An Iterative Regression Model for Estimating Soybean Yields from Environmental Data

ANDRES C. RAVELO AND WAYNE L. DECKER

Department of Atmospheric Science, University of Missouri, Columbia 65211

(Manuscript received 9 April 1980, in final form 27 July 1981)

ABSTRACT

A model was developed for using weather data to estimate the yields of soybeans for varieties adapted to the central United States. The model utilized an iterative regression analysis for relating soybean yields to environmental variables. This technique evaluated the simple and interacting contributions to soybean yield of environmental variables in terms of a time scale related to soybean development (biometeorological time). The environmental variables tested were daily climatological data (rainfall and maximum and minimum air temperatures), derived agrometeorological variables (actual and potential evapotranspiration) and a soil moisture index. The maximum air temperature, potential evapotranspiration and soil moisture index accounted for more of the variability in soybean yields (coefficient of determination of 0.75) than other combinations of the tested variables. For verification of the model, a sample of 20 yields were withheld from the iterative regression analysis and comparisons were made between the yields simulated from the regression equations and the observed yields. The mean difference of 0.98 q ha\(^{-1}\) between observed and estimated yields for the twenty cases did not differ from zero by a statistically significant amount. The standard error of estimates was 4.79 q ha\(^{-1}\). Although this precision provides estimates of field yields which may be used for many practical purposes, the low correlation between the observed and estimated yields for the test cases indicates the need for caution in using this type of analysis.

1. Introduction

Among the major components of the national and international economy, agricultural production is subjected to wide year-to-year fluctuations. These variations are caused largely by fluctuations in weather and climate. Excellent statements on the impacts of variations in climate on food supply have been published by Landsberg (1974), Newman and Rickett (1974) and McQuigg et al. (1973). Reliable crop forecasts may allow for economic and social adjustments to both shortages and surpluses of agricultural commodities (Decker, 1976).

This study was designed to develop a yield model for soybeans [Glycine max (L) Merrill] from a regression analysis which estimates 1) the final yield of soybeans for field size areas using simple weather and derived environmental variables as predictors, and 2) the contribution of the daily values of these predictors to potential yield of soybeans.

2. Data

a. Independent variables

BIOMETEOROLOGICAL TIME SCALE

Calendar time has been used by Runge and Odell (1960) and Thompson (1963, 1970) to study the effects of weather conditions on soybean yield. In this study, a biometeorological time scale (BTS) was used. The BTS, which is defined as the rate of development toward maturity, was first suggested by Robertson (1968) for wheat and it was later adapted to soybeans by Major et al. (1975). The BTS, designated in this study as \( t \), for soybeans recognizes six phenological stages: planting, emergence, beginning of flowering, beginning of pod filling, end of flowering and maturity. These are denoted successively, as \( t = 0, 1, 2, 3, 4 \) and 5. The planting dates were reported for each year at the test sites (USDA, 1957–76). The dates for the stages following planting were computed from the maximum and minimum air temperatures and daylengths with the technique developed by Major et al. (1975). For the analysis reported here a separate BTS value (i.e., value of \( t \)) was computed for each day of the growing season. The planting date of each growing season was assigned a value of \( t = 0 \), the date of emergence \( t = 1.0 \), the date of the beginning of flowering \( t = 2.0 \), etc. The decimal value of the BTS for each day was computed by dividing the number of days which a given date was into the phenological period by the total length in days of the phenological period. For example, if on a particular year soybeans were planted on 1 June and emerged on 7 June; 1 June would have a value of \( t = 0 \), on 2 June \( t = \% = 0.17 \), on 3 June \( t = \% = 0.33 \), etc. If on the same year
the soybeans began to flower on 3 July (27th day after emergence), the \( t \) value on 8 June would be 
\[ 1 + \frac{52}{258} = 1.04, \] on 30 June the \( t \) value would be 
\[ 1 + \frac{2}{328} = 1.88. \] The mean dates of all periods of the 
BTS are shown in Table 1 for the locations used in 
this study.

b. Weather-related variables

Daily precipitation and maximum (TMAX) and 
minimum (TMIN) air temperatures from the weather 
station nearest to the experimental plots (Table 1) 
were tested in the development of the models for 
predicting soybean yields. For a model with physio-
logical significance, plant water stress is a logical 
predictor of grain yields; but direct estimates of plant 
stress are not generally available. Potential water 
stress may be determined from estimates of evaporative 
demand by the atmosphere and the soil moisture 
levels. To develop an index of evaporative demand, 
estimates of potential evapotranspiration (PE) were 
required; and, since radiation (solar and net), sun-
shine duration, atmospheric humidity and pan 
evaporation were not observed at most of the loca-
tions with the soybean yield data, conventional 
methods for estimating potential evapotranspiration 
were not possible. Daily estimates of PE were 
obtained from the multiple-regression technique sug-
gested by Baier and Robertson (1965) and Baier 
(1967):

\[
PE = -5.6715 + 0.136(TMAX) 
+ 0.1065(TRANGE) + 0.0055(Q_0),
\]

where TMAX is the daily maximum temperature 
(°C), TRANGE the range of daily temperatures (°C), 
\( Q_0 \) the daily solar energy at the top of the atmosphere 
(cal cm\(^{-2}\) day\(^{-1}\)), and PE the potential evapotrans-
piration (mm day\(^{-1}\)) as estimated by the method of 
Penman (1956).

In Eq. (1) the temperature variables provided 
estimates of atmospheric stress while \( Q_0 \) added a 
seasonal adjustment in the estimated PE associated 
with the change in the potential for solar energy 
through the growing season. To estimate the regres-
sion statistics of Eq. (1) it was necessary to conduct 
an evaluation using environmental data required for 
the Penman estimate of PE (temperature, humidity 
and net radiation). Such data were available from a 
field station operated by the Missouri Agricultural 
Experiment Station in central Missouri. Using the 
Penman estimates of PE as the dependent variable 
and TMAX, TRANGE and \( Q_0 \) as independent 
variables, the regression statistics of Eq. (1) were 
calculated. Approximately 84% of the variation in 
the daily estimates of PE was explained by the 
equation.

c. Derived variables relating to soil characteristics

The actual evapotranspiration (AE) depends on 
the evaporative demand by the atmosphere (PE), 
the adequacy of soil moisture and the stage of 
development of the crop. To determine these relations-
ships for soybeans, soil moisture measurements 
under a soybean crop, which were reported by the 
USDA (1977), were examined for the years 1974, 
1975 and 1976. The measurements of soil moisture 
were made weekly by 15 cm increments of soil 
depth using gravimetric determinations for the upper 
46 cm and a neutron probe for depths from 46 to 160 
cm. Coefficients for the adjustment of PE to esti-
mates of ET for each crop growth stage and the 
contribution of each soil layer to the evaporative 
demand of the crop canopy were determined by 
iterative comparisons between the computed and 
measured soil moisture. The iteration was accom-
plished by repeatedly changing the fractional con-
tribution of each soil layer to the total evapotrans-
piration. These iterations were continued until 
the best coincidence to the withdrawal of water by 
evapotranspiration from a layer over the 3-year 
period was observed. This technique was applied

<table>
<thead>
<tr>
<th>Location</th>
<th>North (latitude)</th>
<th>West (longitude)</th>
<th>Number of (years)</th>
<th>Planting</th>
<th>Emergence*</th>
<th>Flowering*</th>
<th>Pod filling*</th>
<th>End flowering*</th>
<th>Maturity*</th>
<th>Mean grain yield (q ha(^{-1}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Portageville, MO</td>
<td>36°25'</td>
<td>89°42'</td>
<td>6</td>
<td>5/15</td>
<td>5/23</td>
<td>6/24</td>
<td>7/19</td>
<td>8/13</td>
<td>9/2</td>
<td>23.5</td>
</tr>
<tr>
<td>Mt. Vernon, MO</td>
<td>37°04'</td>
<td>93°53'</td>
<td>4</td>
<td>5/15</td>
<td>5/26</td>
<td>6/22</td>
<td>7/16</td>
<td>8/6</td>
<td>9/1</td>
<td>28.6</td>
</tr>
<tr>
<td>Urbana, IL</td>
<td>40°06'</td>
<td>88°14'</td>
<td>7</td>
<td>5/19</td>
<td>5/30</td>
<td>7/5</td>
<td>8/1</td>
<td>8/12</td>
<td>9/6</td>
<td>31/7</td>
</tr>
<tr>
<td>Carbondale, IL</td>
<td>37°44'</td>
<td>89°12'</td>
<td>13</td>
<td>5/30</td>
<td>6/7</td>
<td>7/8</td>
<td>8/2</td>
<td>8/18</td>
<td>9/10</td>
<td>26.2</td>
</tr>
<tr>
<td>Worthington, IN</td>
<td>39°02'</td>
<td>86°57'</td>
<td>16</td>
<td>5/27</td>
<td>6/5</td>
<td>7/10</td>
<td>8/5</td>
<td>8/16</td>
<td>9/10</td>
<td>30.7</td>
</tr>
<tr>
<td>Evansville, IN</td>
<td>38°03'</td>
<td>87°32'</td>
<td>9</td>
<td>5/25</td>
<td>6/3</td>
<td>7/5</td>
<td>8/1</td>
<td>8/16</td>
<td>9/7</td>
<td>31.2</td>
</tr>
<tr>
<td>Columbus, KS</td>
<td>37°15'</td>
<td>94°52'</td>
<td>10</td>
<td>6/5</td>
<td>6/14</td>
<td>7/13</td>
<td>8/6</td>
<td>8/27</td>
<td>9/15</td>
<td>19.8</td>
</tr>
<tr>
<td>Powhatan, KS</td>
<td>39°40'</td>
<td>95°31'</td>
<td>10</td>
<td>5/30</td>
<td>6/8</td>
<td>7/4</td>
<td>8/4</td>
<td>8/21</td>
<td>9/12</td>
<td>25.5</td>
</tr>
</tbody>
</table>

* As estimated by the method of Major et al. (1975).
TABLE 2. Crop coefficients to adjust PE to AE for soybean growth stages and soil depths. Thickness of each soil zone and soil water retention characteristics (MW: maximum available water) are shown for the versatile budget at Columbia, Missouri.

<table>
<thead>
<tr>
<th>Plant growth stages</th>
<th>Soil zones</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>BTS*</td>
</tr>
<tr>
<td>Planted to</td>
<td></td>
</tr>
<tr>
<td>emergence</td>
<td>0.1-1.0</td>
</tr>
<tr>
<td>Emergence to</td>
<td></td>
</tr>
<tr>
<td>first flower</td>
<td>1.0-2.0</td>
</tr>
<tr>
<td>First flower to</td>
<td></td>
</tr>
<tr>
<td>pod set</td>
<td>2.0-3.0</td>
</tr>
<tr>
<td>First pod to</td>
<td></td>
</tr>
<tr>
<td>end flower</td>
<td>3.0-4.0</td>
</tr>
<tr>
<td>End flower to</td>
<td></td>
</tr>
<tr>
<td>maturity</td>
<td>4.0-5.0</td>
</tr>
<tr>
<td>Soil zone thickness</td>
<td></td>
</tr>
<tr>
<td>(cm)</td>
<td>4.8</td>
</tr>
<tr>
<td>Maximum soil</td>
<td></td>
</tr>
<tr>
<td>water (MW) (cm)</td>
<td>1.10</td>
</tr>
<tr>
<td>Percent of the</td>
<td></td>
</tr>
<tr>
<td>total MW in</td>
<td></td>
</tr>
<tr>
<td>each zone</td>
<td>5.00</td>
</tr>
</tbody>
</table>

* Refers to biological time scale.

separately to each period of the biological time scale. The final coefficients are shown in Table 2. According to Table 2, for the period from planting to emergence of the plants, AE is 60% of PE when soil moisture is nonlimiting, and all of this water is removed from the upper 61.5 cm of soil. On the other hand, during the period from first pod set until the end of flowering, AE is 95% of PE and water is removed from the soil to a depth of 152 cm.

The actual evapotranspiration was further adjusted according to availability of soil moisture by adaptations of methods suggested by Baier and Robertson on (1965) and Baier et al. (1972). The available soil moisture (AW, for the ith soil layer) is defined as the difference between estimated soil moisture and the soil moisture content at the permanent wilting point for the soil. AW has a maximum value (MW), which also is a characteristic of the soil type, and is defined as the difference in the soil water content of a fully watered soil at equilibrium with gravitational forces (i.e., a soil at field capacity) and the water content at the permanent wilting point for the soil. The soil characteristics were obtained for the soils used in this investigation from information supplied by USDA (Current Series) and by the USDA Uniform Soybean Test (1957–76).

Using the values of evapotranspiration, a derived independent variable for the prediction model was defined as the ratio between the actual and potential evapotranspiration (AE/PE). Further, the daily simulated soil moisture values provided an additional independent variable through a soil moisture index (SMI) suggested by Ravelo and Decker (1979). The SMI is the ratio of the actual plant available soil moisture (AW) to the maximum plant available soil moisture (MW). Two soil moisture indices were tested. The first index involved the moisture content of all six soil layers, surface to 152 cm (SM16). Since the root systems of mature soybeans have been reported by Mitchell and Russell (1971) and Raper and Barber (1970) to be primarily in the upper 10–20 cm of soil, a soil moisture index for the upper three layers of soil, surface–24 cm (SMI3), was tested.

d. Dependent variable

Yield observations from the Uniform Soybean Tests for the northern states (USDA, 1957–76) were selected for model development and testing. These observations were randomly divided into one group of 80 to be used in the regression development and a second group of 20 for testing the resulting regression equation.

The mean grain yields reported for all varieties of maturity Group IV, which is the recommended maturity group for the locations in the central United States, were used as the dependent variable in this study. The locations of the soybean test sites are shown in Table 1.

3. The iterative regression model

a. Description of model

The proposed model for relating soybean yields to weather data basically involves the equation

$$Y = \sum_{i=5}^{\infty} V_i V_2 V_3,$$

where $Y$ is the seasonal grain yield in quintals (100 g) per hectare (q ha$^{-1}$), $V_i$ is the summation of daily $Y$ values over biometeorological times, $0 < t < 5$ for soybeans, and $V_1$, $V_2$, $V_3$ are the functions of the selected "independent variables." Each $V$ function is of the general form

$$V_{ij} = (a_{i1} t^j + a_{i2} t^{j+1} + a_{i3} t^{j+2} + a_{i4} t^{j+3}) + (a_{i5} t^j + a_{i6} t^{j+1} + a_{i7} t^{j+2} + a_{i8} t^{j+3}) X_1 + \cdots + (a_{i8} t^j + a_{i9} t^{j+1} + a_{i10} t^{j+2}) X_n,$$

for $i = 1, 2$ and $3, j = 1, 2, \ldots, n$ days, where $a_{i1}, a_{i2}, \ldots, a_{i8}$ are coefficients which are evaluated for each $Y$ by an iterative regression technique (Baier, 1973 and Ravelo, 1978) and $X_j$ represents the environmental variables used for the $j$th day.

In this model, soybean response to each environmental variable can be either linear or quadratic. The biometeorological time is an indicator of the quantitative response of the crop during its life cycle. Therefore, it is assumed a gradual change of the soybean response to the environmental variables occurs over the growing season and the daily
weighting of each variable can be adequately fitted by a fourth-power polynomial as a function of the biometeorological time [Eq. (3)].

b. Application of the model

The coefficients in Eq. (3) used to estimate \( V_1, V_2, V_3 \) were evaluated and tested by an iterative multiple-regression technique adapted from Baier (1973). The intercept with the \( Y \) axis was suppressed because the estimated contribution to the yield accumulation at planting time \( (t = 0) \) must be zero. Following a suggestion by Robertson (1968) the estimated coefficients for Eq. (3) were normalized by adjusting the magnitudes so that the sum of these coefficients for each \( V_i \) equaled unity. This procedure is described by Ravelo (1978). This iterative procedure is continued until the \( R^2 \), the coefficient of determination (CD), and the standard errors of estimate (SEE) do not change appreciably with another iteration. In this study, the CD’s and the SEE’s stabilized after about 42 iterations. However, Ravelo (1978) found that for practical purposes the iterative regression procedure could have been discontinued after ten iterations.

Soybean yields were related to the various combinations of agrometeorological variables and Table 3 gives the coefficients of determination for some of these combinations. Generally, those combinations of variables that include SMI3 produced higher coefficients of determination than combinations using SMI6 as a variable. Models using minimum air temperatures were associated with lower correlations than those with maximum temperatures.

The highest coefficient of determination was associated with the model using maximum air temperature, potential evapotranspiration and the soil moisture index. Variations in the daily input of these variables account for 75% of the total variability in seasonal grain yield at the 10 locations. The close relation between observed and estimated yields for the 80 plantings used in the model development are shown in Fig. 1. The correspondence between predicted and observed is reasonably close over the entire range.

<table>
<thead>
<tr>
<th>Variable combination</th>
<th>Iteration no. 42</th>
</tr>
</thead>
<tbody>
<tr>
<td>TMAX-PE-SMI6</td>
<td>0.62</td>
</tr>
<tr>
<td>TMAX-PE-SMI3</td>
<td>0.75</td>
</tr>
<tr>
<td>TMAX-AE-SMI3</td>
<td>0.71</td>
</tr>
<tr>
<td>TMAX-AE/PE-SMI3</td>
<td>0.74</td>
</tr>
<tr>
<td>TMIN-PE-SMI3</td>
<td>0.63</td>
</tr>
<tr>
<td>TMIN-MAX-PE</td>
<td>0.63</td>
</tr>
<tr>
<td>TMIN-MAX-AE/PE</td>
<td>0.55</td>
</tr>
</tbody>
</table>

![Figure 1](image1.png)

**Fig. 1.** Scatter diagram of observed versus estimated soybean yields (q/ha) for 80 plantings used to develop regression estimates using daily maximum temperature, potential evapotranspiration, and soil moisture index (SMI3) as predictors. The line is the 1:1 correspondence between predicted and observed.

The daily contribution to yield of agrometeorological variables can be evaluated by using the normalized coefficients from the last iteration. To illustrate this, Fig. 2 shows the accumulated “mean” contribution (shown by the dashed line) to the average yield of the daily predictions obtained from the iterative regression for the 80 station-years (1957–76) using the three variables providing the highest correlation. This portion of Fig. 2 indicates the importance of the environmental conditions during the period extending from just prior to flowering through the beginning of the pod-filling period. It appears from this analysis that the weather before or after this

![Figure 2](image2.png)

**Fig. 2.** The accumulated yield contributions for each day. The dashed line shows the mean accumulation while the solid line is for 1962 at Columbia, Missouri. The predictors used were daily maximum temperature, potential evapotranspiration, and soil moisture index (SMI3). 1962 Final observed yield; estimated yield.
transpiration rates (see Fig. 3) caused the reduction in crop prospects. After 1 September, the yield prospects recovered slightly to 22.9 q ha\(^{-1}\) at the maturity date of 13 September. This observed effect of the sequence weather indicated the necessity of considering the entire growing season when simulating soybean yields.

c. Model verification

Twenty station-years, from 10 locations and for the period 1957–76, were selected at random for use as an independent data set for testing the model. The estimated yields using the prediction model were compared with the observed yields from the soybean uniform test in Fig. 4. Although the coefficient of determination was only 0.19, most data points fall relatively close to the 1:1 line. The predicted yields for three station-years (Columbia, 1963; Mt. Vernon, 1968; Urbana, 1966) were considerably lower than the actual yields. Occurrence of a severe bacterial pustule and bacterial blight at Urbana in 1966 and a green stinkbugs infestation at Mt. Vernon in 1968 could explain the high predicted yields for these cases. There is no apparent explanation for Columbia in 1963.

A review of Fig. 4 shows that the selection of test data by a completely random process did not provide test yields over the full range of yields. In the test data the range of yields was from 19.8 to 34.9 q ha\(^{-1}\) while in the sample used to develop the regression equations the range was from 12.4 to 37.6 q ha\(^{-1}\). The low correlation for the data in Fig. 4 may have been because of the restricted range of the sample. To test the impact of the sampling technique, the data from the 100 samples were again divided into 80 yields for computing the regression statistics and 20 yields for testing, but the test cases were obtained by selecting every fifth value from an array of the 100 yields. The complete regression analysis was repeated and the results are summarized in Table 4. The new regression analysis showed that increasing the range of values for the

---

**Table 4. Comparison of regression statistics from two methods for sampling test cases from the 100 plantings.**

<table>
<thead>
<tr>
<th>Regression results for sample of 80 plantings</th>
<th>Test case using sample of 20 plantings</th>
</tr>
</thead>
<tbody>
<tr>
<td>First analysis</td>
<td>Second analysis</td>
</tr>
<tr>
<td>First analysis</td>
<td>Second analysis</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>First analysis</th>
<th>Second analysis</th>
<th>First analysis</th>
<th>Second analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard error of estimate (q ha(^{-1}))</td>
<td>2.8</td>
<td>3.0</td>
<td>4.8</td>
<td>3.8</td>
</tr>
<tr>
<td>Coefficient of determination</td>
<td>0.75</td>
<td>0.70</td>
<td>0.19</td>
<td>0.35</td>
</tr>
<tr>
<td>Mean yield (q ha(^{-1}))</td>
<td>26.9</td>
<td>26.6</td>
<td>26.9</td>
<td>26.7</td>
</tr>
<tr>
<td>Standard deviation of yields (q ha(^{-1}))</td>
<td>5.6</td>
<td>5.4</td>
<td>5.0</td>
<td>4.9</td>
</tr>
</tbody>
</table>

---

**Fig. 4. Scatter diagram of observed versus estimated soybean yields (q/ha) for the 20 plantings used for testing the regression equation.**
test cases greatly improved both the coefficient of determination and the standard error of estimate. The CD for the second test was nearly double that of the first test.

4. Conclusions

The model produced estimates of the seasonal grain yields which were compared with the observed yield. The standard error of estimate for the primary analysis reported was 2.8 q ha⁻¹ as compared with 4.8 q ha⁻¹ for the control data. The difference between the mean of the observed and estimated yields was statistically non-significant. According to the standard error of estimate, using the first test case, one would expect to estimate yields for other locations to within 3.2 q ha⁻¹ about half the time. This precision would provide estimates of field yields for many practical uses.

The amount of the variation in yields explained by the regression equation for the test years is quite low (19% for the first test and 33% for the second test). The low correlation coefficient for the independent test reduces the confidence in yield estimates using this technique. To directly utilize mathematical models of this type one must be able to utilize the relationships for estimating yields for different years and locations. The poor performance of the iterative regression technique with independent data may be a characteristic of many regression analyses.

It appears clear from Fig. 2 that the daily weather for the period extending about two weeks just prior to the onset of flowering and extending until about a week after the initiation of pod filling is critical to soybean production. The existence of a critical period at this time is probably because of the interaction between soybean response and the climate in mid-America. From Table 1 it is apparent that the critical period corresponds closely to the month of July when high temperatures and drought are frequent in the Central United States.

To improve the precision of the model in explaining the variability of soybean yields, some possible modifications are suggested. The model considers the effect of only three selected agrometeorological variables but it can be easily modified to account for up to five variables. Technology trend, such as the amount of fertilizer applications, could be included in the model. The crop stage predictions might be improved by including soil moisture and solar radiation as determinant factors of plant phenological events.

The soybean-weather analysis model as proposed here requires more testing and verification. However, results are encouraging for utilizing this model to predict soybean yield as the season progresses. This model can be adapted to estimating final yields at any BTS interval. For this purpose, weather data for the current year provide the current accumulative effects on the final yield, and a projection of normal daily weather events from the selected BTS interval to the end of the season will provide the final yield estimate.

REFERENCES


——, Current Series: Soil Survey Interpretation Form 5. Soil Cons. Ser., Midwest Regional Technical Center, Lincoln, NE.

——, 1977: Soil moisture observations at Fleetwood Farm, Columbia, Missouri. Watershed Research Center, 3 pp.