ARTIFICIAL NEURAL NETWORK MODEL AS A DATA ANALYSIS TOOL IN PRECISION FARMING

A. Irmak, J. W. Jones, W. D. Batchelor, S. Irmak, K. J. Boote, J. O. Paz

ABSTRACT. *Spatial variation in landscape and soil properties combined with temporal variations in weather can result in yield patterns that change annually within a field. The complexity of interactions between a number of yield-limiting factors makes it difficult to accurately attribute yield losses to conditions that occur within a field. In this research, a back-propagation neural network (BPNN) model was developed to predict the spatial distribution of soybean yields and to understand the causes of yield variability. First, we developed a BPNN model by relating soybean yield to topography, soil, weather, and site factors and evaluated model predictions for the same field for independent years. We also explored the potential use of BPNN for predicting yields in independent fields. Finally, we evaluated the ability of the BPNN to attribute yield losses due to soybean cyst nematodes (SCN), soil pH, and weeds. A total of 14 input datasets with combinations of four controlling factors (topographic, soil fertility, weather, and site) were used. For each objective, data from fields in Iowa were used for training the BPNN, while a portion of the data was withheld to verify the accuracy of yield predictions. All BPNN models had fully connected feed-forward architecture with a back-propagation weight adjustment algorithm. When tested for a particular field, the BPNN captured the major patterns of yield variability in independent years; the root mean square error of prediction (RMSEP) was 14.2% of actual yield. When the BPNN was trained with inputs from five fields, the RMSEP at test sites was 11.2% of actual yield. When the BPNN was used to attribute yield losses to soil pH, SCN, and weed populations, standard errors were 92, 262, and 171 kg ha^{-1}, respectively. The technique showed that the BPNN could predict spatial yield variability with an RMSEP of about 14%.*

Keywords. *Artificial neural network, Precision farming, Spatial yield variability.*

Spatial variation in landscape and soil properties combined with temporal variations in weather can result in crop yield patterns that vary annually within a field. A number of yield-limiting factors create this temporal and spatial variability (i.e., water stress, diseases, weeds, soil fertility). In the past several years, researchers have used empirical or process-based tools to identify various yield-limiting factors within fields and analyze causes of yield variability. Crop growth models like the CROPGRO-Soybean (Hoogenboom et al., 1994) and CERES-Maize (Jones and Kiniry, 1986) have been used to identify the magnitude of yield loss attributed to yield-limiting factors such as water stress, diseases, weeds, and soil pH (Paz et al., 1998, 2001a; Irmak et al., 2002a). While this approach has considerable promise, its success depends on the accuracy of relationships that are incorporated into the crop model and their transferability to other years and locations (Irmak et al., 2002a).

Another approach to estimate yield losses due to yield-limiting factors is artificial neural networks (ANNs). Unlike analytical approaches, ANNs require no explicit mathematical equation and no limiting assumptions of normality or linearity (MathWorks, 2005). The advantage of an ANN over more traditional physiology-based crop models is that most of the intense computations take place during the training process. Once the ANN is trained for a particular system, its operation is relatively fast and unknown input patterns can be rapidly identified in a real-time environment (Keller et al., 1994). ANNs have been used in a wide range of data processing applications such as image recognition (Burks et al., 1999; Gliever and Slaughter, 2001; Noh et al., 2004), land use change/classification (Carpenter et al., 1997; Gopal et al., 1999; Carpenter et al., 1999a, 1999b; Shock et al., 2002), land drainage engineering (Yang et al., 1996, 1997), and crop evapotranspiration calculation (Odhiambo et al., 2001a, 2001b) as well as to predict yield for a new set of input conditions (Drummond et al.; 1995; Shearer et al., 1999; Wilkerson et al., 1999; Liu et al., 2001), support the use of mechanistic simulation tools by providing initial condition

Submitted for review in August 2004 as manuscript number BE 5490; approved for publication by the Biological Engineering Division of ASABE in October 2006.

A contribution of University of Nebraska Agricultural Research Division, Lincoln, NE 65583, Journal Series No. 14616. Florida Agricultural Experiment Station Journal Series No. R-08851.

The authors are **Ayse Irmak, ASABE Member Engineer,** Research Assistant Professor, Department of Biological Systems Engineering, University of Nebraska-Lincoln, Lincoln, Nebraska; **James W. Jones, ASABE Fellow,** Distinguished Professor, Department of Agricultural and Biological Engineering, University of Florida, Gainesville, Florida; **William D. Batchelor, ASABE Member Engineer,** Professor and Head, Department of Agricultural and Biological Engineering, Mississippi State University, Mississippi State, Mississippi; **Suat Irmak, ASABE Member Engineer,** Assistant Professor, Department of Biological Systems Engineering, University of Nebraska-Lincoln, Lincoln, Nebraska; **Kenneth J. Boote,** Professor, Department of Agronomy, University of Florida, Gainesville, Florida; and **Joel O. Paz, ASABE Member Engineer,** Public Service Assistant, Department of Biological and Agricultural Engineering, University of Georgia, Griffin, Georgia. **Corresponding author:** Ayse Irmak, Department of Biological Systems Engineering, University of Nebraska-Lincoln, 253 L. W. Chase Hall, Lincoln, NE 68583-0726; phone: 402-472-5351; fax: 402-472-6338; e-mail: airmak2@unl.edu.

  ISSN 0001–2351

values or site-specific parameters (Braga, 2000), and guide parameter estimation in agricultural machinery models (Pinto et al., 1999).

A well-trained ANN model should include key factors that quantify the inherent variability associated with crop production to capture the essence of meaningful relationships. In site-specific management (SSM), the inputs may include soil, landscape, weather, and crop variables. Drummond et al. (1995) used a feed-forward, back-propagation neural network (BPNN) design for predicting corn and soybean yield. Their BPNN model included inputs such as soil phosphorus, magnesium, potassium, pH, organic matter, and topsoil depth. The BPNN was found useful as an aid in understanding yield variability, although Drummond's network model needed further improvements for increasing accuracy, such as incorporating weather information. Shearer et al. (1999) developed an ANN to determine if fertility, elevation, electrical conductivity, or satellite image features may be used in conjunction with standard BPNN topologies to predict spatial variability in corn grain yield. The BPNN model showed promise in predicting spatial yield variability but was not able to predict high yields, thereby reducing the predictive power and utility of the ANN. Braga (2000) used a BPNN to predict the spatial patterns of corn yield using topographic and agronomic variables as well as seasonal rainfall as input. His work resulted in a reasonable level of predictability of the spatially variable grain yield, but he did not have enough data to check the accuracy of predictions in other years or in other fields. Liu et al. (2001) included soil factors, management factors, and monthly rainfall as inputs in their BPNN. Their results showed that: (1) the BPNN was able to capture the expected interactions between rainfall and the amount of applied nitrogen fertilizer, (2) corn yields could be predicted with 80% accuracy using the BPNN, and (3) calculated yield trends were realistic, i.e., yields showed the expected increase, flattened, and then decreased as various input factors were increased.

In order to use inputs efficiently and optimize profitability, growers need to know the tradeoffs between the costs of the input and its application versus yield and income losses associated with failure to control those factors. The main question addressed in this research was whether an ANN model, trained to fit spatially distributed yield data, could then be used to quantify the losses in yield associated with different yield-limiting factors that occur in a field. Another question was whether an ANN trained for one field could be used for the same field in different years or in independent fields. For this study, data on spatial variability of soybean yields in Iowa were used to address these questions. Specific objectives were to:

- Develop an ANN model relating soybean yield to topography, soil fertility, weather, and site factors, and evaluate model predictions for the same field for independent years.
- Evaluate the uncertainty associated with using an ANN to predict yield in independent fields during the same year.
- Evaluate the ability of the ANN approach to attribute yield losses due to soybean cyst nematodes (SCN), soil pH, and weed stress factors.

Materials and Methods

Field Description

Data from six fields in three counties in Iowa were used to develop and evaluate the ANN models. These fields were used to study the effects of soil, water, and pest interactions on yield variability (Paz et al., 1998, 2001a, 2001b; Irmak et al., 2001, 2002a, 2002b). The McGarvey (41.93° N, 94.07° W) and Heck fields (41.94° N, 94.08° W) are located near Perry City in Boone County. The Keiper Ray 100 (42.01° N, 91.80° W) and Ray 80 (42.00° N, 91.80° W) fields are near Cedar Rapids in Linn County and are adjacent to each other. The Kusel North (41.90° N, 95.00° W) and Kusel South (41.89° N, 95.00° W) fields are in the western part of Iowa in Carrol County (fig. 1). All fields used a corn (*Zea mays* L.) – soybean (*Glycine max* L.) rotation, and all field operations were conducted by the cooperating farmers.

Input Layers of Information

In this research, we assumed that soybean yields vary within and across fields due to four main types of environmental factors and their interactions: weather, soil fertility,

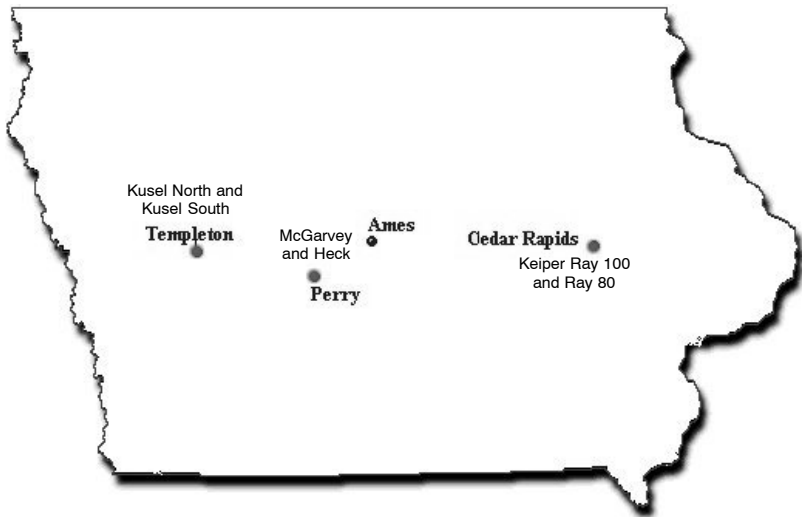

Figure 1. Location of six fields in Iowa from which data were used to build the BPNN models.

Table 1. Layers of information for topography, soil fertility, weather, and site factors in ANN development. These factors were assumed to be main cause for within field variability.

Inputs				
Topographic Factors	Soil Fertility	Weather Factors	Site Factors	Outputs
Elevation	Soil P	May rainfall	SCN	Yield
Slope	Soil K	June rainfall	Weed	
Wetness index	Soil pH	Early July rainfall		
	Soil CEC	Late July rainfall		
		August rainfall		

landscape features, and site factors. Functionally, this is expressed as:

$$Y = f(TF, FF, WF, SF) \quad (1)$$

where

Y = soybean yield (kg ha^{-1})
TF = topographic factors
FF = soil fertility factors
WF = weather factors
SF = site factors.

There were a total of 14 input data layers for each field (table 1). A description of inputs and data preparation is given below.

Soybean Yield Data

Soybean yields at field sites were measured with a combine-mounted yield monitor for all years. Yields were averaged within each grid cell using all values of yield monitor data in each cell. Grid cells in which yield data were missing for any year were eliminated. There was only one season of data available for the Kusel fields, but the other fields had two or more seasons. Each field was a different size, but the grid cell size was the same for each (50 × 50 m). The number of grid cells in each field and the minimum, average, and maximum yields for each soybean growing season are listed in table 2.

Weather Factors

Water requirements of soybean vary with stage of development. Soybeans reach peak water use during grain filling, when plants are more sensitive to drought. During this period, large yield losses may occur on non-irrigated soybean grown in soils with low water holding capacities when rainfall is low (Ritchie et al., 1996). We assumed that rainfall is the primary cause of annual yield variability in our study.

Weather data were collected with a LICOR 1000 system at the field sites. Missing weather data were obtained from the Midwest Regional Climate Center for Cedar Rapids and Perry weather stations to fill gaps in observed data for the Keiper and Heck fields, respectively. These stations are approximately 10 km from the experimental fields. Rainfall was partitioned into five periods, similar to Liu et al. (2001), and referred to as May, June, early July, late July, and August rainfall. This segmentation was based on planting date. May rainfall was the rainfall in the 30 days following planting, June rainfall was the rainfall in the next 30 days, early July rainfall in the next 15 days, and late July rainfall was in the next 15 days. Cumulative rainfall values (mm) for the five periods are given in table 3. The rainfall partitioning allowed us to study the effects of rainfall patterns on soybean yield. It also allowed us to account for excess water, especially if it caused poor germination or reduced plant population due to flooding early in the season or reduced yield due to delays in harvest dates (Irmak et al., 2005). Seasonal rainfall, calculated from planting to harvest (table 3), was not used for ANN training.

Topographic Factors

Topographically derived hydrological indices are indicators of locations in a field that may be prone to excess water accumulation or to water stress. Spatial relationships between topographic features and hydrologic processes affect water flow and accumulation, which have major effects on soil water balance (Moore and Grayson, 1991; Garbrecht and Martz, 2000).

Three hydrological indices were used as input layers in the ANN to account for the effects of topographic attributes on yield: elevation (m), slope (%), and wetness index (WI). For all fields, DEM data were available, and elevation varied among fields. Elevation change (difference between minimum and maximum points) was about 9, 9, 20, 15, 5, and 2.5 m for Keiper Ray 100, Keiper Ray 80, Kusel North, Kusel South, McGarvey, and Heck fields, respectively (table 4). Elevation data for each field were normalized based on the minimum value before use in the ANN.

Table 2. Number of grids (0.2 ha size in each) and measured minimum, average, and maximum soybean yield (kg ha^{-1}) at field sites for each soybean growing season.

		Yield (kg ha^{-1})			
Field	Year	No. of Grids	Min.	Avg.	Max.
Keiper Ray 80	1995	60	3131	3543	4106
	1997	60	3795	4259	4499
Keiper Ray 100	1996	48	2663	3400	3947
	1998	48	3231	4009	4579
Kusel North	1998	59	2534	3015	3477
Kusel South	1997	59	2947	3260	3755
McGarvey	1995	77	1875	3021	3398
	1997	77	996	2126	3256
	1999	77	2105	2826	3354
Heck	1998	98	1358	3103	3906
	2000	98	2359	2753	3082

Table 3. Rainfall (mm) was partitioned into 5 groups at field sites: May, June, early July, late July and August. The Segmentation of rainfall during the growing period was based on the planting date.

		Rainfall Amounts (mm)					
Field	Year	May	June	Early July	Late July	August	Total[a]
Ray 80	1995	127	74.6	27.2	36.2	48.8	364.1
	1997	94	115.9	19.1	0.5	228.8	556.4
Ray 100	1996	216.2	63.6	40.1	12	68.6	471.6
	1998	123	147.5	18	94	104	727.5
Kusel North	1998	140.8	196.2	185.0	20.0	104.0	667.0
Kusel South	1997	38.1	117	12	49	121	511.1
McGarvey	1995	122.9	73.7	64.9	116.7	97.2	507
	1997	54.0	82.0	4.0	37.0	47.0	224
	1999	135.3	165.6	113.1	134.7	177.3	726
Heck	1998	121	262	72	34	103	609
	2000	89.9	153.9	41.7	41.8	27.4	374.1

[a] Total seasonal rainfall (mm) at field sites from planting (May) to end of August. This data was not used in building the BPNN model.

Table 4. Range of elevation data (minimum and maximum) for six fields in Iowa (m). The elevation and its secondary attributes (slope and wetness index) were used in BPNN model to account for the effects of topographic factors on soybean yield variability.

Field	Minimum (m)	Maximum (m)
Keiper Ray 80	28.42	37.22
Keiper Ray 100	20.81	29.29
Kusel North	15.50	34.36
Kusel South	11.44	26.87
McGarvey	98.38	104.02
Heck	98.04	101.11

The wetness index (WI) has been shown to characterize spatial distribution of zones of surface saturation and soil water content in landscape (Moore et al., 1988, 1992). WI is the logarithm of the ratio of specific catchment area and slope:

$$WI = \ln\left(\frac{A}{\tan\beta}\right) \quad (2)$$

where A is specific catchment area (upslope area per unit width of contour), and β is slope gradient (radians). ArcView GIS 3.2 software was used to compute elevation, slope, and wetness index for each cell. Inverse distance-weighted interpolation was used to predict the values for any unmeasured locations. The mean value was then computed for each cell using the zonal statistics tool of the same software.

Soil Fertility Factors

Four soil fertility factors were included in the ANN: soil pH, phosphorus, potassium, and cation exchange capacity (CEC) (table 1). The soil fertility factors were obtained from analysis of soil samples taken from each grid cell in all fields (T. E. Fenton, unpublished data). Two years of soil nutrient analysis data were available for the McGarvey and Heck fields (1997 and 2000). The samples were collected in the center of every other grid cell, and kriging was used to interpolate values for non-sampled cells. Then zonal statistics were used to calculate the mean value for each grid cell. For other fields, samples were collected for each grid in the field. Two years of yield data were collected for the Keiper fields, but soil samples were collected just once in each grid cell. Therefore, we assumed that the soil fertility data for the Keiper fields were similar across years.

Site Factors

Site characteristics consist of other yield-limiting factors that occur in complex combinations across fields, such as pests, diseases, and weeds. We included two site factors: soybean cyst nematode (SCN eggs/100 cc soil at planting), and weed density scale (unitless). SCN egg counts, collected from each grid in the McGarvey and Heck fields before planting each spring, varied widely within and between fields. The Keiper and Kusel fields did not have SCN problems. The weed distribution was observed in all fields and ranked from none (density = 0) to very high (density = 4).

ANN Architecture and Parameter Selection

There are several types of ANN models differentiated by the method of connecting nodes, the methods of computing weights, the number of nodes in hidden layers, and the type of transfer function between layers. The architecture determines how weights are interconnected in the network and which learning rules may be used (MathWorks, 2005). Selection of learning rules is important because it affects what input function, transfer function, and parameters will be used for the ANN model. The back-propagation learning rule used to train the feed-forward network architect is a universal approximator (Haykin, 1994) and one of the most widely used in SSM (Drummond et al.; 1995; Shearer et al., 1999; Wilkerson et al., 1999; Braga, 2000; Liu et al., 2001). Given sufficient hidden nodes, multi-layer-feed-forward network architectures can approximate any function of interest to any desired degree of accuracy (White et al., 1992). In this research, a standard fully connected, feed-forward, back-propagation neural network (BPNN) design was used, as shown in figure 2. A sigmoid function was used as an input transfer function in all cases. Details of the BPNN implementation are described by Dayhoff (1990). The architecture of the BPNN used in this study was as follows:

- Number of layers = 3 (input, hidden, and output).
- Number of neurons in the hidden layer = 3 to 15.
- Type of activation functions = sigmoid for hidden layer, linear for output layer.
- Number of nodes in input layer = 14 (table 1).
- Batch size = 154, 563, 543, and 120 for cases 1, 2, 3, and 4, respectively (table 5).
- Number of nodes in output layer = 1 (fig. 2).
- Network error type = mean square error.

The model included 14 nodes in the input vector, 15 nodes in the hidden layer (maximum nodes = 15), and one node in the output vector (soybean yield). The ThinksPro neural network software package was used for computations (Logical Designs Consulting, Inc., available at: www.logicaldesigns.com/ThinksPro.htm). The default initial weights from ThinksPro were used for the hidden and output layers together with a constant learning rate of 0.01 for both layers. We used trial and error to select the optimum parameter values (connection weights) that would give the most accurate results.

Training and Testing Data Preparation

The training dataset is a group of inputs from grid cells used for fitting the connection weights of the BPNN model (learning). The testing dataset includes a set of examples from grid cells used to assess the performance of a trained BPNN. We developed three case studies using the same model structure (fig. 2). For each case study, we prepared training and testing datasets from the fields in Iowa (table 5). Details of training and test preparation for each case study are summarized below.

Case 1: Predicting Within-Field Spatial Variability in Independent Years

The McGarvey field was the only site with three years of data that allowed us to test the performance of the BPNN for predicting yield in an independent year after training it with two years of data. Data from 1995 and 1997 together were used to build the McGarvey BPNN model 1 (table 5). After the BPNN was trained with two years of data, its performance was assessed in 1999, an independent year.

Case 2: Predicting Spatial Yield Variability in Independent Fields

A BPNN was trained using data from four fields collected from 1995 to 2000 (table 5), resulting in a total of seven

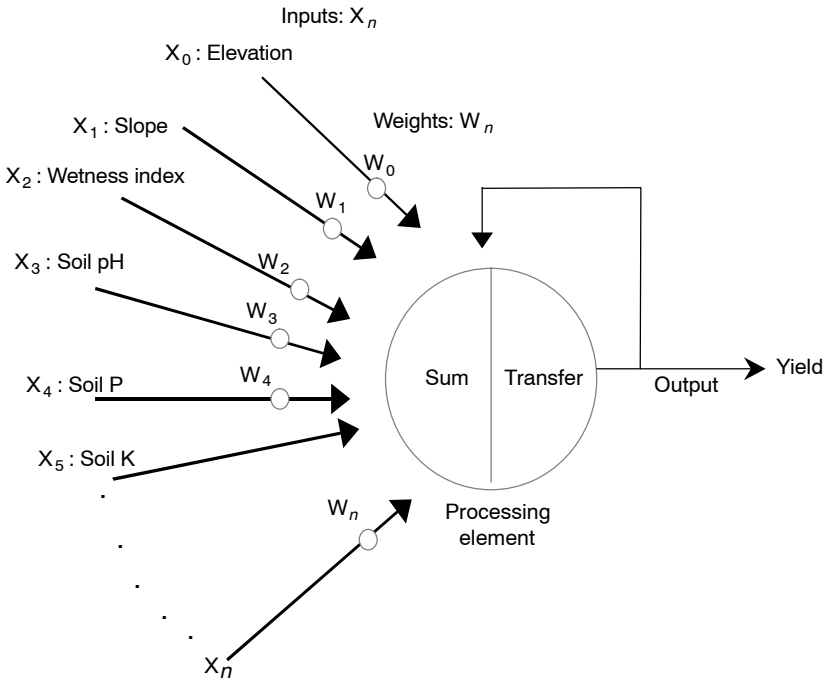

Figure 2. Schematic representation of fully connected, feed-forward, back-propagation neural network (BPNN) design for calculating soybean yield.

year-by-field combinations. The data were obtained from the Kusel (North and South), McGarvey, and Heck fields (table 5). Once the ANN was trained, its performance was evaluated for predicting yield in two fields over two years that were not used in the training: the Ray 100 (1996 and 1998) and Ray 80 (1995 and 1997) fields (fig. 1).

Case 3: Attributing Yield Losses to Stress Factors

The last model was built to determine if the BPNN could attribute yield losses due to three stress factors (table 5). We used the dataset created by Irmak et al. (2002a), who developed a combined crop model-regression approach to estimate yield reduction due to four factors: water, soybean cyst nematodes, soil pH, and weeds. Irmak et al. (2002a) used a two-year data set from the Keiper Ray 80 field in Iowa in which yield-decreasing factors of soybean cyst nematodes, soil pH, and weeds were artificially imposed on 24 of 60 grid cells and a combination of multiple stresses was imposed on 8 grid cells over two years.

Table 5. Three different cases of ANN modeling. In all three cases, the ANN models had fully connected feed-forward architecture, back-propagation learning, and one hidden layer. The field names, batch size for the training dataset, validation years, and total number of grids (*N*) at validation sites are given for training and testing years.

	Training		Testing		
Case	Batch Size	Field	Field	Testing Years	Total *N*
1	154	McGarvey	McGarvey	1999	77
2	543	Kusel North	Keiper Ray 100	1996, 1998	96
		Kusel South	Keiper Ray 80	1995, 1997	120
		McGarvey			216
		Heck			
3	120	Ray 80[a]	Keiper Ray 80[b]	1995, 1997	120

[a] Yield data from the Keiper Ray 80 field is modified as a result of imposing artificial stress factors (Irmak et al., 2002a).
[b] Actual dataset from the Keiper Ray 80 field.

For each stress factor, a yield loss was computed using the relationships between the level of imposed factor and yield from existing literature (Mortensen et al., 1999; Niblack et al., 1993; Paz, 2000). Modified yield in each grid cell was then calculated by subtracting the loss from the actual yield with the assumption that the yield-limiting factors were independent and additive:

$$Y_m = Y_a - \sum_{i=1}^{3}(1-\lambda_i).Y_a \quad (3)$$

where

Y_m = grid modified yield after stress is imposed (kg ha^{-1})
λ_i = yield reduction factor (0 to 1) due to stress factor i, where i = 1 for pH, i = 2 for SCN, and i = 3 for weeds
Y_a = grid actual yield for a given year (kg ha^{-1}).

This equation was applied only to grid cells in which the stress factors were imposed artificially. Thus, a total of 32 grid cells had reduced yields due to imposed stress factors. The rest of the grid cells (28) did not have stress imposed, and yield values remained those that were observed. The final dataset included both treated and non-treated grid cell yield values. This resulted in a modified dataset with 32 damaged grids and 28 undamaged grids. The details of this dataset are documented by Irmak et al. (2002a). This modified dataset was then used in case 3 for evaluating the ANN's ability to attribute yield losses across two seasons. As in the first two case studies, the BPNN model in case 3 had 14 nodes in the input layer and one node (yield) in the output layer. The only difference was that a portion of the input and output dataset was modified across two seasons. The steps to quantify the effects of multiple stress factors with the BPNN technique are summarized as follows:

Step 1: Train the model using the modified dataset for the Keiper Ray 80 field (Irmak et al., 2002a). The inputs and outputs are from 120 grid cells across two seasons. Obtain

spatial distribution of yield under stressed (due to SCN, soil pH, and weeds) and non-stressed conditions.

Step 2: For testing, use the original input and output dataset of the Keiper Ray 80 field for the 120 grid cells (table 2). Use the trained BPNN from step 1 to estimate yields for each grid cell. This will give yield predictions under non-stressed conditions because the dataset did not include any artificially imposed stress factors.

Step 3: Subtract the fitted yield in step 1 for each grid cell from the yield estimate (step 2) for which damage was assumed to be zero. Calculate the error between the estimated and imposed yield losses for each stress factor at each grid for comparisons.

Step 4: Repeat this procedure for each imposed stress factor (SCN, soil pH, and weeds) to estimate yield losses and corresponding BPNN prediction errors.

Data Processing

Input Preprocessing

There were large variations in inputs and yields within and among the six fields studied. Therefore, each input was normalized before building the BPNN models. The input range and data for each field were described in detail by Paz (2000) and Irmak (2002). For this article, we used the mean/standard deviation transformation to normalize the inputs:

$$X'_{ijk} = \frac{X_{ijk} - \overline{X}_{jk}}{SD_{jk}} \tag{4}$$

where X'_{ijk} is the normalized input for grid i, year j, and field k; X_{ijk} is the input for grid cell i, year j, and field k without normalization; $\overline{X}_{jk}$ is the mean value of all grid cells in year j for field k; and SD_{jk} is the standard deviation. Elevation was first transformed by subtracting the lowest point in each field from each grid cell elevation within the same field and then normalized (eq. 4).

Output Layer (Yield)

The output variable for the ANN models was soybean yield. Due to temporal and spatial variations in yield (table 2), the yield data were normalized for the BPNN models in cases 1 and 2 using equation 5:

$$y'_{ijk} = \frac{y_{ijk}}{y_{\max,k}} \tag{5}$$

where y'_{ijk} is the normalized yield for grid cell i, year j, and field k (0 to 1), y_{ijk} is the measured yield for grid i, year j, and field k (kg ha^{-1}), and $y_{\max,k}$ is the measured maximum yield for each field k across years (kg ha^{-1}). Since the maximum potential yield greatly varies across fields, we assumed that normalizing yield for each field would improve the accuracy of BPNN predictions for other fields. For case 3, yields were not normalized because only one field was used.

Statistical Analysis

Measured yield data were used for the output layer during BPNN training and testing. The results are analyzed using root mean square error of fitting (RMSE) and of prediction (RMSEP):

$$\text{RMSE} = \sqrt{\frac{1}{ygf}\sum_{k=1}^{f}\sum_{j=1}^{y}\sum_{i=1}^{g}(y'_{ijk} - \hat{y}'_{ijk})^2} \tag{6}$$

where RMSE is the root mean square of error between measured and predicted normalized yield (0 to 1); f, y, and g are the total number of fields, years, and grid cells, respectively; y'_{ijk} is predicted normalized yield (0–1); and $\hat{y}'_{ijk}$ is measured normalized yeild (0–1).

RMSE is unitless in equation 6, since we used normalized yield value (0 to 1) to train the BPNN and to evaluate predictions. The measured and predicted yields were then de-normalized to obtain the corresponding training and testing statistics (in units of kg ha^{-1}) for each year for all six fields. The statistics used to compare results were: (1) mean absolute error, (2) % error ([mean predicted yield − mean measured yield]/mean measured yield × 100), (3) RMSE, (4) % RMSE based on the actual mean yield, and (5) r^2 of a linear regression line.

Results and Discussions

Case 1: Predicting within-Field Yield Variability in Independent Years

Our first goal was test the ability of the BPNN model to predict spatial distribution of yield variability within a field in independent years. The 1995 and 1997 data from the McGarvey field were combined to train the model for this case (case 1: number of grid cells = 154, table 5). The trained model was used to evaluate its ability to predict yields in 1999, an independent year, in the same field. There was considerable temporal and spatial yield variation in the McGarvey field. Mean yields in the field were 3021, 2126, and 2826 kg ha^{-1} for 1995, 1997, and 1999, respectively (table 2). The lower yields in 1997 were probably due to the lower amount and timing of rainfall that occurred during the growing season (table 3). Soybean plants received little rainfall during both vegetative and reproductive phases (early grain filling period) in 1997 due to drought. These stress patterns led to low yields.

Figure 3 shows de-normalized predicted versus actual yield data for training and testing datasets for this case. Yield predictions matched observed data well for the training dataset (1995 and 1997). The RMSE of fitting was 122 and 141 kg ha^{-1} for 1995 and 1997, respectively (table 6). The yield predictions from testing were shown as Δ in figure 3. For these independent data, the model overpredicted yields for most grid cells in the field. Yield prediction error (RMSEP) was as high as +21.8% (overprediction) and as low as −14.3% (underprediction) of actual yield. On average, the RMSEP was 298 kg ha^{-1} (10.4% of mean actual yield) for 1999, the independent validation year, with a slope of regression line of 0.91 and r^2 of 0.57 (fig. 3, table 6). These results indicated that the BPNN captured the major patterns of spatial yield variability for an independent year in the same field.

Case 2: Testing ANN Performance for Independent Fields

For the second case, our objective was to assess the performance of the BPNN for predicting spatial distribution of yields for independent fields. The BPNN model was first trained using the Kusel North, Kusel South, McGarvey, and Heck data. The trained BPNN was used to predict yields for the Ray 100 and Ray 80 fields. Soybean yield variations were higher in the Ray 100 field than in the Ray 80 field for both years (table 2). For instance, yield variations in Ray 100 were

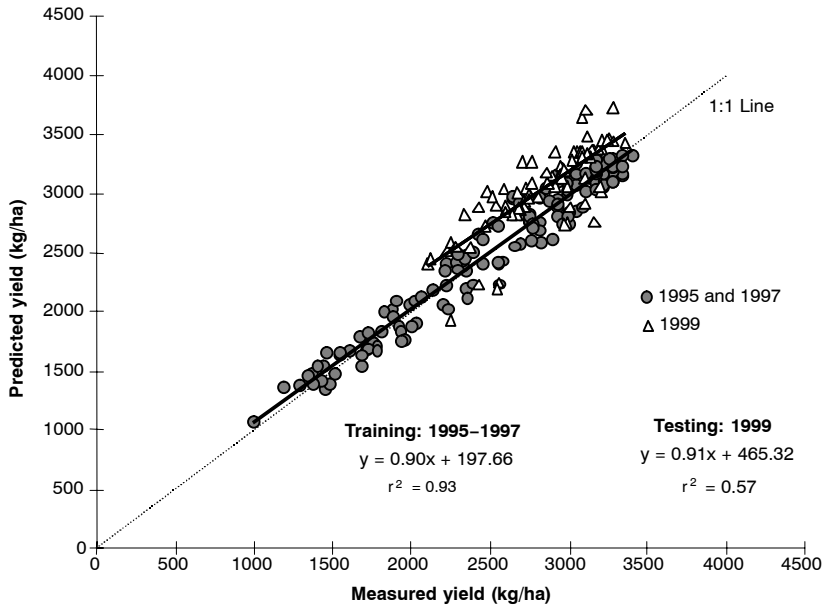

Figure 3. Predicted vs. measured yield (kg ha^{-1}) from BPNN training using data from the McGarvey field (case 1). The trained BPNN was used to test the results for an independent year (1999) for the same field.

Table 6. ANN performance for an individual year on the field level for predicting yield in the McGarvey field (case 1). Statistics include number of grid cells, mean measured (M) and predicted (P) yield, mean absolute error (MAE), % error, root mean square of error (RMSE), and r^2 of regression line.

			Yield (kg ha^{-1})		Error			
	Year	N	Measured	Predicted	MAE (kg ha^{-1})	(P–M)/M × 100	RMSE[a] (kg ha^{-1})	r^2
Training	95	77	3021	3048	98	0.9	122 (4.1)	0.85
	97	77	2135	2125	123	−0.5	141 (6.6)	0.93
Testing	99	77	2818	3021	273	7.2	298 (10.4)	0.57

[a] Analysis of data from validation set (i.e., RMSEP = 298 kg ha^{-1}).

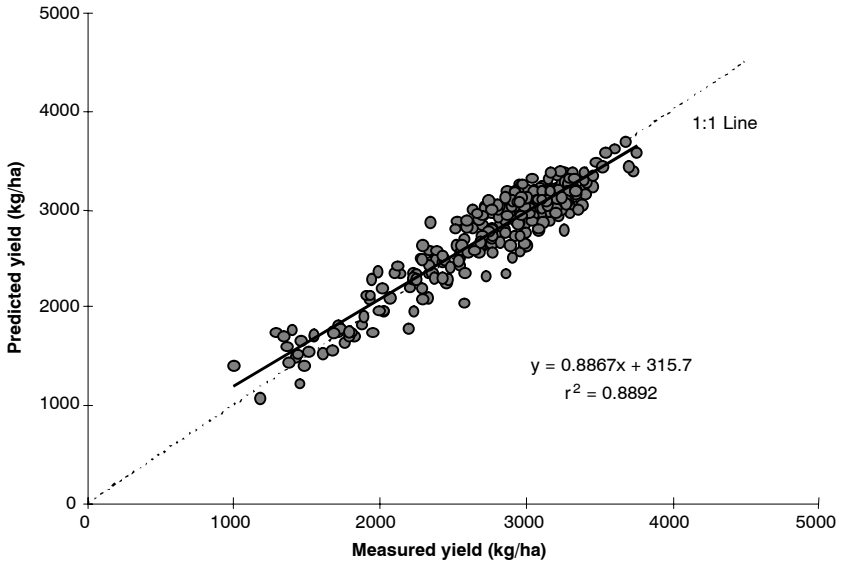

Figure 4. Predicted vs. measured de-normalized yield from BPNN training using data from four fields collected during 1995 to 2000 (total of seven year × field combinations; case 2 training).

32.5% and 29.4% in 1996 and 1998, respectively, with average yield 3400 and 4009 kg ha^{-1}. For the Ray 80 field, average yields were 3544 and 4259 kg ha^{-1}, with 23.7% and 15.6% variation for 1995 and 1997, respectively.

Figure 4 shows predicted and measured de-normalized yields for all four fields that were used to train the model for case 2. Predicted yields matched observed data well. The model explained 89% of the yield variation. On average, the

Table 7. BPNN performance on individual field basis for predicting soybean yield (case 2). The BPNN was trained using data from four fields, and the trained BPNN was used to predict yields for Ray 100 and Ray 80 fields.

				Average Yield (kg ha^{-1})		Error			
	Fields	Year	N	Measured	Predicted	MAE (kg ha^{-1})	(P−M)/M × 100	RMSE[a] (kg ha^{-1})	r^2
Training	Kusel North	1998	59	3015	3000	108	−0.5	137 (4.5)	0.48
	Kusel South	1997	57	3260	3268	108	0.2	139 (4.3)	0.49
	McGarvey	Field[b]	231	2658	2654	4	−0.1	189 (7.1)	0.88
	Heck	Field[b]	196	2928	2925	130	−0.1	177 (6.0)	0.83
	All fields	All years	543	2857	2854	134	−0.1	175 (6.1)	0.87
Testing[c]	Ray 100	96	48	3400	3632	370	6.8	466 (13.7)	0.48
		98	48	4009	4025	303	0.4	417 (10.4)	0.12
		Field[b]	96	3705	3829	337	3.4	442 (11.9)	0.36
	Ray 80	95	60	3544	3703	255	4.5	354 (10.0)	0.23
		97	60	4259	4415	178	3.7	226 (5.3)	0.14
		Field[b]	120	3901	4059	217	4.0	297 (7.6)	0.68

[a] Data in parenthesis shows the % RMSE or % RMSEP based on the actual yield.
[b] Analysis is based on combination of data from all years, field level.
[c] Analysis of data from validation set (i.e., RMSEP = 466 kg ha^{-1} for 1995 in the Ray 100 field).

percentage error was less than 1% of mean actual yield for all fields that were calibrated (table 7). Again, the percentage error was calculated as 100 × (mean predicted yield – mean measured yield)/mean measured yield. The RMSE values were 4.5%, 4.3%, 7.1%, and 6.0% of mean actual yields for the Kusel North, Kusel South, McGarvey, and Heck fields, respectively.

The trained BPNN gave similar results for both years for the Ray 100 testing site (table 7, fig. 5). There was a mixture of under- and overpredictions, except for a few grid cells with poorly predicted yields in 1998 in this field. The percentage error for the Ray 100 field was 6.8% and 0.4% of actual yields in 1996 and 1998, respectively. For the Ray 80 field, the ANN usually overpredicted yields for both years. At the field level, the RMSEP values were 442 kg ha^{-1} (11.9% of mean actual yield) and 297 kg ha^{-1} (7.6% of mean actual yields) for the Ray 100 and Ray 80 fields, respectively. The r^2 values were slightly better for the Ray 100 field than for the Ray 80 field. After taking into account all factors, the model explained 36% and 68% of yield variability for the Ray 100 and Ray 80 fields, respectively.

In summary, results showed that the ANN model usually resulted in overpredictions for independent fields. The model gave higher yields, especially for high-yielding grid cells. The RMSEP was less than 14% of mean actual yield for the worst case. The r^2 values were generally low and not consistent from year to year in the same field.

Case 3: Attribution of Yield Losses to Stress Factors

Case 3 uses the dataset developed by Irmak et al. (2002a) to determine if the BPNN could attribute yield losses due to three stress factors. Figure 6 shows errors ([predicted yield loss – imposed yield loss]/imposed yield loss) versus imposed yield losses for all three variables for 1995 and 1997 at the Keiper Ray 80 field. The model tended to overpredict yield when the level of pH damage was greater (greater imposed yield loss) in 1995. The ANN model underestimated yield loss consistently in 1997 (fig. 6a). On average, the mean absolute error (MAE) was 60 and 106 kg ha^{-1} for 1995 and 1997, respectively. The error was as high as 20.6% and as low

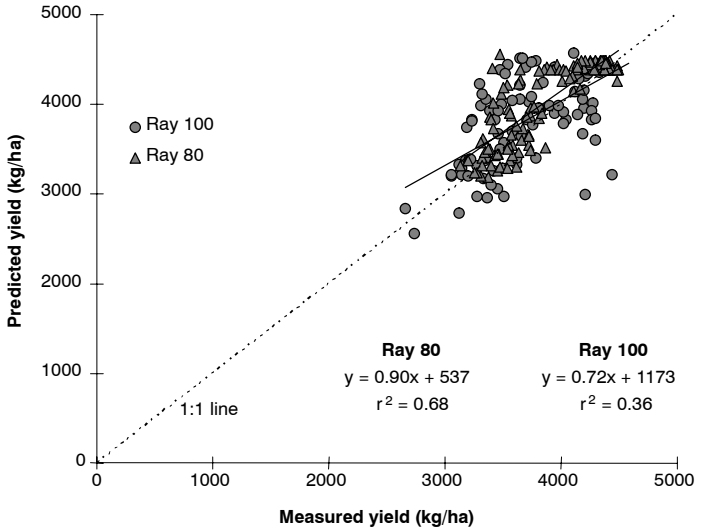

Figure 5. Predicted vs. measured de-normalized yield (kg ha^{-1}) for the Ray 100 and Ray 80 fields (case 2 testing). Predictions are based on the BPNN trained using data from four fields collected during 1995 to 2000 (total of seven year × field combinations).

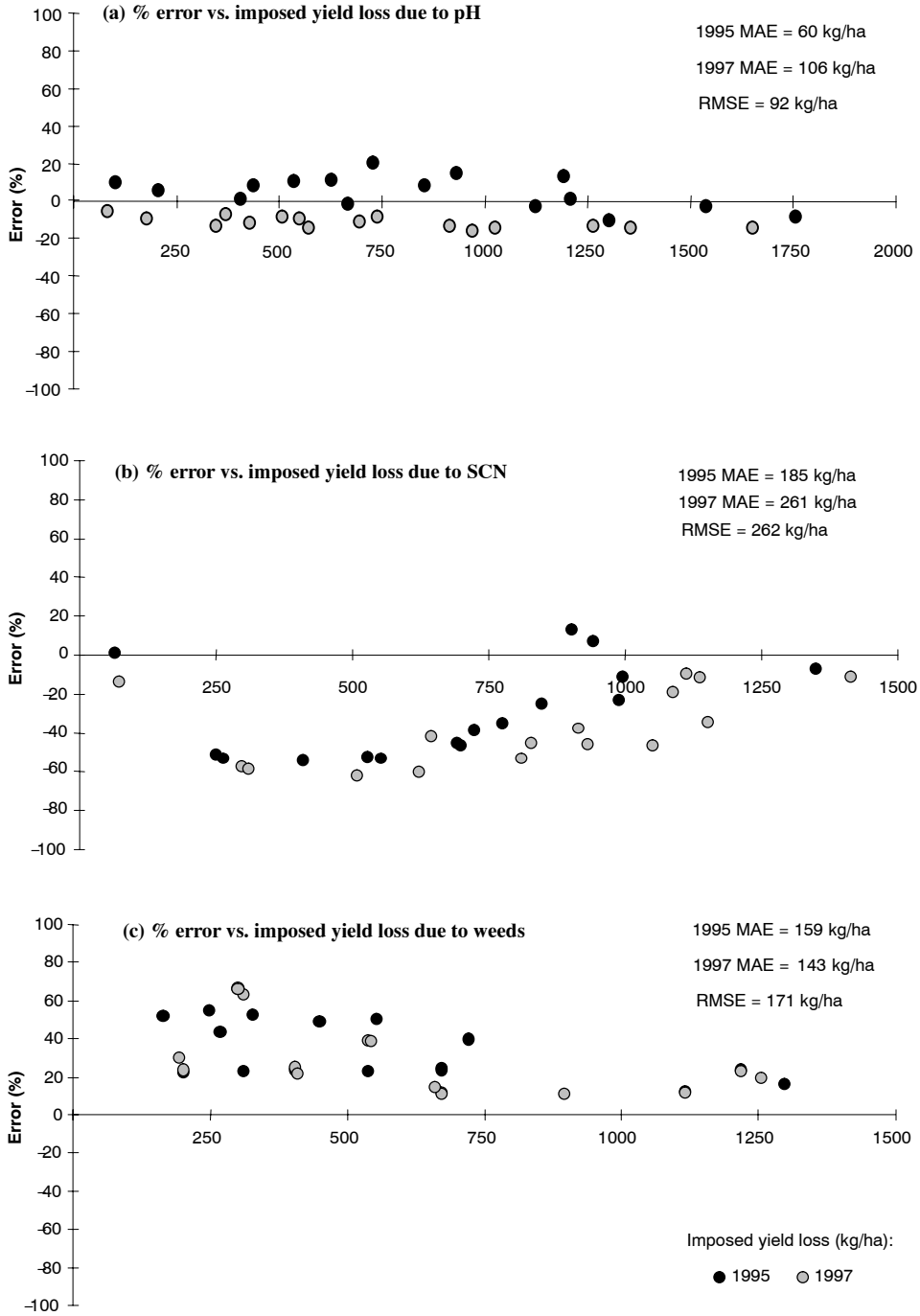

Figure 6. Percent error vs. imposed yield loss in each grid due to effects of (a) pH, (b) SCN, and (c) weeds in the Keiper Ray 80 field using the BPNN approach (case 3). MAE is mean absolute error.

as −10.1% of imposed yield loss in 1995, while it was as low as −15.7% of imposed yield loss in 1997. Mean RMSE was 92 kg ha^{-1} over two years of data, indicating that the accuracy of the method was within about 6% of the maximum imposed yield loss due to soil pH.

The ANN model underestimated yield loss by as much as 400 kg ha^{-1} for some of the grid cells that had imposed SCN damage (fig. 6b). The ANN underestimated yield loss consistently in 1997, and similar trends were obtained in 1995 except in two grid cells. The errors were as high as 13.1% and as low as –53.6% of imposed yield loss in 1995. On average, the MAE was 185 and 261 kg ha^{-1} for 1995 and 1997, respectively. The mean RMSE was 262 kg ha^{-1} over two years of data, or about 19% of the maximum imposed SCN damage. :–

The model overpredicted the effects of weeds on yield for both years (fig. 6c). The error was as high as 54.8% in 1995 and 61.0% in 1997. On average, the MAE was 159 and 143 kg ha^{-1} for 1995 and 1997, respectively. The mean RMSE of 171 kg ha^{-1} over two years of data indicates that the accuracy was within 14% of the highest imposed yield loss due to weeds.

Overall, our results indicated that when an ANN model was used for diagnostic purposes, standard errors of attribu-

tion were 92, 262, and 171 kg ha^{-1} for losses to soybean yields due to soil pH, SCN, and weeds, respectively. Irmak et al. (2002a) showed that the combined crop modeling-regression approach overpredicted the yield loss due to soil pH and weeds for 1995 and 1997, but underestimated yield loss consistently in 1997 for all grid cells that had imposed SCN yield loss in the same field. The standard errors of attribution in their study were 185, 220, and 163 kg ha^{-1} for soil pH, SCN, and weeds, respectively.

Conclusions and Discussion

In this research, a BPNN model was used to predict spatial distribution of soybean yields and to understand causes of spatial yield variability. For each objective, data from different fields in Iowa were used for training, while a portion of data was withheld to test the accuracy of yield predictions. The following conclusions were drawn:

- A BPNN model trained for a particular field (case 1) was able to capture the major patterns of spatial yield variability for the same field in independent years, i.e., RMSEP was 10.4% of actual yield for an independent year (table 6). The model explained 57% of yield variability for independent years in the same field after taking into account topography, soil fertility, weather, and site factors in this study.
- A BPNN model trained for a wide range of inputs from different fields (case 2) usually overestimated yield for independent fields. The model explained 36% to 68% of yield variability for independent fields (table 7).
- When a BPNN model was used for diagnostic purposes (case 3), results were better than those found by Irmak et al. (2002a) for quantifying yield losses due to soil pH. For the other variables, the combined crop model-empirical approach performed better.
- The BPNN trained for a specific field was more accurate for predicting yield in independent years than using a BPNN trained in other fields (comparison of case 1 vs. case 2).
- We found that the segmentation of rainfall input for BPNN was an important input. We also tried seasonal total rainfall in analysis, but the results were not improved relative to those presented in this article.

Readers should also note that it is unlikely that spatial data, like those used in this study, would be available for building an ANN model under most commonly practiced farming conditions. Therefore, it would be useful to build and test models for a producer's field using readily available information. In addition, it is important to select input variables for an ANN that have causal relationships with the output variable. An ANN trained with such inputs might mislead the user because the training might set connection weights that should not exist at all.

Another important consideration is that that use of an ANN for predicting yields for adjacent or nearby fields may set weights for inputs that are not yield limiting or critical in the training field. Therefore, the ANN model should be built for each field because the connection weights are field-specific.

References

Braga, R. P. 2000. Predicting the spatial pattern of grain yield under water limiting conditions. PhD diss. Gainesville, Fla.: University of Florida.

Burks, T. F., S. A. Shearer, R. S. Gates, and K. D. Donohue. 1999. Backpropagation neural network design and evaluation for classifying weed species using color image texture features. ASAE Paper No. 993047. St. Joseph, Mich.: ASAE.

Carpenter, G. A., M. N. Gjaja, S. Gopal, and C. E. Woodcock. 1997. ART neural networks for remote sensing: Vegetation classification from Landsat TM and terrain data. *IEEE Trans. Geosci. Remote Sensing* 35(2): 308-325.

Carpenter, G. A., S. Gopal, S. Macomber, S. Martens, and C. E. Woodcock. 1999a. A neural network method for mixture estimation for vegetation mapping. *Remote Sensing of Environ.* 70(2): 138-152.

Carpenter, G. A., S. Gopal, S. Macomber, S. Martens, C. E. Woodcock, and J. Franklin 1999b. A neural network method for efficient vegetation mapping. *Remote Sensing of Environ.* 70(3): 326-338.

Dayhoff, J. E. 1990. *Neural Network Architectures: An Introduction*. New York, N.Y.: Van Nostrand Reinhold.

Drummond, S. T., K. A. Sudduth, and S. J. Birrell. 1995. Analysis and correlation methods for spatial data. ASAE Paper No. 951335. St. Joseph, Mich.: ASAE.

Garbrecht, J., and L. W. Martz. 2000. Digital elevation model issues in water resources modeling. In *Hydrologic and Hydraulic Modeling Support with Geographic Information Systems*, 1-28. D. Maidment and D. Djokic, eds. Redlands, Cal.: Environmental Systems Research Institute, Inc.

Gliever, C., and D. C. Slaughter. 2001. Crop versus weed recognition with artificial neural networks. ASAE Paper No. 013104. St. Joseph, Mich.: ASAE.

Gopal, S., C. E. Woodcock, and A. H. Strahler. 1999. Fuzzy neural network classification of global land cover from a 1° AVHRR data set. *Remote Sensing of Environ.* 67(2): 230-243.

Haykin, S. 1994. *Neural Networks: A Comprehensive Foundation.* New York, N.Y.: Maxwell Macmillan College Publishing.

Hoogenboom, G., J. W. Jones, P. W. Wilkens, W. D. Batchelor, W. T. Bowen, L. A. Hunt, N. Pickering, U. Singh, D. C. Godwin, B. Baer, K. J. Boote, J. T. Ritchie, and J. W. White. 1994. Crop models. In *DSSAT* 3: 95-244. G. Y. Tsuji, G. Uehara, and S. Balas, eds. Honolulu, Hawaii: University of Hawaii.

Irmak, A. 2002. Linking multiple layers of information for understanding soybean yield variability. PhD diss. Gainesville, Fla.: University of Florida.

Irmak, A., J. W. Jones, W. D. Batchelor, and J. O. Paz. 2001. Estimating spatially variable soil properties for crop model use in precision farming. *Trans. ASAE* 42(6): 1867-1877.

Irmak, A., J. W. Jones, W. D. Batchelor, and J. O. Paz. 2002a. Linking multiple layers of information for diagnosing causes of spatial yield variability in soybean. *Trans. ASAE* 45(3): 839-849.

Irmak, A., W. D. Batchelor, J. W. Jones, S. Irmak, J. O. Paz, H. Beck, and M. Egli. 2002b. Relationship between plant available soil water and yield for explaining within-field soybean yield variability. *Applied Eng. in Agric.* 18(4): 471-482.

Irmak, A., J. W. Jones, and S. S. Japtap. 2005. Evaluation of the CROPGRO-soybean model for assessing climate impacts on regional soybean yields. *Trans. ASAE* 48(6): 2343-2353.

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

Keller, P. E., R. T. Kouzes, and L. J. Kangas. 1994. Three neural network based sensor systems for environmental monitoring. In *Proc. IEEE Electro94 Conference.* Piscataway, N.J.: IEEE.

Liu, J., C. E. Goering, and L. Tian. 2001. Neural network for setting target corn yields. *Trans. ASAE* 44(3): 705-713.

MathWorks. 2005. Neural network toolbox. Natick, Mass. The MathWorks, Inc. Available at: www.mathworks.com/products/neuralnet/.

Moore, I. D., and R. B. Grayson. 1991. Terrain-based catchment partitioning and runoff prediction using vector elevation data. *Water Resources Res.* 27(6): 1177-1191.

Moore, I. D., G. J. Burch, and D. H. Mackenzie. 1988. Topographic effects on distribution of surface soil water and the location of ephemeral gullies. *Trans. ASAE* 31(4): 1098-1107.

Moore, I. D., P. E. Gessler, G. A. Nielsen, and G. A. Peterson. 1992. Terrain analysis for soil specific crop management. In *Soil Specific Crop Management*, 27-55. P. C. Robert et al., eds. Minneapolis, Minn.: ASA-CSSA-SSSA.

Mortensen, D. A, A. R. Martin, F. W. Roeth, T. E. Harvill, R. W. Klein, M. Milner, G. A. Wicks, R. G. Wilson, D. L. Holshouser, D. J. Lyon, P. J. Shea, and J. T. Rawlinson. 1999. *WeedSOFT User's Manual*. Version 4. Lincoln, Neb.: University of Nebraska, Department of Agronomy.

Niblack, T. L., R. D. Heinz, G. S. Smith, and P. A. Donald. 1993. Distribution, density, and diversity of *Heterodera glycines* in Missouri. *J. Nematology* 25(4S): 880-886.

Noh, H. K., Q. Zhang, and S. Han. 2004. A neural network model of nitrogen stress assessment using a multispectral corn nitrogen deficiency sensor. ASAE Paper No. 041132. St. Joseph, Mich.: ASAE.

Odhiambo, L. O., R. E. Yoder, D. C. Yoder, and J. W. Hines. 2001a. Optimization of fuzzy evapotranspiration model through neural training with input-output examples. *Trans. ASAE* 44(6): 1625-1633.

Odhiambo, L. O., R. E. Yoder, and D. C. Yoder. 2001b. Estimation of reference crop evapotranspiration using fuzzy state models. *Trans. ASAE* 44(3): 543-550.

Paz, J. O. 2000. Analysis of spatial yield variability and economics of prescriptions for precision agriculture: A crop modeling approach. PhD diss. Ames, Iowa: Iowa State University.

Paz, J. O., W. D. Batchelor, T. S. Colvin, S. D. Logsdon, T. C. Kaspar, and D. L. Karlen. 1998. Analysis of water stress effects causing spatial yield variability in soybeans. *Trans. ASAE* 41(5): 1527-1534.

Paz, J. O., W. D. Batchelor, G. L. Tylka, and R. G. Hartzler. 2001a. A modeling approach to quantify the effects of spatial soybean yield limiting factors. *Trans. ASAE* 44(5): 1329-1334.

Paz, J. O., W. D. Batchelor, and G. L. Tylka. 2001b. Method to use crop growth models to estimate potential return for variable-rate management in soybeans. *Trans. ASAE* 44(5): 1335-1341.

Pinto, F. A. C., J. F. Reid, Q. Zang, and N. Noguchi. 1999. Guidance parameter determination using artificial neural network classifier. ASAE Paper No. 993004. St. Joseph, Mich.: ASAE.

Ritchie, S. W., J. J. Hanway, and H. E. Thompson. 1996. How a soybean plant develops. Special Report 53. Ames, Iowa: Iowa State University Cooperative Extension Service.

Shearer, S. A., J. A., Thomasson, T. G. Mueller, J. P. Fulton, S. F. Higgins, and S. Samson. 1999. Yield prediction using a neural network classifier trained using soil landscape features and soil fertility data. ASAE Paper No. 993042. St. Joseph, Mich.: ASAE.

Shock, B. M., G. A. Carpenter, S. Gopal, and C. E. Woodcock. 2002. ARTMAP neural network classification of land use change. In *Proc. World Congress of Computers in Agriculture and Natural Resources*, 22-28. F. S. Zazueta and J. Xin, eds. ASAE Paper No. 701P0301. St. Joseph, Mich.: ASAE.

White, H., A. R. Gallant, K. Hornik, M. Stinchcombe, and J. Wooldridge. 1992. *Artificial Neural Networks: Approximation and Learning Theory.* Cambridge, Mass.: Blackwell Publishers.

Wilkerson, J. B., R. Sui, W. E. Hart, L. R. Wilhelm, and D. D. Howard. 1999. Artificial neural networks for determining nitrogen status in corn. ASAE Paper No. 993042. St. Joseph, Mich.: ASAE.

Yang, C. C., S. O. Prasher, and R. Lacroix. 1996. Applications of artificial neural networks to land drainage engineering. *Trans. ASAE* 39(2): 525-533.

Yang, C. C., S. O. Prasher, R., Lacroix, S. Sreekanth, N. K. Patni, and L. Masse. 1997. Artificial neural network model for subsurface-drained farmlands. *J. Irrig. and Drain. Eng*. 123(4): 285-292.