THE EFFECT OF WEATHER STATION DENSITY ON CROP YIELD FORECASTS

M. Denice McCormick

Follow this and additional works at: http://newprairiepress.org/agstatconference

Part of the Agriculture Commons, and the Applied Statistics Commons

This work is licensed under a Creative Commons Attribution-Noncommercial-No Derivative Works 4.0 License.

Recommended Citation

This is brought to you for free and open access by the Conferences at New Prairie Press. It has been accepted for inclusion in Conference on Applied Statistics in Agriculture by an authorized administrator of New Prairie Press. For more information, please contact cads@k-state.edu.
The National Agricultural Statistics Service (NASS) uses regression models to forecast yield for crops such as corn, soybeans and winter wheat. Analyses were conducted on the use of precipitation data in these regression models (McCormick and Birkett 1992, and McCormick 1993). Precipitation data are obtained from two sources. The National Climatic Data Center (NCDC) supplies historic precipitation data used for developing regression model parameters. The Climate Analysis Center (CAC) supplies current year precipitation data that are used as regression model input. CAC weather station density is sparse across the U.S. in many major agricultural areas compared to NCDC weather station density. As a result, significant differences exist between NCDC and CAC regional precipitation terms for corn and winter wheat. This paper evaluates the effect on forecast accuracy when only CAC data are used versus NCDC data for model development. Results indicate that regional corn, soybean and winter wheat models based on CAC data are just as accurate as NCDC based models.

KEY WORDS: Precipitation data, regression models, forecast accuracy.

I. INTRODUCTION

The National Agricultural Statistics Service (NASS) evaluated the addition of precipitation terms to regression models used to forecast crop yields at the regional level (McCormick and Birkett, 1992, and McCormick, 1993). Previously, these models were based solely on survey data related terms. The survey data were collected from randomly selected sample plots in randomly selected fields in States within each crop region. For fall harvested crops, the surveys are conducted monthly starting August 1. For winter wheat, the initial monthly survey is conducted May 1. Linear or quadratic regression models are developed for each month and crop, based on historic data relating a regional survey variable to final yield. Data from the current survey are then aggregated to the regional level and used in the model to forecast final yield.

NASS used two different sources of precipitation data for this research effort. The Climate Analysis Center (CAC) of the National Weather Service provides data in a real-time mode for the current year from a fairly sparse network of about eight to twelve automated stations per state. The National Climatic Data Center (NCDC) of the National Weather Service provides historic data from a much denser network of about 80 to 120 stations per state, about ten times the number that CAC provides, with almost one station in every county. The NCDC network includes the CAC automated station network plus a large number of other manually operated stations. However, there is about a four month lag until this data becomes available. In previous research and applications of this approach by NASS, historic NCDC data were used to develop the regression model parameters and then CAC data were used as input to the model for the current year forecast. NCDC data for the current year were obtained when available to develop new models for the next year.

The sparse density of CAC stations and the location of stations in relation to crop growing areas are concerns to NASS regarding its use in crop yield forecasting. However, the advantages of easy access and real-time availability of CAC data are attractive. The sparse CAC network within states still provides 85 stations for current year corn
forecasting, 71 stations for soybean forecasting, and 179 stations for winter wheat forecasting at the regional (multi-state) level. The average number of stations per State is approximately nine for corn and soybean States and eleven for winter wheat States. Most States have at least one CAC station per Agricultural Statistics District (ASD). An ASD, previously called "Crop Reporting District", is a region consisting of contiguous counties within each State that is characterized by similar agricultural characteristics, such as homogeneous soil, cropping practices and climatic patterns (Arends, W., etal 1983). One station per ASD is assumed to provide adequate geographic coverage for forecasting regional crop yields.

This study examines the reliability of regional level models developed using data from CAC stations, as compared to NCDC-based models, for corn, soybeans, and winter wheat. It considers data for twelve years, 1980 to 1991. Corn States covered are Illinois (IL), Indiana (IN), Iowa (IA), Michigan (MI), Minnesota (MN), Missouri (MO), Nebraska (NE), Ohio (OH), South Dakota (SD), and Wisconsin (WI). Soybean States are Arkansas (AR), Illinois (IL), Indiana (IN), Iowa (IA), Minnesota (MN), Missouri (MO), Nebraska (NE), and Ohio (OH), and winter wheat States are Arkansas (AR), Colorado (CO), Idaho (ID), Illinois (IL), Indiana (IN), Kansas (KS), Missouri (MO), Montana (MT), Nebraska (NE), Ohio (OH), Oklahoma (OK), Oregon (OR), South Dakota (SD), Texas (TX), and Washington (WA).

To aid in the interpretation of model evaluations, differences between CAC and NCDC regional and State level precipitation terms were examined. This type of comparison had been done previously. Kaiser and Sebaugh (1984) used a paired t-test to compare real-time and historic weather data to forecast barley and spring wheat yields in North Dakota from 1970 to 1979. Historic values were derived from NCDC data. Current year values were derived from data from the Climate Analysis Branch (CAB), a forerunner of the CAC. Kaiser and Sebaugh concluded that the values derived from real time data were not equivalent to NCDC derived values. Therefore, it was not practical to use CAB data for real-time crop yield forecasting at that time.

II. DATA

Current year precipitation data are obtained directly from the computers at the Climate Analysis Center (CAC) of the National Weather Service each month over a modem to a personal computer. CAC is the best known source of real-time data available. Data obtained, however, are preliminary, characterized by few edit or quality control checks. Automated weather stations that supply the data are sparsely located across the country with a tendency to be located at airports and Army or Air Force airfields, in urban rather than agricultural areas. The network is capable of providing hourly, daily, weekly or monthly accumulated data. NASS uses CAC's monthly summarized data for each weather station. It then summarizes the data to monthly totals on ASD, State, and regional levels. Table 1 (Appendix A) provides a summary of CAC station coverage by crop and State based on 1991 harvested acreage. The coverage statistic \( C_s \) is the percentage of harvested acres by crop within a State that is represented by CAC weather stations out of the total State harvested acres:

\[
C_s = \frac{\sum_{d=1}^{D_s} I_{sd}(A_{sd})}{100} \quad \text{for crop} \quad \text{and State} \quad \text{with CAC weather stations}
\]
where

\( C_s \) = the State coverage percentage for State \( s \) for a particular crop,

\( D_s \) = number of districts in State \( s \),

\( I_{sd} \) = 1, for State \( s \), district \( d \) if the district is covered by an automated station,

\( 0 \), otherwise, and

\( A_{sd} \) = total harvested acres, for state \( s \), district \( d \).

Harvested acres of a district are considered covered if the district contains a CAC station. The under-coverage statistic is then \( (100 - C_s) \). Regional coverage statistics \( C_r \) are also provided:

\[
C_r = \frac{\sum_{s=1}^{S} \sum_{d=1}^{D_s} I_{sd} A_{sd}}{\sum_{s=1}^{S} \sum_{d=1}^{D_s} A_{sd}} \times 100
\] (2.2)

where

\( S \) = number of states in region for a particular crop.

Historic weather data are supplied by the National Climatic Data Center (NCDC). In addition to the automated station reports, this data set includes reports from manually operated stations. Manual operators are often volunteer local weather reporters such as local farmers and agricultural extension agents. NCDC also uses quality control measures to edit the data, but cannot provide the data in a real-time mode for use in current year crop yield forecasts.

The precipitation data are summarized to regional totals for particular months or groups of months (McCormick 1993). Different months are used for different crops and monthly forecasts as documented in Table 2 (Appendix A).

The survey data collected in the survey are also aggregated to the regional level. The procedures used to aggregate both the precipitation and survey data are documented in Appendix B.

III. METHODOLOGY

The multiple linear regression models used for forecasting crop yields are of the form:

Model 1: \( Y_t = \beta_0 + \beta_1 Z_t + \epsilon_t \)

Model 2: \( Y_t = \beta_0 + \beta_1 Z_t + \beta_2 Z_t^2 + \epsilon_t \)

Model 3: \( Y_t = \beta_0 + \beta_1 Z_t + \beta_2 P_t + \epsilon_t \)

Model 4: \( Y_t = \beta_0 + \beta_1 Z_t + \beta_2 Z_t^2 + \beta_3 P_t + \epsilon_t \)

where:
Applied Statistics in Agriculture

the regional aggregated precipitation variable as defined in Appendix B, and

the regional aggregated survey variable involving plot counts and measurements in randomly selected fields, as defined in Appendix B.

Model 2 is the official model used by NASS to forecast August corn and soybeans and September soybeans. However, Model 1 is the official model used to forecast September corn and May winter wheat.

Models 3 and 4 each use one monthly precipitation term. Analysis was conducted previously (McCormick 1993) to determine which month or combination of months from the growing season provided optimal forecasting capability. Also, models with multiple monthly precipitation terms were examined. In order to evaluate the reliability of the CAC models, the regression parameters for Models 3 and 4 were developed:

1. based on full NCDC data 1980-1991, and

So the CAC subset used for this analysis reflects the density of the CAC network.

The reliability of the models was then evaluated based on the following criteria.

Model Evaluation Criteria

The primary model evaluation criterium is the set of prediction intervals (PI) that correspond to the years with the minimum, median, and maximum yields over the twelve years in the study for each crop. These years for corn were 1983, 1989, and 1986, respectively; for soybeans 1988, 1981, and 1990, respectively; and for winter wheat were 1989, 1985, and 1983 respectively. A second criterium is the adjusted coefficient of determination, R^2, which provides a measure of correspondence between predicted and actual yields. Both the PI and R^2 are based on the sum of squared differences from the least squares analysis used to derive the model parameters.

1. The prediction interval (PI) refers to half the confidence interval length for the predicted value of a future \( Y \) for a given future year \( o \).

That is, at the \( \alpha \) significance level:

\[
PI = t(1-\frac{\alpha}{2};n-1-p)SD(\hat{\beta}_o),
\]

where

\[
SD(\hat{\beta}_o) = s[(X'_o(X'_oX_o)^{-1}X_o) + 1]^{\frac{1}{2}}.
\]

\( s = (\text{residual MSE})^{\frac{1}{2}}, \)
relevant p-dimensional row vector of independent variables for year 0 (for example, in Model 3: \( p=3, x_o = [1, Z_o, P_o] \)).

\( x_o \) = relevant \((n-1) \times p\) matrix of independent variables (excludes \( x_o \)),

\( n \) = number of years, and

\( p \) = number of parameters.

The \( x_o \) matrix excludes the row vector \( x_o \), so that the PI reflects the accuracy expected in an operational model where current year data are not included in the model development. A significance level of 0.32 provided \( t \) values near 1.0. Consequently, the future \( Y \) will fall within the calculated PI of the predicted \( Y \) approximately 68% of the time.

2. \( R_a^2 \) is used as a goodness-of-fit test for each model with an adjustment made for the corresponding degrees of freedom (Draper and Smith 1981).

\( R_a^2 \) is calculated as:

\[
R_a^2 = 1 - \frac{(RSS_p)/(n-p)}{(CTSS)/(n-1)}, \tag{3.3}
\]

where

\( RSS_p \) = the residual sum of squares taking the changing number of parameters into account,

\( CTSS \) = the corrected total sum of squares,

\( n \) = the number of years, and

\( p \) = the number of parameters.

**Outlier Identification**

Since the purpose of the models is to make forecasts, the \( r_{student} \) statistic (also called the studentized residual) was used to help identify outliers to be excluded from the model (Belsley, Kuh and Welsh 1980). It is similar to the standardized residual:

\[
I_{si} = \frac{I_i}{s\sqrt{1-h_i}}, \tag{3.4}
\]

where

\( r_i \) = \( i^{th} \) residual,

\( s \) = (residual MSE)\(^{1/2} \), and

\( h_i \) = \( x_i'(X'X)^{-1}x_i \)

S is replaced by \( s(i) \). \( s(i) \) is the estimate of \( \sigma \) with the \( i^{th} \) observation deleted.

In a forecasting model, \( r_{student} \) measures how many prediction standard errors the forecast is from the observed \( Y \). Observations with absolute values of studentized greater than 3.0 were identified as outliers. The \( r_{student} \) statistic is distributed closely to the \( t \)-distribution with \( n-p-1 \) degrees of freedom (See Notes for Tables 3, 4 and 5 (Appendix A) on outlier removal).
Precipitation Term Comparisons

For additional insight into model interpretations, paired t-tests were used to compare regional monthly precipitation totals calculated from the full standard NCDC data set to totals calculated from CAC automated stations from 1980 to 1991. Comparisons were also made within regions for individual years over States and for individual States over years. Refer to Tables 6, 7, 8 (Appendix A).

IV. RESULTS

Tables 3, 4 and 5 (Appendix A) present the prediction intervals and \( R^2 \) for each crop comparing the official linear or quadratic model using survey data only versus the addition of an NCDC-based or CAC-based optimal monthly precipitation term. In each table, the prediction intervals relate to the years with minimum, median, and maximum regional yields.

The prediction intervals and \( R^2 \) values from the NCDC and CAC models tend to be very similar for corn and soybeans. Surprisingly, the wheat CAC model has smaller prediction intervals than the NCDC model, even though the CAC model is based on less precipitation data. The forecasting models containing a precipitation term provide smaller prediction intervals for corn and winter wheat than the official models. The \( R^2 \) and PI values for soybeans are similar for precipitation and official models.

Mean differences between the NCDC and CAC monthly precipitation totals for the regions (over years), years (over States) and States (over years) were calculated. Differences were calculated by subtracting CAC from NCDC values.

Tables 6, 7, and 8 (Appendix A) present the results at the regional level for the monthly precipitation terms used in the forecast models, and then the results for years or States that had significant differences for each crop.

V. SUMMARY

Regional crop yield forecast models using precipitation terms based on CAC data are more desirable than NCDC based models since CAC data are available real-time at low cost. This assumes real-time unedited CAC data have similar accuracy and quality as the edited NCDC data. This analysis indicates that just due to the density of the stations, significant differences do exist between terms derived from both data sources. NCDC has a much denser network of stations having almost ten times the number of the CAC network. A valid concern was that this could create problems in forecast accuracy.

Precipitation terms were derived from both sources to develop model parameters for making corn, soybean and winter wheat yield predictions. Results from this study showed:

1. Regional crop yield models developed using CAC based terms were just as accurate as those based on NCDC terms. CAC models capture the impact of precipitation on final yield comparably to NCDC models even though CAC values are based on only about one-tenth the number of stations as NCDC values and only provide coverage on approximately 70% of crop harvested acres.

2. CAC regional precipitation values are significantly different than NCDC values for corn and winter wheat at the \( \alpha = 0.10 \) level of significance. Corn mean differences for June and July, at 0.09
inches for both months are relatively small and of no practical importance. The May forecast period winter wheat mean difference is larger, at 0.39 inches. Soybean mean differences are 0.09 and 0.02 inches for the August and September forecasts, respectively. Neither is significantly different.

3. Using CAC-derived precipitation terms, forecast accuracy was improved 47% for the first forecast in May for winter wheat and 21% for the first forecast in August for corn over official model results that use only survey data derived terms. There was a 43% improvement in forecast accuracy for the second forecast of corn yield in September. No improvement occurred for either the first or second forecast for soybeans in August or September over results from the official models.

Recommendations

The density of the CAC network appears to be adequate for making regional yield forecasts in spite of the fact that the network is based on a reduced number of stations that provide about 70% coverage over Agricultural Statistics Districts. This conclusion is the basis of the following recommendations:

1. The CAC should continue to strive to improve coverage. In the meantime, NASS should evaluate possible gains in forecast accuracy from data imputation schemes whereby data from missing districts are imputed based upon data from nearby or surrounding automated stations.

2. State level models using both data sources should be evaluated prior to the implementation of CAC (data only) models to ensure that CAC station coverage is as adequate on the State level as it is on the regional level.

3. Also, before discontinuing access to NCDC data, it is necessary to evaluate the quality of CAC values by comparing CAC data to edited data now available for 1992.

VI. REFERENCES


VII. APPENDIX A

Table 1: CAC Coverage

<table>
<thead>
<tr>
<th>State</th>
<th>RANK 91¹</th>
<th>#Stations</th>
<th>C¹</th>
<th>100-C¹</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CORN:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IL</td>
<td>2</td>
<td>10</td>
<td>64</td>
<td>36</td>
</tr>
<tr>
<td>IN</td>
<td>5</td>
<td>6</td>
<td>69</td>
<td>31</td>
</tr>
<tr>
<td>IA</td>
<td>1</td>
<td>8</td>
<td>87</td>
<td>13</td>
</tr>
<tr>
<td>MI</td>
<td>17</td>
<td>7</td>
<td>54</td>
<td>46</td>
</tr>
<tr>
<td>MN</td>
<td>4</td>
<td>10</td>
<td>62</td>
<td>38</td>
</tr>
<tr>
<td>MO</td>
<td>3</td>
<td>9</td>
<td>79</td>
<td>21</td>
</tr>
<tr>
<td>OH</td>
<td>13</td>
<td>74</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td>8</td>
<td>65</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>WI</td>
<td>7</td>
<td>65</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td><strong>REGION TOTAL:</strong></td>
<td><strong>85</strong></td>
<td><strong>72</strong></td>
<td><strong>28</strong></td>
<td></td>
</tr>
<tr>
<td><strong>SOYBEANS:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AR</td>
<td>8</td>
<td>54</td>
<td>46</td>
<td></td>
</tr>
<tr>
<td>IL</td>
<td>1</td>
<td>10</td>
<td>54</td>
<td>46</td>
</tr>
<tr>
<td>IN</td>
<td>4</td>
<td>6</td>
<td>70</td>
<td>30</td>
</tr>
<tr>
<td>IA</td>
<td>2</td>
<td>8</td>
<td>86</td>
<td>14</td>
</tr>
<tr>
<td>MN</td>
<td>3</td>
<td>7</td>
<td>46</td>
<td>54</td>
</tr>
<tr>
<td>MO</td>
<td>10</td>
<td>59</td>
<td>41</td>
<td></td>
</tr>
<tr>
<td>NE</td>
<td>9</td>
<td>74</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>OH</td>
<td>5</td>
<td>77</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td><strong>REGION TOTAL:</strong></td>
<td><strong>71</strong></td>
<td><strong>65</strong></td>
<td><strong>35</strong></td>
<td></td>
</tr>
<tr>
<td><strong>WINTER WHEAT:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AR</td>
<td>8</td>
<td>53</td>
<td>47</td>
<td></td>
</tr>
<tr>
<td>CO</td>
<td>4</td>
<td>10</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>ID</td>
<td>5</td>
<td>73</td>
<td>27</td>
<td></td>
</tr>
<tr>
<td>IL</td>
<td>10</td>
<td>38</td>
<td>62</td>
<td></td>
</tr>
<tr>
<td>IN</td>
<td>6</td>
<td>77</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>KS</td>
<td>12</td>
<td>86</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>MO</td>
<td>10</td>
<td>58</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>MT</td>
<td>11</td>
<td>100</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>NE</td>
<td>9</td>
<td>72</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>OH</td>
<td>13</td>
<td>76</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>OK</td>
<td>2</td>
<td>9</td>
<td>83</td>
<td>17</td>
</tr>
<tr>
<td>OR</td>
<td>12</td>
<td>54</td>
<td>46</td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td>8</td>
<td>66</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>TX</td>
<td>3</td>
<td>41</td>
<td>75</td>
<td>25</td>
</tr>
<tr>
<td>WA</td>
<td>15</td>
<td>100</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><strong>REGION TOTAL:</strong></td>
<td><strong>179</strong></td>
<td><strong>79</strong></td>
<td><strong>21</strong></td>
<td></td>
</tr>
</tbody>
</table>

¹ Rank 91 = rank of 5 largest states based on 1991 harvested acres.
² State weather station coverage percentage as defined in Equation 2.1.
Table 2: Precipitation Term Month

<table>
<thead>
<tr>
<th>Crop</th>
<th>Forecast Date</th>
<th>Months used in Precipitation Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corn</td>
<td>August 1</td>
<td>July</td>
</tr>
<tr>
<td></td>
<td>September 1</td>
<td>June</td>
</tr>
<tr>
<td>Soybeans</td>
<td>August 1</td>
<td>July</td>
</tr>
<tr>
<td></td>
<td>September 1</td>
<td>August</td>
</tr>
<tr>
<td>Winter Wheat</td>
<td>May 1</td>
<td>August (lag 1 year) + March (current year)</td>
</tr>
</tbody>
</table>

Table 3: Corn Model Forecast Results

<table>
<thead>
<tr>
<th>Model</th>
<th>$R_a^2$</th>
<th>Prediction Intervals</th>
<th>min</th>
<th>med</th>
<th>max$^1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUGUST:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OFFICIAL</td>
<td>.88</td>
<td></td>
<td>6.5</td>
<td>5.3</td>
<td>5.3</td>
</tr>
<tr>
<td>JUL NCDC</td>
<td>.92</td>
<td></td>
<td>5.4</td>
<td>4.4</td>
<td>4.6</td>
</tr>
<tr>
<td>JUL CAC</td>
<td>.92</td>
<td></td>
<td>5.4</td>
<td>4.2</td>
<td>4.4</td>
</tr>
<tr>
<td>SEPTEMBER:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OFFICIAL</td>
<td>.93</td>
<td></td>
<td>4.8</td>
<td>4.0</td>
<td>4.1</td>
</tr>
<tr>
<td>JUN NCDC</td>
<td>.98</td>
<td></td>
<td>2.5</td>
<td>2.1</td>
<td>2.1</td>
</tr>
<tr>
<td>JUN CAC</td>
<td>.98</td>
<td></td>
<td>2.7</td>
<td>2.3</td>
<td>2.3</td>
</tr>
</tbody>
</table>

MINIMUM: 1983 82.0 bu.
MEDIAN: 1989 120.4 bu.
MAXIMUM: 1986 125.2 bu.

$^1$ Prediction intervals as defined in Equation 3.1 are reported for the years where the minimum, median and maximum crop yield occurred. Those years are listed above. For corn, each prediction interval is reported with outlier year 1988 removed from the models.
### Table 4: Soybean Model Forecast Results

<table>
<thead>
<tr>
<th>Model</th>
<th>$R_a^2$</th>
<th>Prediction Intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>min</td>
</tr>
<tr>
<td>AUGUST:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OFFICIAL</td>
<td>.63</td>
<td>2.9</td>
</tr>
<tr>
<td>JUL NCDC</td>
<td>.64</td>
<td>3.0</td>
</tr>
<tr>
<td>JUL CAC</td>
<td>.62</td>
<td>3.1</td>
</tr>
<tr>
<td>SEPTEMBER:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OFFICIAL</td>
<td>.93</td>
<td>1.4</td>
</tr>
<tr>
<td>AUG NCDC</td>
<td>.92</td>
<td>1.5</td>
</tr>
<tr>
<td>AUG CAC</td>
<td>.92</td>
<td>1.6</td>
</tr>
</tbody>
</table>

**MINIMUM:** 1988 27.8 bu.  
**MEDIAN:** 1983 29.0 bu.  
**MAXIMUM:** 1990 37.4 bu.

$^1$ Prediction intervals as defined in Equation 3.1 are reported for the years where the minimum, median and maximum crop yield occurred. Those years are listed above. No outlier years were detected for soybeans.

### Table 5: Winter Wheat Model Forecast Results

<table>
<thead>
<tr>
<th>Model</th>
<th>$R_a^2$</th>
<th>Prediction Intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>min</td>
</tr>
<tr>
<td>MAY:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OFFICIAL</td>
<td>.88</td>
<td>1.7</td>
</tr>
<tr>
<td>AUG+MAR NCDC</td>
<td>.94</td>
<td>1.2</td>
</tr>
<tr>
<td>AUG+MAR CAC</td>
<td>.97</td>
<td>0.9</td>
</tr>
</tbody>
</table>

**MINIMUM:** 1989 32.9 bu.  
**MEDIAN:** 1985 37.5 bu.  
**MAXIMUM:** 1983 42.4 bu.

$^1$ Prediction intervals as defined in Equation 3.1 are reported for the years where the minimum, median and maximum crop yield occurred. Those years are listed above. For winter wheat, each prediction interval is reported with outlier year 1987 removed from the models.
### Table 6: Paired Comparison t-test Corn

| MONTH | MEAN DIFFERENCE | t-VALUE | Prob>|t| | Cr |
|-------|----------------|---------|--------|-----|
| REGION: | | | | |
| JUN | 0.09 | 1.9 | 0.09 | 72% |
| JUL | 0.09 | 3.7 | 0.00 | |
| YEARS: | | | | |
| 1982 | JUN | 0.16 | 2.2 | 0.05 | 72% |
| 1985 | JUN | 0.09 | 1.9 | 0.09 | |
| STATES: | | | | |
| IN | JUN | 0.19 | 2.2 | 0.05 | 69% |
| WI | JUN | 0.17 | 1.8 | 0.10 | 65% |
| | JUL | 0.33 | 2.1 | 0.06 | |

### Table 7: Paired Comparison t-test Soybeans

| MONTH | MEAN DIFFERENCE | t-VALUE | Prob>|t| | Cr |
|-------|----------------|---------|--------|-----|
| REGION: | | | | |
| JUL | 0.09 | 0.9 | 0.38 | 65% |
| AUG | 0.02 | 0.2 | 0.85 | |
| YEARS: | | | | |
| 1988 | AUG | -0.61 | -2.4 | 0.03 | 65% |
| STATES: | | | | |
| IA | JUL | 0.18 | 1.8 | 0.09 | 86% |
| NE | AUG | 0.90 | 4.9 | 0.00 | 74% |
Table 8: Paired Comparison t-test Wheat

| MONTH | MEAN DIFFERENCE | t-VALUE | Prob>|t| | Cr |
|-------|----------------|---------|--------|------|
| AUG+MAR | 0.39 | 6.1 | 0.00 | 79% |

YEARS:

| YEAR | MONTH | MEAN DIFFERENCE | t-VALUE | Prob>|t| | Cr |
|------|-------|----------------|---------|--------|------|
| 1985 | AUG+MAR | 0.61 | 2.4 | 0.03 | 79% |
| 1987 | 0.62 | 2.5 | 0.02 |
| 1991 | 0.42 | 2.2 | 0.05 |

STATES:

| STATE | MONTH | MEAN DIFFERENCE | t-VALUE | Prob>|t| | Cr |
|-------|-------|----------------|---------|--------|------|
| ID | AUG+MAR | 0.55 | 5.8 | 0.00 | 73% |
| IL | 0.77 | 2.3 | 0.06 | 38% |
| OR | 0.76 | 3.4 | 0.01 | 54% |
| SD | 0.54 | 3.4 | 0.01 | 66% |
| WA | 1.58 | 6.8 | 0.00 | 100% |

VIII. APPENDIX B

The following procedures are used to aggregate the precipitation and survey data to regional totals.

Precipitation Data

Precipitation variables used in the models represent total precipitation for a particular month at the regional level. The variable is constructed as follows:

\[ p_t = \frac{\sum_{s=1}^{S} A_{ts} R_{ts}}{\sum_{s=1}^{S} A_{ts}} \]  

(8.1)

where

- \( p_t \) = estimated monthly accumulated precipitation for the region for year t,
- \( S \) = the number of States covered,
- \( A_{ts} \) = the acres for harvest for year t, State s, and
- \( R_{ts} \) = the estimated monthly accumulated precipitation for year t, State s, where
\[ R_{tsd} = \frac{\sum_{d=1}^{D_s} A_{tsd} E_{tsd}}{\sum_{d=1}^{D_s} A_{tsd}} \]  

where:

- \( A_{tsd} \) is the acres for harvest for year \( t \), State \( s \), district \( d \), and
- \( D_s \) is the number of districts per State \( s \).
- \( E_{tsd} \) is the average monthly station accumulated precipitation for year \( t \), State \( s \), district \( d \), where

\[ E_{tsd} = \frac{1}{W_{tsd}} \sum_{w=1}^{W_{tsd}} U_{tsdw} \]  

where:

- \( W_{tsd} \) is the number of weather stations for year \( t \), State \( s \), district \( d \), and
- \( U_{tsdw} \) is the accumulated monthly precipitation for year \( t \), State \( s \), district \( d \), weather station \( w \).

**Survey Data**

The construction of the independent variables for the regional regression models for both soybeans and corn is discussed by Birkett (1990, 1993). For soybeans for the month of August, the independent variable \((Z_t)\) is the estimated number of lateral branches per eighteen square feet. For September, the independent variable is the estimated number pods with beans per eighteen square feet. These regional level estimates for soybeans are constructed as follows:

\[ Z_t = \frac{\sum_{s=1}^{S} A_{ts} F_{ts}}{\sum_{s=1}^{S} A_{ts}} \]  

where:

- \( A_{ts} \) is the acres for harvest for year \( t \), State \( s \), and
- \( F_{ts} \) is the mean number of lateral branches per 18 sq. feet for year \( t \), State \( s \), for August or
- mean estimated pods per 18 sq. feet for year \( t \), State \( s \), for September where

\[ F_{ts} = \frac{1}{m_{ts}} \sum_{j=1}^{m_{ts}} B_{tsj} \]  

where:

- \( m_{ts} \) is the number of samples in \( J_s \) in year \( t \), State \( s \),
- \( J_s \) is the subset of samples classified in maturity categories 2-6 (or 1-6 in southern States), year \( t \), State \( s \),
- \( B_{tsj} \) is plants per 18 square feet for year \( t \), State \( s \), sample \( j \),
- \( L_{tsj} \) is lateral branches per plant year \( t \), State \( s \), sample \( j \), for August, and
- estimated pods with beans per plant per 18 sq. feet, year \( t \), State \( s \), sample \( j \), for September.
Corn independent variables \((Z_t)\) are more complex as they are a function of both plant counts and average kernel row length per square foot. \(C_u\) is substituted for \(F_u\) in equation (8.4). In August, it is calculated as:

\[
C_{ts} = \frac{1}{m_{ts}} \sum_{j=1}^{m_{ts}} (U_{tsj} + V_{tsj}) \bar{K}_{tsj}, \tag{8.6}
\]

where

\(C_u = \) a function of the number of stalks with ears, the number of ears with kernels, and the average kernel row length per square foot,

\(U_{uj} = \) number of stalks with ears, per sq. ft., year \(t\), State \(s\), sample \(j\),

\(V_{uj} = \) number of ears with kernels, per sq. ft., year \(t\), State \(s\), sample \(j\), and

\(\bar{K}_{uj} = \) the average kernel row length per ear, year \(t\), State \(s\), sample \(j\).

In September, \(C_u\) is calculated as:

\[
C_{ts} = \frac{1}{m_{ts}} \sum_{j=1}^{m_{ts}} V_{tsj} \bar{K}_{tsj}, \tag{8.7}
\]

For both forecasts, data are used from the subset of samples in maturity categories 3-6 for year \(t\), State \(s\).

The winter wheat independent variable \((Z_t)\) for May is a function of number of stalks with emerged or late boot heads and weight per head per square foot. Here, \(H_u\) is substituted for \(F_u\) in equation (8.4). It is calculated as:

\[
H_{ts} = \frac{1}{m_{ts}} \sum_{j=1}^{m_{ts}} Q_{tsj} G_{tsj}, \tag{8.8}
\]

where

\(H_u = \) a function of number of stalks with emerged or late boot heads times weight per head per sq. ft. year \(t\), State \(s\),

\(Q_{uj} = \) number of stalks with emerged or late boot heads per sq. ft., year \(t\), State \(s\), sample \(j\), and

\(G_{uj} = \) weight per head year \(t\), State \(s\), sample \(j\).

**Yield Data**

The regional yield values included in this study were calculated as follows:

\[
Y_t = \frac{\sum_{s=1}^{S} A_{ts} Y_{ts}}{\sum_{s=1}^{S} A_{ts}}, \tag{8.9}
\]

where

\(Y_t = \) final regional yield for year \(t\), and

\(Y_u = \) NASS State yield for year \(t\), State \(s\).