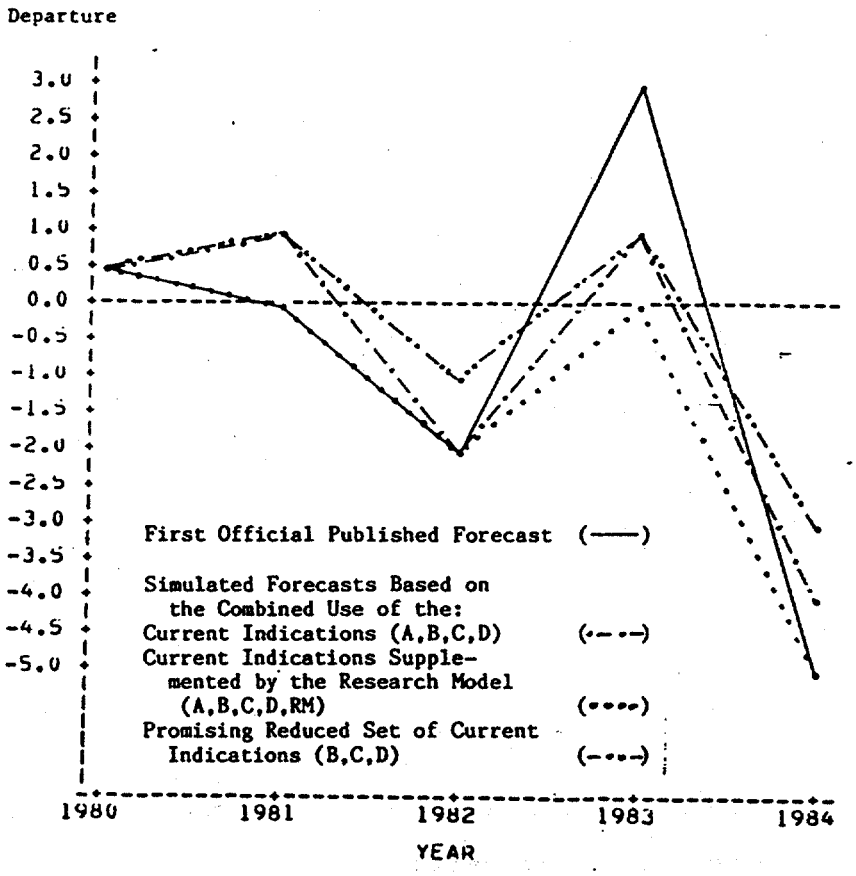

Figure 4. Departures of actual and simulated forecasts, based on selected alternative combinations of model and survey indications, from the final official yield estimate (forecast minus final) for the first wheat yield forecast in a major producing U. S. state, 1980-1984. Departures are in bushels per acre.

provides a more accurate forecast (by one bushel) in each of the two years when the full current set exhibits the most inaccuracy. This greater accuracy for simulated forecasts based on use of indications B, C, and D is also reflected in Table 7. The other reduced set of three current indications exhibiting some improvement, over the entire five-year period, from the full set of current indications was based on use of indications A, B and D. When the research model is used to supplement these two reduced sets of current indications, it is found to be less effective than using the model along with the full set of current indications.

In 1984 the final yield estimate was much higher than all of the indications, except indication D. The year was very different in many respects and was characterized by a substantial rebound in the acreage of many crops from the government program reduction of the previous year. The research model performed poorly, particularly when contrasted with its excellent performance over a previous ten year test period. As such, the 1984 experience and the additional data it provides present an opportunity for improving the model and the 1984 results suggest such an improvement is needed if the model is to consistently contribute to more accurate Crop Reporting Board forecasts.

CONCLUSION

Evaluation of crop yield models involves the evaluation of individual models for eight criteria that desirable models would possess. Because there are only comparative standards against which individual models can be assessed, the comparison of alternative or competing models is required to identify the best available models. These comparisons may concentrate on yield indication reliability, as presented in this paper, but the other seven criteria are also important and should be considered. Finally, the best available model may not be good enough to make a positive difference in the official forecasts. Or a promising model, thought to be inferior to one still being "perfected," may be very useful in improving the published forecasts. The ability to simulate forecasts based on the current indications, allows "what if" questions about the use of these models to be answered. It also allows one to assess the impact of changes in the set of indications, which may be entirely apart from considering the use of any new models or methods. These questions have been difficult to address in an objective way, particularly when "someone's" indication along with the survey from which it is derived is considered for elimination.

Nonparametric Test - Wilcoxon Signed Rank:

H_0: One model does not perform better than the other model.

H_a: One model performs better than the other model.

Procedure to compute test statistic, T:

1. Compute the D_i.
2. Assign ranks to $|D_i|$.
3. Assign signs to Rank ($|D_i|$) corresponding to the signs of D_i.
4. Let T = the absolute value of the sum of the ranks with the less frequent sign (corresponding to non-zero D_i).

Reject H_0 if $T < T_{\alpha}$ (1 tailed),n.

Paired-Sample Statistical Tests Comparing the Performance of Two Crop Yield Models

Definition of Terms

$\hat{Y}_{1_i}$ = Yield as predicted by model 1 for year i.

Y_{2_i} = Yield is predicted by model 2 for year i.

$d_{1_i} = \hat{Y}_{1_i} = Y_i$ = Difference between model 1 predicted and actual yield for year i.

$d_{2_i} = \hat{Y}_{2_i} - Y_i$ = Difference between model 2 predicted and actual yield for year i.

$D_i = |d_{1_i}| - |d_{2_i}|$.

Rank ($|D_i|$) = Ranks of the absolute values of D_i assigned in ascending order (smallest value of $|D_i|$ = rank 1, ..., largest value of $|D_i|$ = rank n). If two or more years have the same value for $|D_i|$, assign each year the average of the ranks.

Parametric Test - Student t:

$H_0: \mu_D = 0$

$H_a: \mu_D \neq 0$

Test Statistic = $t = \dfrac{\overline{D}}{s_{\overline{D}}}$, where

$\overline{D} = 1/n \Sigma D_i$,

$s_{\overline{D}} = (s_D^2/n)^{\frac{1}{2}}$, and

$s_D^2 = \{ \Sigma D_i^2 - 1/n (\Sigma D_i)^2 \} / (n-1)$.

Reject H_0 if $t > t_{\alpha, (n-1)}$.

APPENDIX

Measures of Model Performance

Definition of Terms:

Y_i = Yield as reported by U.S.D.A. for year i ("true" or "actual" yield).

$\hat{Y}_i$ = Yield as predicted by a model for year i.

$d_i = \hat{Y}_i - Y_i$ = difference between predicted and actual yield for year i.

$RD_i = 100\ d_i/Y_i$ = relative difference for year i.

$i = 1,\ldots, n$ = number of test years and $\Sigma = \sum_{i=1}^{n}$ = summation over the test years.

$\overline{Y} = 1/n\ \Sigma\ Y_i$ = average actual yield.

Measures of Yield Indication Reliability:

Bias = $B = 1/n\ \Sigma\ d_i = \overline{d}$.

Relative Bias = $RB = 100\ B/\overline{Y}$.

Mean Square Error = $MSE = 1/n\ \Sigma\ d_i^2$.

Root Mean Square Error = $RMSE = (MSE)^{\frac{1}{2}}$.

Relative Root Mean Square Error = $RRMSE = 100\ RMSE/\overline{Y}$.

Variance = $Var = 1/n\ \Sigma\ (d_i - \overline{d})^2$.

Standard Deviation = $SD = (Var)^{\frac{1}{2}}$.

Relative Standard Deviation = $RSD = 100\ SD/(\overline{Y} + \overline{d})$.

Mean Square Error = Variance + $(Bias)^2$.

Pearson r between $\hat{Y}_i$ and Y_i:

$$r = \frac{\left[\Sigma \hat{Y}_i Y_i - \frac{(\Sigma \hat{Y}_i)(\Sigma Y_i)}{n}\right]}{\left[\left(\Sigma \hat{Y}_i^2 - \frac{(\Sigma \hat{Y}_i)^2}{n}\right)\left(\Sigma Y_i^2 - \frac{(\Sigma Y_i)^2}{n}\right)\right]^{\frac{1}{2}}}$$

Not all models need to be individually evaluated (see Barnett, et al., 1980); many of those evaluated require only the most superficial comparison before they are relegated to the "also-ran" category, and only promising models need to be assessed for their impact on official forecasts. Unfortunately, only models or methods which can be retrospectively simulated based on the necessary data inputs can be assessed for their impact on the official forecasts. However, methods based on field or sample data collection can be readily assessed for the impact of their discontinuance.

REFERENCES

Barnett, Thomas L., Sharon K. LeDuc, Raymond P. Motha, Wendell W. Wilson, 1980. Criteria for Identifying Candidate Yield Models. AgRISTARS Yield Model Development Project, Document YMD-1-1-1(80-2.1).

Sebaugh, Jeanne L., 1981a. Evaluation of "Straw Man" Model 1, The Simple Linear Model, for Soybean Yields in Iowa, Illinois and Indiana. AgRISTARS Yield Model Development Project, Document YMD-1-3-2(81-03.1) and Statistical Reporting Service Staff Report AGESS810304.

Sebaugh, Jeanne L., 1981b. Comparison of One, Two and Three Line Segment "Straw Man" Models for Soybean Yields in Iowa, Illinois and Indiana. AgRISTARS Yield Model Development Project, Document YMD-1-3-2(81-05.1) and Statistical Reporting Service Staff Report AGESS810514.

Statistical Reporting Service, U. S. Department of Agriculture, 1983. Scope and Methods of the Statistical Reporting Service. Miscellaneous Publication No. 1308.

Wilson, Wendell W., Thomas L. Barnett, Sharon K. LeDuc, Fred B. Warren, 1980. Crop Yield Model Test and Evaluation Criteria. AgRISTARS Yield Model Development Project, Document YMD-1-1-2(80-2.1).

Wilson, Wendell W. and Jeanne L. Sebaugh, 1981. Established Criteria and Selected Methods for Evaluating Crop Yield Models in the AgRISTARS Program. American Statistical Association 1981 Proceedings of the Section on Survey Research Methods, 24-31.