CROP YIELD PREDICTION USING ARTIFICIAL NEURAL NETWORKS AND GENETIC ALGORITHMS

by

C. MAXWELL MARTIN

(Under the Direction of Gerrit Hoogenboom)

ABSTRACT

Previous research has established that large-scale climatological phenomena influence local weather conditions in various parts of the world. These weather conditions have a direct effect on crop yield. Consequently, much research has been done exploring the connections between large-scale climatological phenomena and crop yield. Artificial neural networks have been demonstrated to be powerful tools for modeling and prediction, and can be combined with genetic algorithms to increase their effectiveness. The goal of the research presented in this thesis was to develop artificial neural network models using genetic algorithm-selected inputs in order to predict southeastern US maize yield at various points throughout the year.

INDEX WORDS: Artificial Neural Networks, Genetic Algorithms, Crop Yield, Maize, Teleconnections, Sea Surface Temperature, Decision Support Systems

CROP YIELD PREDICTION USING ARTIFICIAL NEURAL NETWORKS AND GENETIC ALGORITHMS

by

C. MAXWELL MARTIN

B.A., The University of Georgia, 2009

B.S., The University of Georgia, 2009

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2009

CROP YIELD PREDICTION USING ARTIFICIAL NEURAL NETWORKS AND GENETIC ALGORITHMS

by

C. MAXWELL MARTIN

Major Professor:	Gerrit Hoogenboom
Committee:	Ron W. McClendon Walter D. Potter

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2009

DEDICATION

This thesis is dedicated to my parents, Charles and Valerie Martin, without whose love and support this thesis never could have been written.

ACKNOWLEDGEMENTS

I would like to thank the members of my committee for their patience and guidance, both during the writing and revision of this thesis and throughout the course of my research with them. In particular I wish to thank Dr. McClendon for kindly offering me the research position I have held during my time in this program, Dr. Hoogenboom for guiding the course of my studies and helping to improve the quality of my research with his expertise, and Dr. Potter for encouraging me to sign up for the Artificial Intelligence program and helping to smooth the path of my academic career with his kindness and enthusiasm. I also wish to thank Dr. Joel Paz for the invaluable contributions he made to my studies while a part of our research group. Additionally, I wish to thank my fellow students who have offered assistance in the course of my research, specifically Brian Smith for his assistance with the finer points of the Java neural network code, and Kevin Crowell and Bob Chevalier for their insightful comments during research group meetings.

This work was conducted under the auspices of the Southeast Climate Consortium (SECC; www.SEClimate.org) and supported by a partnership with the United States Department of Agriculture-Risk Management Agency (USDA-RMA), by grants from the US National Oceanic and Atmospheric Administration-Climate Program Office (NOAA-CPO) and USDA Cooperative State Research, Education and Extension Services (USDA-CSREES) and by State and Federal funds allocated to Georgia Agricultural Experiment Stations Hatch project GEO01654.

TABLE OF CONTENTS

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

By influencing regional weather patterns, large-scale meteorological phenomena can have a significant impact on agricultural production. Large-scale weather patterns such as the El Niño Southern Oscillation (ENSO) and the Pacific-North American pattern have been linked by research to local weather patterns in various locations around the world (Philander, 1990; Trenberth, 1997; Leathers et al., 1991). In turn, specific climatic conditions such as fluctuations in precipitation have been shown to have strong influences on crop failures in the United States (Ibarra and Hewitt, 1999), demonstrating that weather patterns may be valuable for modeling and predicting crop yield. Such predictions could be used by crop managers to minimize losses when unfavorable conditions may occur. Additionally, these predictions could be used to maximize crop prediction when potential exists for favorable growing conditions.

The links between meteorological phenomena and crop yield have been studied by researchers before, generally using methods of statistical analysis such as correlation analysis (Baigorria et al., 2008) or linear correlation (Travasso et al., 2008). Many of these studies have focused on specific regional impacts of the ENSO phenomenon. More recently, Martinez et al. (2009) studied the impacts of ENSO and several other meteorological phenomena on yield of maize in the southeastern US. Using a combination of linear correlation analysis and principal component regression, Martinez et al. (2009) created models of crop yield based on multiple

different indices of meteorological phenomena. Their research found the Pacific-North American pattern to be most strongly correlated with southeastern US maize yield.

Artificial neural networks (ANNs) are computational modeling tools which can be used to classify and predict data. ANNs have been applied to agricultural and meteorological research in the past with great success (Jain et al., 2003, 2006; Smith et al., 2007, 2009). Genetic algorithms (GAs) are a tool for computational optimization which can be combined with ANNs in various ways. Specifically, in cases where a large number of potential inputs are available, they can be used to select optimal subsets of inputs for model development (Guo and Uhrig, 1992).

The goal of the research presented in this thesis was the development of ANN models using GA-selected inputs for the prediction of maize yield in the southeastern US based on indices of large-scale meteorological phenomena. In order to achieve this goal, ANN models were developed using inputs from four indices of meteorological phenomena. In order to determine the subset of these inputs which would produce ANN models with the lowest error, a GA was used to search the input space. Searches were run for various prediction dates throughout the year to determine the effect of decreasing the amount of information available to the model on its accuracy. The results of these searches were then compared to the results of prior research on the use of indices of meteorological phenomena to predict crop yield.

Chapter 1 introduces the idea that indices of large-scale meteorological phenomena can be used to model and predict crop yield, providing an overview of previous research based around this idea. This chapter also discusses the specific computational modeling and optimization tools used in the research presented in this thesis. Chapter 2 describes the initial development of the ANN models used for predicting crop yield. The potential inputs to these

ANN models are then searched using a GA. Chapter 3 describes the effects of moving the prediction date of the ANN models earlier in the year, reducing the amount of information available for use in making predictions. The results of these searches are analyzed and compared to those of other researchers. Chapter 4 summarizes the results of the research conducted for this study and suggests possible avenues for future research.

CHAPTER 2

A GENETIC ALGORITHM & NEURAL NETWORK HYBRID FOR PREDICTING CROP YIELD BASED ON SEA SURFACE TEMPERATURE[1]

[1] Martin, C. M., R. W. McClendon, J. Paz, and G. Hoogenboom. To be submitted to *Expert Systems With Applications*.

ABSTRACT

Large-scale meteorological and climatological phenomena have been shown to impact crop growth, development, and yield, but these effects are highly complex. Research to explore the correlations between various climate indices and crop yield has been performed largely using statistical methods. The goal of this study was to determine the correlation between sea surface temperature and maize yield in the southeastern US. Specific objectives included determining preferred model parameters, finding the subset of available inputs which minimized model error, and determining the predictive accuracy of a final model by applying it to an independent evaluation data set. Artificial neural network (ANN) models were developed to predict maize yield using inputs from four sea surface temperature indices. A genetic algorithm (GA) search was performed to select the preferred input variables for the ANN model in order to minimize the error. This GA used multiple instantiations of the ANN models as its fitness function. The search utilized a novel tiered evaluation system where different data sets were used to evaluate individual solutions for selection and detecting convergence. The results of this search were then used to select the best set of ANN inputs to develop a final model for the prediction of southeastern US maize yield based on sea surface temperature. Initial ANN models using all available inputs were able to achieve a mean absolute error of 1558 kilograms per hectare (kg/ha) on the selection set. The final model had a mean absolute error of 1045 kg/ha on the selection set, and an error of 1840 kg/ha for the independent evaluation set. The final model error on the evaluation set was comparable with the overall data set average standard deviation of 1923 kg/ha. Using the GA-selected inputs produced an improvement in model error for both the development and selection sets, indicating that the tiered fitness scheme developed for this research is worthy of further investigation. This indicated that the approach outlined in this study

could be of great benefit to crop managers. Further research into applying the techniques presented here to other types of data would determine whether they are of general use for other applications.

INTRODUCTION

Research has shown that large-scale meteorological and climatological phenomena such as the El Niño-Southern Oscillation (ENSO) can influence local weather conditions in various locations around the world (Philander, 1990; Trenberth, 1997). Some of this research has explored the impact of the ENSO pattern on crop yield in the southeastern US (e.g. Handler, 1990; Hansen et al., 1998, 1999; Izaurralde et al., 1999). Other research has suggested that indices such as tropical North Atlantic sea surface temperature (Enfield, 1996) and the Pacific-North American pattern (Leathers et al., 1991) are correlated with local weather conditions. Recently, Martinez et al. (2009) studied the correlations between the values of four indices of sea surface temperature and southern US maize yield using principal component analysis, a statistical analysis technique. Martinez et al. (2009) used the values of the indices for tropical North Atlantic sea surface temperature, the Pacific-North American pattern, and the Bermuda high index in addition to the Japanese Meteorological Association index, which measures the behavior of the ENSO phenomenon. This approach, which was more inclusive than prior research related to ENSO, allowed for the possibility that meteorological phenomena other than ENSO could also have a strong impact on crop yield. Additionally, their approach could be applied to areas where ENSO does not have a strong impact. However, Martinez et al. (2009) focused only on exploring correlations and made no attempt to create a predictive model. Additionally, their entire data set was used in model development, with no data held back for independent evaluation. Their research indicated that the Pacific-North American pattern had the

greatest correlation with crop yield residuals. Other studies have explored the impact of diverse meteorological phenomena on different crops throughout the world, using various statistical methods. Everingham et al. (2003) used Monte Carlo procedures to examine correlations between the southern oscillation index and sugarcane yield in Australia. D'Arrigo and Wilson (2008) developed regression models based on the relation between equatorial Pacific sea surface temperature and the growth of rice in Indonesia. Travasso et al. (2009) used cumulative probability distribution functions to analyze potential correlations between sea surface temperature anomalies and yield for maize, sunflower, and soybeans in Argentina.

Artificial neural networks (ANNs) have been applied to model complex relations, and they have demonstrated the capability to handle a large number of inputs and generalize correlations (Bose and Liang, 1996; Haykin, 1999). Jain et al. (2003; 2006) used artificial neural networks for the prediction of air temperatures, specifically in order to predict the possibility of frost. In order to increase model accuracy, Jain et al. (2003; 2006) tested various combinations of inputs to the ANN models. Smith et al. (2007; 2009) improved upon this air temperature prediction system by developing multiple instantiations of the same ANN models, and compared these in order to select the most accurate model. Shank et al. (2008a, 2008b) applied ANNs to predict dewpoint temperatures. In their approach, they developed an ensemble model, with a number of separate ANNs that predicted for different time periods ranging from one to twelve hours in advance.

Genetic algorithms (GAs) have been used as a tool to enhance the performance of ANN models. Dasgupta and McGregor (1992) applied GAs to design different ANN architectures for specific applications. Their GA operated on two different levels, selecting both the connection structure of the ANN and its weight values. Yang and Honavar (1998) used GAs to select subsets

of features for pattern classification to develop a special type of ANN models based on constructive learning algorithms. Henderson et al. (1998) applied a GA to choose parameter values such as ANN learning rate, momentum, and number of hidden nodes. GAs have been used in tandem with ANNs in a number of other ways, such as training ANNs directly in place of the traditional backpropagation algorithm and constructing novel neural network architectures (Whitley, 1995). Guo and Uhrig (1992) showed that in cases where a large amount of information is available as potential input into the model, this large number of inputs can have a negative effect on the ANN models, overwhelming them with excessive and unnecessary information. For their study, Guo and Uhrig (1992) used a GA search to select the optimal set of inputs over 20 potential inputs. The fitness scheme used by Guo and Uhrig (1992) penalized solutions with a larger number of inputs in an attempt to develop smaller, more efficient networks to reduce training time. However, penalizing solutions based on the number of inputs could also cause the GA to select a sub-optimal set of inputs to the ANN model.

The goal of the research presented herein was to develop a hybrid ANN model with GA-selected inputs for predicting maize yield within the southeastern U.S. based on large-scale climate indices. The specific objectives included 1) to determine the preferred ANN architecture and parameters, 2) to determine the subset of inputs which minimized the error in predicting crop yield, and 3) to determine the predictive accuracy of the final ANN model when applied to an independent evaluation data set.

METHODOLOGY

A. Data

ANN models were developed to predict maize yield based on prior values of sea surface temperature indices and the latitudinal and longitudinal coordinates of the districts for which the

predictions were being made. Data from 94 counties in seven crop reporting districts in Georgia, Florida, and Alabama were used. Simulated yield data were used in order to remove possible influences on crop growth such as technological advances and changing chemical inputs and production practices. The yield data were generated using the Cropping System Model (CSM; Jones et al., 2003) of the Decision Support System for Agrotechnology Transfer (DSSAT; Hoogenboom et al., 2004), as described by Persson et al. (2009a, 2009b). The simulated yield data were generated for the same locations as used by Martinez et al. (2009), containing maize yield data from 129 counties in the southeastern US that have significant maize production. Although the data for this study were taken from the same area as the study by Martinez et al. (2009), their study was based on detrended observed yield, rather than simulated yield. The simulated crop yield data were divided geographically by established crop reporting districts. Those districts containing a smaller number of counties were not included, reducing the total number of counties in the data set from 129 to 94. Data from the remaining seven largest crop reporting districts (two in Georgia and Florida, three in Alabama) were used. These data were taken from the years 1951 to 2006, since 1950 was the earliest year for which sea surface temperature index values were available.

Some limitations were placed on the CSM when generating data for this study. Simulated yield data from three planting dates were available from the data generated by Persson et al. (2009a, 2009b). However, only data from the middle planting date (the 76th day of the year, i.e., March 17 (March 16 during leap years)) were used. Also, only yield using the highest level of fertilizer were included in the dataset to avoid any nutrient stresses. Additionally, the simulations for only one cultivar of maize were included: Pioneer 31G98. This is one of the most commonly grown hybrids in the region. As described in Persson et al. (2009a, 2009b), independent

simulations were performed for each county, and within each county simulations were performed using the three most common soil types. Only the results for the most common soil type representative of maize production within each county were included in this study. Simulated yield values from each county in the area covered by the study were averaged across all counties contained within each of the seven crop reporting districts. These averaged annual yield values were used as the target values for the ANN models. The units specified for the simulated yield were kilograms of dry matter per hectare.

Values from four sea surface temperature indices were used as inputs. These indices included the Japan Meteorological Agency (JMA) index (Center for Ocean-Atmospheric Prediction Studies, 2009), the North Atlantic Oscillation (NAO) index (National Oceanic and Atmospheric Administration, 2009), the Pacific-North American (PNA) teleconnection pattern (National Oceanic and Atmospheric Administration, 2009), and the Oceanic Niño Index (ONI) (National Oceanic and Atmospheric Administration, 2009). These indices measure various large-scale climatological phenomena, such as the dominant ENSO. Each index consists of one value per month, though values of the ONI are averages taken from a three-month window. As an example, the March ONI value would be an average of the values for the months of February, March, and April. The final harvest of maize in the southeastern US generally occurs during the summer months (Martinez et al., 2009). Therefore, data from June were the latest data used for predictions. Consequently, input data for a specific year were taken as beginning with July of the previous year and ending with June of the current year. For example, when developing a prediction for the year 1970, the model would have data for all four sea surface temperature indices from July 1969 to June 1970 available as inputs. The remaining two inputs for the ANN model consisted of the geographical coordinates of the center of the crop reporting district.

Each pattern that was presented to the ANN represented information relating to one of the seven crop reporting districts for one of the 55 available years. Thus a total of 385 patterns were available for the study. Each pattern had one target output value, which was the simulated yield for a crop reporting district calculated by averaging yield across all counties within that crop reporting district. The input values included were the prior values of four indices and the latitudinal and longitudinal coordinates of the crop reporting district. The values of the four sea surface temperature indices for each of the preceding 12 months provided a total of 48 inputs. The inclusion of the two inputs for the geographical location of the region provided 50 inputs per pattern. All values that were used either as inputs or as the target output were scaled to the range 0.1 to 0.9.

The 385 patterns were partitioned into three data sets consisting of the development set, the selection set, and the evaluation set. The development set was used to train the ANN models and determine the fitness of individuals for the GA, while the selection set was used to determine when training should be terminated to avoid overtraining and to select ANN parameters. The evaluation set was not used during model development and was held back for evaluation of the final model. Of the 56 available years within the entire dataset, 33 were partitioned into the development set (approximately 60 %), 14 were partitioned into the evaluation set (approximately 25 %), and 9 were partitioned into the selection set (approximately 15 %). With regards to ENSO, each year was designated with one of three statuses: El Niño, La Niña, and Neutral. Consequently, patterns were partitioned into separate data sets by year, so that each partitioned data set would retain the same proportions of El Niño, La Niña, and Neutral years as the entire dataset. The year-by-year partitioning is shown in Table 2.1.

B. Model Development

Model development consisted of three phases: the selection of ANN architecture and associated parameters, the GA search to determine the preferred inputs to the ANN model, and the development of a single final ANN model based on the inputs selected by the GA search. Prior to the use of the GA, multiple ANN models were developed and compared to determine the architecture and parameters which would minimize model error on the selection set. Once this model had been obtained, the GA search was initiated to determine the subset of available inputs which would minimize the mean absolute error (MAE) of the ANN model. This GA was a basic generational GA, using operators such as point crossover and bit flip mutation. The fitness measure used by the GA involved developing ANN models using the parameters chosen in the first phase of model development. These models were developed using a specific set of inputs, and their error values were used for the GA fitness scheme. Finally, these GA selected inputs were used to develop a number of ANN models. The model with the lowest error on the selection set was chosen as the final model.

Phase I: ANN Architecture Determination

The Java ANN model development software that was used in this study was based on the code developed by Smith et al. (2007). In addition to the Ward network architecture (Ward Systems Group, 1993) used in that study, the library was modified to include standard three-layer networks which used the logistic function for nodes in the hidden and output layers. The ANNs were trained using the error backpropagation (EBP) algorithm described in Haykin (1999). Under the EBP algorithm, models are instantiated by randomizing the network weights and training set order. A given model within this system describes an architecture based on a set of inputs and outputs. Accordingly, the various models that were developed in the process of

determining a final model differed according to which of the 50 available inputs were used in addition to their architecture and parameters. Therefore, the individual instantiations of any such model differed only in terms of random initial weights and the order in which patterns were presented to the network during the training process.

The ANN parameters and settings differentiating the potential models in this research were the number of hidden nodes and the ANN learning rate. Two network architectures were compared to determine which would be best suited for the final model. The 3-layer standard EBP neural network (Haykin, 1999) was tested against the Ward network architecture (Ward Systems Group, 1993). Ward networks involve a single hidden layer consisting of three slabs of nodes. The nodes within a slab use a particular activation function. Trials were conducted to compare these ANN architectures and selected ANN parameters. All 50 available inputs were included in this phase of the research.

Phase II: Input Selection

Once the preferred ANN architecture and parameters had been selected, they were used as part of a GA search to select the inputs for the model. This search was conducted to determine which subset of the available inputs would produce the minimum MAE. The fitness function for this GA involved developing three instantiations of an ANN model using a subset of the available inputs. The ANN parameters and settings determined during the previous phase of model development were used for the models developed in the fitness function. For the GA search, a general GA library was written in Python. This was interfaced with the Java ANN code using Jython, a library which allows the manipulation of Java objects and libraries from within Python. Each individual within the population was represented as a bit string, with each bit representing a potential climate input. This bit string represented the genome of a given

individual. Since four indices were used, each with 12 monthly values, this bit string was of length 48. A bit set to 1 in a particular position indicated that the corresponding input was used, and a 0 indicated that a particular input was not used in the model developed based on the genome of that individual. Hence, the number of 1's contained within the bit string would indicate the total number of active sea surface temperature inputs for a model based on a particular individual. The two geographical inputs were always used in model development.

The fitness function for the GA involved evaluating multiple instantiations of the ANN model specified by the individual's bit string genome. The tiered fitness scheme developed for this study involved performing two fitness evaluations: one for the selection process, and another for the termination condition. If an individual was being evaluated for the selection process, its error values on the development set would be used for its fitness value. If an individual was being evaluated for the GA convergence condition, its error values on the selection set would be considered. Similar to the way in which the selection set was used to determine when to stop the network training process to prevent the backpropagation algorithm from overfitting to the ANN development set, this novel tiered fitness scheme was implemented to prevent the GA search from overfitting to any particular data set. For a fitness evaluation of each individual within the GA population, three instantiations of the model were developed using the EBP algorithm. Three independent instantiations were developed in order to balance out the stochastic nature of the EBP algorithm. After training, each instantiation was applied to the ANN development set to determine its error. The lowest of these three error values was then assigned to the individual as its fitness value.

The selection scheme used for the GA search was binary tournament selection, which involves selecting two random individuals from the population and evaluating their fitness

values. The individual with the better fitness value (lower error) was then selected as a parent for mating and the process was repeated to select a second parent. The crossover operator used was two-point crossover. This process involves choosing two random points along the length of the bit string genome, and splitting the two parent individuals along these points to recombine their genetic material into a new individual. The mutation operator used was the traditional bit flip mutation, where any allele to be mutated has its value reversed, i.e. 0 becomes 1, 1 becomes 0. During the mating process, selected individuals had an 80% chance of undergoing crossover, and each bit in a newly created individual's bit string genome had a 10% chance of undergoing mutation.

Since each fitness evaluation involved developing multiple instantiations of the specified ANN model and each model had to be trained for a relatively long amount of time, fitness calculations were the most computationally expensive aspect of the GA search. As a result, the population size of the GA was limited to 100, similar to the process used by Guo and Uhrig (1992). In order to avoid the possibility of low diversity arising from such a small population size, population seeding was used. Population seeding is a process wherein specific individuals known to have relatively good fitness are inserted into a GA population in order to ensure that their genetic material is present for the algorithm to utilize (Julstrom 1994). In order to ensure that each of the potential inputs was active in some member of the initial population, five individuals with genomes consisting only of 1's (all inputs active) were inserted into the initial population. A generational GA (Holland, 1975) was used. In each generation, the individual with the best fitness according to the selection fitness function was evaluated on the selection set to test the termination condition. This termination condition caused the GA to halt if there had been

no new minimum in the selection set error during the past 10 generations, indicating that the GA had converged.

Once 20 GA runs had been completed, their selected inputs were compared. This involved compiling the bit string genome of the best individual from the final generation of each run, and averaging the values for each bit. This yielded a set of proportions, one for each of the 48 bits in the bit string, indicating how frequently that input was selected by the GA in its final solution. A threshold value was selected to transform these proportions back into a binary string indicating which sea surface temperature inputs the final model should use. Any input with a proportion below this threshold would not be used in the final model, while any input with a proportion greater than or equal to this threshold would be used.

Phase III: Final Model Evaluation

A number of different threshold values were tested to determine which produced the model with the lowest error on the development and selection sets. This selected threshold value was used to determine the inputs for developing the final ANN model. Models using these inputs were tested using several different learning rates to choose final model parameters. Fifty instantiations of this model were developed, and the instantiation with the lowest error on the selection set was chosen as the final model. This ANN model was then applied to the evaluation set to determine its accuracy.

Trials from Phase I of model development testing the different learning rates and number of hidden nodes were run on a dual-core Intel Core 2 Duo computer. Trials were run on Ward networks with one, three, and five hidden nodes per slab, as well as standard EBP networks with one, three, and five total hidden nodes. Learning rates of 0.1, 0.3, and 0.6 were tested. Networks were allowed to train for up to 1000 epochs, and required to train for a minimum of 500 epochs.

The development set and selection set error values were compared between the results from all parameter settings to determine the ANN model settings for the GA search fitness function and the final model.

For Phase II, to search for inputs on the ANN models specified by Phase I, 20 trials of the GA were run on an 8-core Xeon 2.0 GHz server. These trials were specified to terminate when the lowest selection set error in the population had not decreased for 10 generations. The final solutions with the lowest selection set error from each run were saved. The bit string genomes of all final solutions were compiled and averaged, yielding the proportion of inclusion for each input within these final solutions. These proportion values were then used to determine binary sets of outputs via a number of threshold values. Threshold values of 0.1 to 0.9 with a step size of 0.1 were used to develop model specifications. The different models based on these threshold values were then compared. Ten instantiations of each model were developed, and their error values on both the development and selection sets were compared.

In Phase III, these error values were used to select a threshold value to determine inputs for the final model. A flowchart illustrating this process is shown in Figure 2.1. These inputs were tested using learning rates of 0.1, 0.3, 0.6, and 0.9, with each parameter value evaluated using 10 instantiations. The setting which produced the lowest selection set error was chosen for the final model. Fifty instantiations of this model were then developed, and the instantiation with the lowest selection set error was chosen as the final model and tested on the independent evaluation set.

RESULTS AND DISCUSSION

Phase I

The ANN MAE for the development and selection data sets for each of the models included in the ANN architecture and parameter tests are shown in Table 2.2. Each entry in this table represents an average of ten ANN instantiations developed using the specified parameters. Due to the importance of the ability to generalize, parameters and architecture were selected based on selection set error values. Selection set MAE values for Ward network models ranged between 1558 and 1780, while selection set MAE values for EBP models ranged between 1572 and 1627. The ANN model with the lowest average MAE on the selection set was the Ward network with one hidden node per slab and a 0.6 learning rate, for an MAE of 1558 kg/ha. These same settings resulted in an MAE of 1302 kg/ha for the development set. This was not one of the lowest development set error values, but each of the settings which produced a lower development set error had a higher selection set error. Consequently, the Ward network architecture was used for all subsequent model development. These results are consistent with those of past studies (Smith et al. 2007; Shank et al. 2008a, 2008b) where Ward networks were found to be preferable to standard EBP networks.

Phase II

Twenty GA searches were run using the architecture and parameters chosen in the first phase for their fitness function. The GA searches ran for an average of 25 generations before terminating, though the longest two searches ran for over 50 generations.. The average fitness from each generation for all 20 runs is charted in Figure 2.2. The average population fitness for most runs was minimized to approximately the same level – between 1000 and 1100 kg/ha (with some runs dropping below 1000 kg/ha). Even the runs which continued for a larger number of

generations converged to this fitness range. This suggested that longer GA runs would not have produced a decrease in final solution fitness.

The proportion of inclusion for each of the 48 possible climate indices is shown in Table 2.3. These proportions are the result of averaging the final solutions of all 20 GA runs. Some index values were selected for the final solutions more frequently than others, with the NAO values for March and April and the PNA values for October and March being selected in 95% of runs. Thirteen of the 48 available inputs were selected by the GA in 70% or more of the searches run. Additionally, some index values were rarely active in the final solutions, with one value (the PNA value for November) not present in any of the final solutions. Ten out of 48 inputs were selected by the GA in 30% or less of the runs completed. The fact that the GA selection scheme displayed clear preferences for about half of the available index values shows that the choice of inputs is important to network performance for this application. Also of significance is the fact that the JMA and ONI were both selected most frequently during roughly the same period, from December to February. This is consistent with the fact that both of these indices measure the ENSO phenomenon. The fact that ONI values are taken from three month moving average windows suggests that January may have been the most influential of these three months, with its strength affecting the proportions for both December and February. However, this is not consistent with the results of similar research by Martinez et al. (2009), who found that ENSO values from July to September of the year prior to maize harvest were most correlated with yield values. None of the May or June inputs were selected in more than 60 % of GA runs, indicating that it may not be necessary to include inputs from the last two months in the search.

Threshold values considered for determining inputs to the ANN model ranged from 0.1 through 0.9 with a step size of 0.1. The development and selection set error values for models

developed using these inputs are shown in Table 2.4. A threshold value of 0.6 produced the lowest error on the selection set, with an MAE of 1346 kg/ha. Consequently, the inputs specified by this threshold value were used for the final model. These results demonstrated that the GA was successful in reducing the overall error for both the development and selection sets, as the MAE values for both data sets were lower than when trained using all inputs. Using all inputs, the best Phase I model achieved an MAE of 1558 kg/ha for the selection set, in comparison with the model using the GA-selected inputs, which had a selection set MAE of only 1345 kg/ha. The reduction in development set error when using GA-selected inputs was even greater: while the Phase I model had a development set MAE of 1302 kg/ha, the 0.6 threshold model average development set MAE was 927 kg/ha.

The inputs specified by the chosen threshold value of 0.6 are shown in Table 2.5. Interestingly, many of the inputs which Martinez et al. (2009) found to have the highest degree of correlation with maize yield were not included in these inputs. For instance, Martinez et al. (2009) identified July to September of the previous year as the time period for which the JMA index showed the strongest correlation, yet these inputs were not active in the final solution. Similarly, their study found that the PNA was most strongly correlated during the period of December to February, none of which were active in the inputs shown in Table 2.5. This may indicate that the GA search was identifying and exploiting different connections and correlations than those which Martinez et al.'s (2009) principal component analysis identified. These differences in results could also be due to the differences between the data used for this study and the data used by Martinez et al. (2009), such as the fact that this study is based on simulated yield values, whereas the research done by Martinez et al. was based on detrended, observed yield values.

Phase III

Models using the inputs specified by the 0.6 threshold value were tested using learning rates of 0.1, 0.3, 0.6, and 0.9. Ten trials were run for each learning rate. The mean MAE for each dataset from these trials are shown in Table 2.6, where it can be seen that the different learning rates appear to have had little effect on MAE. The setting which produced the lowest selection set error was a learning rate of 0.3, giving an average selection set error of 1365 kg/ha. The learning rate with the highest error for the selection set (0.6) had an MAE of 1389 kg/ha, only 24 kg/ha higher than the error produced by models using a 0.3 learning rate. Modifying the learning rate also did not result in a dramatic impact on model accuracy for the development set, with only 137 kg/ha separating the model with the best selection set MAE (0.3 learning rate, 941 kg/ha) from the model with the worst selection set MAE (0.9 learning rate, 1078 kg/ha). None of the error values for selection set from the learning rate trials differed by more than 50 kg/ha from the error of the initial 0.6 threshold model for the selection set (1346 kg/ha), indicating that changing the learning rate had little impact on model performance once model inputs had been specified by a threshold value.

Fifty instantiations of the 0.6 threshold model were then developed using a learning rate of 0.3. The best instantiation of the model based on this threshold value was chosen based on selection set error, and was tested on the final evaluation set. The evaluation set was presented to this instantiation of the final model only once, in feed forward mode only. Figure 2.3 shows a scatter plot for the development set that compares the final model prediction to the target simulated yield values. Since each target value is the average of the simulated yield for each

county in a particular district in a specific year, there is considerable variance in the data. This is depicted within the figure by the error bars, which show a range of +/- one standard deviation values resulting from averaging all of the simulated yield within each district. Across all three data sets, the average standard deviation value was 1959 kg/ha, reflecting a generally high level of variability within the data, caused by averaging the yield values from each of the counties within each crop reporting district. As can be seen from Figure 2.3, the final model was able to predict most values of the development set correctly within one standard deviation. Figure 2.4 is a plot of the final model performance on the selection set. Though the predictions for this data set were not as accurate as those for the development set, most predictions still came within one standard deviation of the simulated target value. This held true even for many predictions of the high and low extremes of simulated yield values. Figure 2.5 is a plot showing the final model performance on the evaluation set. As this figure shows, predictions were less accurate for the evaluation set in general than for the development or selection sets. However, many of the predictions still fell within one standard deviation of the simulated yield values. It is possible that this reduction in model accuracy on the evaluation set was due to overfitting during one or more of the stages of model development, despite the efforts to avoid overfitting through the GA's tiered fitness scheme.

The error values for the final model are shown in Table 2.7. The error of the final model for the development set was 792 kg/ha, while the error for the selection set was 1045 kg/ha. Both of these values were lower than those for the initial model using all 48 available climate inputs, which had a development set error of 1302 kg/ha and a selection set error of 558 kg/ha (Table 2.2). This demonstrates that the use of GA-selected inputs did indeed benefit the model, allowing it to predict simulated yield values with greater accuracy for both the development and selection

data sets. The MAE of the final model for the evaluation data set was higher than the MAE for the development and selection data sets at 1840 kg/ha. However, this value was still lower than the evaluation set average standard deviation of 1923 kg/ha, as well as the overall standard deviation value of 1959 kg/ha. As such, the MAE for the evaluation data set is comparable to the level of variability within the data.

SUMMARY AND CONCLUSIONS

ANNs were applied to sea surface temperature-based climate indices to predict maize yield in the southeastern U.S. A GA search was used to determine which inputs were necessary to develop a model with minimal error. This GA search used a tiered fitness scheme where different data sets were used for fitness evaluations, depending on whether these evaluations were for the purpose of selection or convergence detection. Through applying a threshold value to the inputs selected by multiple GA searches, a final set of inputs was determined in order to create a final ANN model. This model achieved an MAE of 792 kg/ha for the development set and an MAE of 1045 kg/ha for the selection set, both lower than the values resulting from models developed using all available inputs. For the evaluation data set, the final model achieved an MAE of 1840 kg/ha, compared with the data set average standard deviation of 1923 kg/ha.

The success of the tiered selection scheme in improving the ANN model performance on the network development and selection sets indicates that it may be a valuable tool for further research, and could be applied to GAs in other research areas. Additionally, the preferences expressed by the GA for certain inputs over others (as shown by the proportions of inclusion in the final solutions) shows that the GA is an effective tool for determining which inputs are most valuable to an ANN model. However, the MAE of the final model applied to evaluation set was not as low as for the development or selection sets, which could be due to the limited size of the

data set (with the development set consisting of only 231 patterns, and the selection set only 72), or possibly due to the selection set not being representative of the overall data. It is also possible that overfitting occurred during the model development process, a possibility which further research into this area could explore. Still, this error value was comparable with the overall variability within the data. Further research applying these techniques to different data could determine the generalizability of the approach outlined in this study. Also, the fact that none of the inputs from the last two months of data were selected in more than 60 % of GA runs suggests possible research into earlier prediction dates.

REFERENCES

Bose, N. K. and P. Liang, 1996. Neural network fundamentals with graphs, algorithms, and applications. In McGraw-Hill Series in Electrical and Computer Engineering, ed. S. W. Director. New York, NY: McGraw-Hill.

Center for Ocean-Atmospheric Prediction Studies. 2009. Monthly JMA Index. Florida State University. ftp://www.coaps.fsu.edu/pub/JMA_SST_Index/. Accessed on June 10, 2009.

D'Arrigo, R., and R. Wilson, 2008. El Niño and Indian Ocean influences on Indonesian drought: implications for forecasting rainfall and crop productivity. International Journal of Climatology 28(5): 611-616.

Dasgupta, D., and D. R. McGregor, 1992. Designing application-specific neural networks using the structured genetic algorithm. COGANN-92: International Workshop on Combinations of Genetic Algorithms and Neural Networks, Baltimore, MD: 87-96.

Enfield, D. B., 1996. Relationships of inter-American rainfall to tropical Atlantic and Pacific SST variability. Geophysical Research Letters 23(23): 3305-3308.

Everingham, Y. L., R. C. Muchow, R. C. Stone, and D. H. Coomans, 2003. Using southern oscillation index phases to forecast sugarcane yields: a case study for Northeastern Australia. International Journal of Climatology 23(10): 1211-1218.

Guo, Z., and R. E. Uhrig, 1992. Using genetic algorithms to select inputs for neural networks. COGANN-92: International Workshop on Combinations of Genetic Algorithms and Neural Networks, Baltimore, MD: 223-234.

Handler, P, 1990. USA corn yields, the El Niño and agricultural drought: 1867-1988. International Journal of Climatology 10(8): 819-828.

Hansen, J. W., A. W. Hodges, and J. W. Jones, 1998. ENSO Influences on agriculture in the southeastern United States. Journal of Climate 11(3): 404-411.

Hansen, J. W., J. W. Jones, C. F. Kiker, A. W. Hodges, 1999. El Niño-Southern Oscillation impacts on winter vegetable production in Florida. Journal of Climate 12(1): 92-102.

Haykin, S, 1999. Neural Networks: A Comprehensive Foundation (Second Edition). Upper Saddle River, NJ: Prentice Hall.

Henderson, C. E., W. D. Potter, R. W. McClendon, and G. Hoogenboom, 1998. Using a genetic algorithm to select parameters for a neural network that predicts aflatoxin contamination in peanuts. In Methodology and Tools in Knowledge-Based Systems, by Tim Hendtlass, et al., 460-469. Berlin: Springer.

Holland, J. H, 1975. Adaptation in Neural and Artificial Systems. Ann Arbor, MI: University of Michigan Press.

Hoogenboom, G., J.W. Jones, P.W. Wilkens, C.H. Porter, W.D. Batchelor, L.A. Hunt, K.J. Boote, U. Singh, O. Uryasev, W.T. Bowen, A.J. Gijsman, A. du Toit, J.W. White, and G.Y. Tsuji, 2004. Decision Support System for Agrotechnology Transfer Version 4.0 [CD-ROM]. Honolulu, HI: University of Hawaii.

Izaurralde, R. C., N. J. Rosenberg, R. A. Brown, D. M. Legler, M. T. Lopez, R. Srinivasan, 1999. Modeled effects of moderate and strong 'Los Niños' on crop productivity in North America. Agricultural and Forest Meteorology 94(3): 259-268.

Jain, A., R. W. McClendon, G. Hoogenboom, and R. Ramyaa, 2003. Prediction of frost for fruit protection using artificial neural networks. American Society of Agricultural Engineers, St. Joseph, MI, ASAE Paper 03-3075.

Jain, A., R. W. McClendon, and G. Hoogenboom, 2006. Freeze prediction for specific locations using artificial neural networks. Transactions of the ASABE 49(6): 1955-1962.

Jones, J. W., G. Hoogenboom, C. H. Porter, K. J. Boote, W. D. Batchelor, L. A Hunt, P. W. Wilkens, U. Singh, A. J. Gijsman, and J. T. Ritchie, 2003. The DSSAT cropping system model. European Journal of Agronomy 18(3): 235-265.

Julstrom, B. A, 1994. Seeding the population: improved performance in a genetic algorithm for the rectilinear Steiner problem. Proceedings of the 1994 ACM Symposium on Applied Computing. Phoenix, AZ: 222-226.

Leathers, D. J., B. Yarnal, M. A. Palecki, 1991. The Pacific/North American pattern and United States climate. Journal of Climate 4(5): 517-528.

Martinez, C. J., G. A. Baigorria, and J. W. Jones, 2009. Use of climate indices to predict corn yields in southeast USA. International Journal of Climatology 20(11): 1680-1691.

National Oceanic and Atmospheric Adminstration, 2009a. Monthly mean North Atlantic Oscillation index. NOAA Climate Prediction Center. http://www.cpc.ncep.noaa.gov/products/precip/CWlink/pna/nao.shtml. Accessed on June 10, 2009; Verified on November 16, 2009.

National Oceanic and Atmospheric Adminstration, 2009b. Monthly mean Pacific-North American Pattern index. NOAA Climate Prediction Center. http://www.cpc.ncep.noaa.gov/products/precip/CWlink/pna/pna.shtml. Accessed on June 10, 2009; Verified on November 16, 2009.

National Oceanic and Atmospheric Adminstration, 2009c. Monthly Oceanic Niño Index. NOAA Climate Prediction Center. http://www.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ensoyears.shtml. Accessed on June 10, 2009; Verified on November 16, 2009.

Persson, T., A. Garcia y Garcia, J. O. Paz, J. W. Jones, and G. Hoogenboom, 2009a. Net energy value of maize ethanol as a response to different climate and soil conditions in the southeastern USA. Biomass & Bioenergy 33(8): 1-10.

Persson, T., A. Garcia y Garcia, J. Paz, J. Jones, and G. Hoogenboom, 2009b. Maize ethanol feedstock production and net energy value as affected by climate variability and crop management practices. Agricultural Systems 100(1): 11-21.

Philander, S. G, 1990. El Niño, La Niña, and The Southern Oscillation. San Diego, CA: Academic Press.

Shank, D. B., G. Hoogenboom, and R. W. McClendon, 2008a. Dewpoint temperature prediction using artificial neural networks. Journal of Applied Meteorology & Climatology 47(6): 1757-1769.

Shank, D. B., R. W. McClendon, J. Paz, and G. Hoogenboom, 2008b. Ensemble artificial neural networks for prediction of dew point temperature. Applied Artificial Intelligence 22(6): 523-542.

Smith, B. A., R. W. McClendon, and G. Hoogenboom, 2007. Improving air temperature prediction with artificial neural networks. International Journal of Computational Intelligence 3(3): 179-186.

Smith. B. A., G. Hoogenboom, and R. W. McClendon, 2009. Artificial neural networks for automated year-round temperature prediction. Computers and Electronics in Agriculture 68(1): 52-61.

Travasso, M. I., G. O. Magrin, M. O. Grondona, and G. R. Rodriguez, 2009. The use of SST and SOI anomalies as indicators of crop yield variability. International Journal of Climatology 29: 23-29.

Trenberth, K. E, 1997. The definition of El Niño. Bulletin of the American Meteorological Society 78(12}: 2771-2777.

Ward Systems Group, 1993. Manual of Neuroshell 2. Frederick, MD.

Whitley, D, 1995. Genetic algorithms and neural networks. In Genetic Algorithms in Engineering and Computer Science, edited by J. Periaux and G. Winter, 191-201. John Wiley & Sons Ltd.

Yang, J., and V. Honavar, 1998. Feature subset selection using a genetic algorithm. IEEE Intelligent Systems 13(2): 44-49.

Table 2.1: Partitioning and El Niño Southern Oscillation (ENSO) status for each year.

Data Set	Neutral	El Niño	La Niña
Model Development	1951, 1961, 1963, 1969, 1980, 1982, 1985, 1986, 1991, 1993, 1994, 1995, 1996, 1997, 2002, 2005, 2006	1958, 1964, 1970, 1973, 1983, 1987, 1992, 2003	1956, 1957, 1965, 1971, 1972, 1975, 1976, 1989
Selection	1954, 1962, 1979, 1984, 2004	1966, 1998	1974, 1999
Evaluation	1953, 1959, 1960, 1967, 1978, 1981, 1990, 2001	1952, 1977, 1988	1955, 1968, 2000

Table 2.2: Mean Absolute Error for various artificial neural network models and data sets. Each entry is an average for 10 instantiations developed with these settings using all 50 inputs.

	Development set MAE (kg/ha)					
	Number of hidden nodes (EBP)			Number of hidden nodes per slab (Ward Net)		
Learning rate	1	3	5	1	3	5
0.1	1292	1349	1307	1273	1318	1288
0.3	1195	1205	1296	1188	1358	1210
0.6	1202	1204	1305	1302	1143	997

	Selection set MAE (kg/ha)					
	Number of hidden nodes (EBP)			Number of hidden nodes per slab (Ward Net)		
Learning rate	1	3	5	1	3	5
0.1	1623	1597	1626	1606	1719	1780
0.3	1575	1575	1627	1596	1650	1725
0.6	1572	1603	1578	1558	1657	1673

Table 2.3 – Proportions of inclusion for each month of each sea surface temperature index (Japan Meteorological Agency (JMA) index, Oceanic Niño Index (ONI), North Atlantic Oscillation (NAO) index, and the Pacific-North American (PNA) teleconnection index), averaged from 20 GA runs. Sea surface temperature data ranged from July of the year prior to harvest to June immediately preceding harvest.

	Proportion of inclusion											
Index	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May	Jun.
JMA	0.35	0.55	0.25	0.50	0.65	0.65	**0.75**	**0.85**	0.60	0.25	0.60	0.30
ONI	0.40	0.60	0.55	0.35	0.25	**0.75**	**0.75**	**0.85**	0.60	0.40	0.45	0.60
NAO	0.25	0.65	0.55	0.35	0.25	0.60	**0.80**	0.10	**0.95**	**0.95**	0.20	0.45
PNA	**0.70**	0.60	**0.90**	**0.95**	0.00	0.30	0.45	0.45	**0.95**	**0.85**	0.55	0.40

Table 2.4 – Mean absolute error for models developed using inputs specified by threshold values based on GA results.

Threshold value	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Development set MAE (kg/ha)	1154	1034	889	1020	1034	**927**	1089	1082	1257
Selection set MAE (kg/ha)	1561	1452	1400	1391	1416	**1346**	1417	1488	1510

Table 2.5 – The inputs chosen for the final model for each index (Japan Meteorological Agency (JMA) index, Oceanic Niño Index (ONI), North Atlantic Oscillation (NAO) index, and the Pacific-North American (PNA) teleconnection index), threshold value of 0.6.

Index	Jul.	Aug	Sep.	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May	Jun.
JMA	0	0	0	0	1	1	1	1	1	0	1	0
ONI	0	1	0	0	0	1	1	1	1	0	0	1
NAO	0	1	0	0	0	1	1	0	1	1	0	0
PNA	1	1	1	1	0	0	0	0	1	1	0	0

Table 2.6 – Mean Absolute Error of models for different learning rates, inputs specified by the 0.6 threshold.

Learning Rate	0.1	0.3	0.6	0.9
Development Set MAE (kg/ha)	1017	**941**	1001	1078
Selection Set MAE (kg/ha)	1369	**1365**	1389	1375

Table 2.7 - Mean Absolute Error for final model, selected based on selection set error from 50 separately trained instantiations.

MAE (kg/ha)		
Development	Selection	Evaluation
792	1045	1840

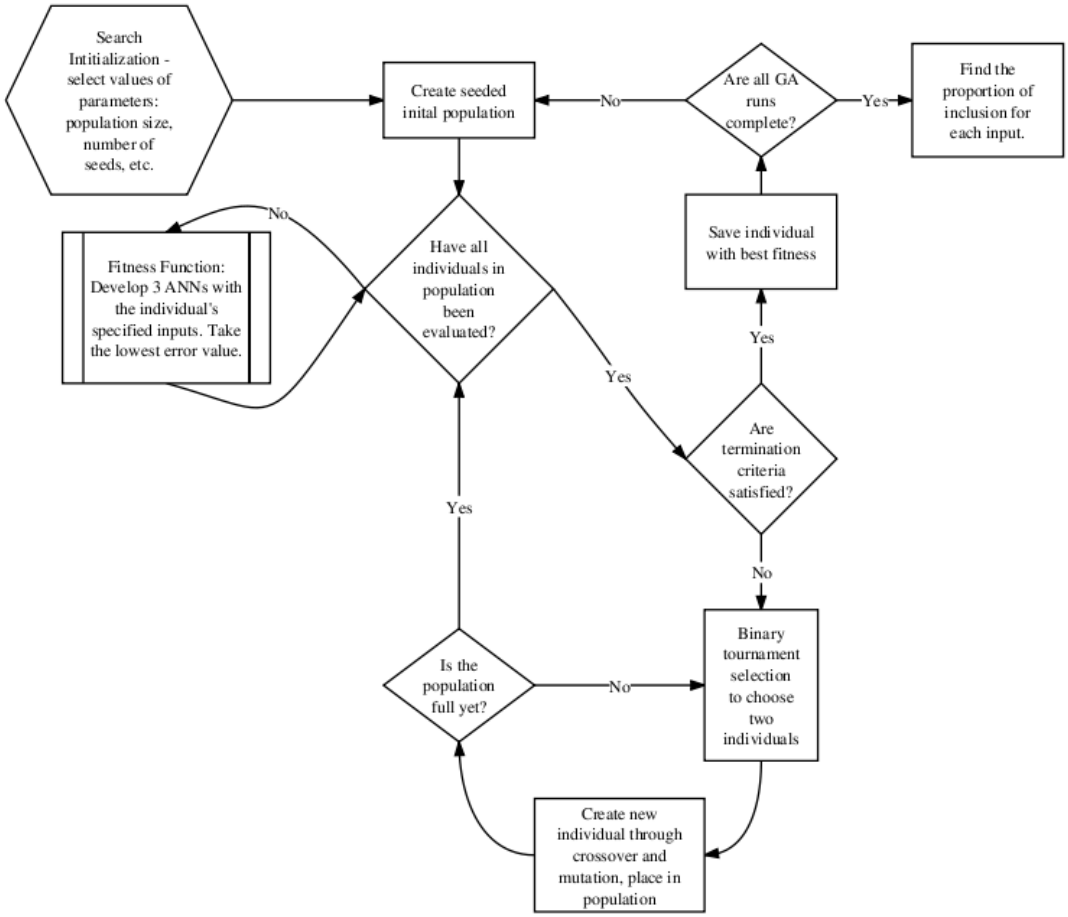

Figure 2.1 - Flowchart depicting the GA search process.

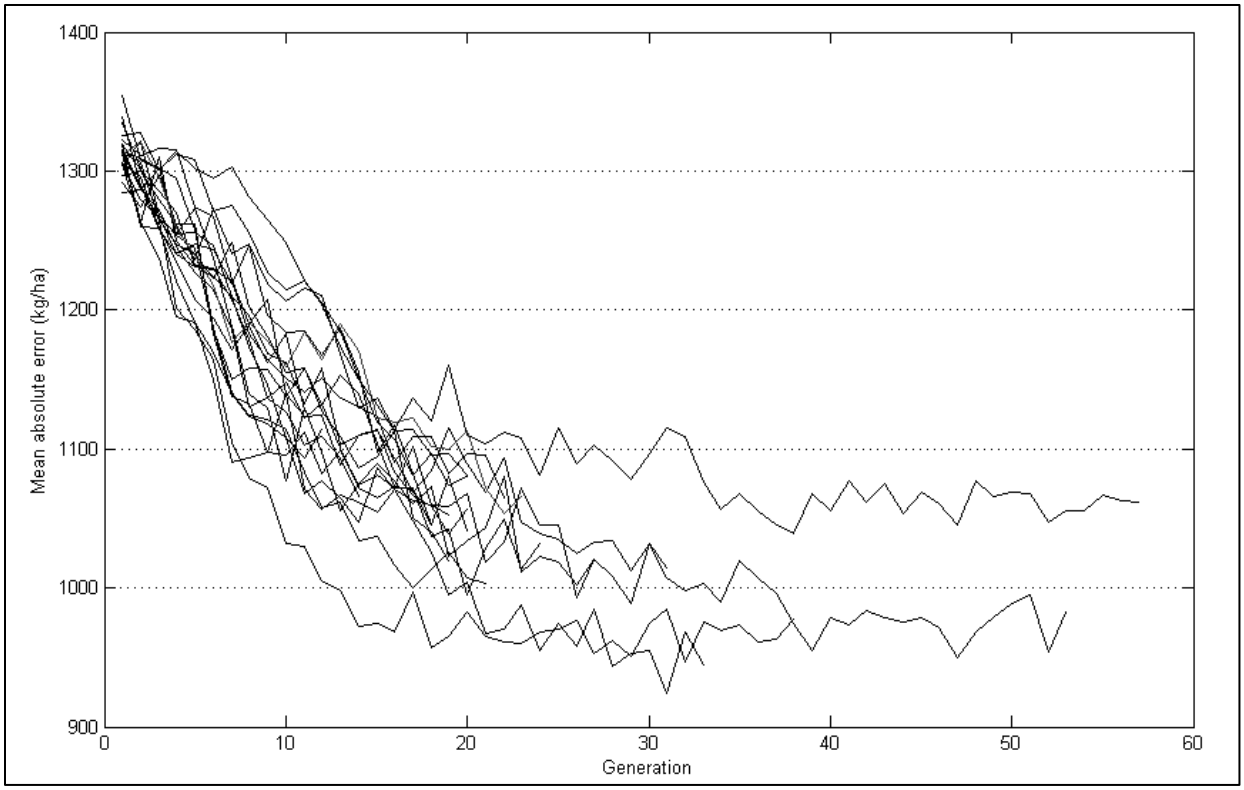

Figure 2.2 – Development set Mean Absolute Error for each generation from each of the 20 GA runs.

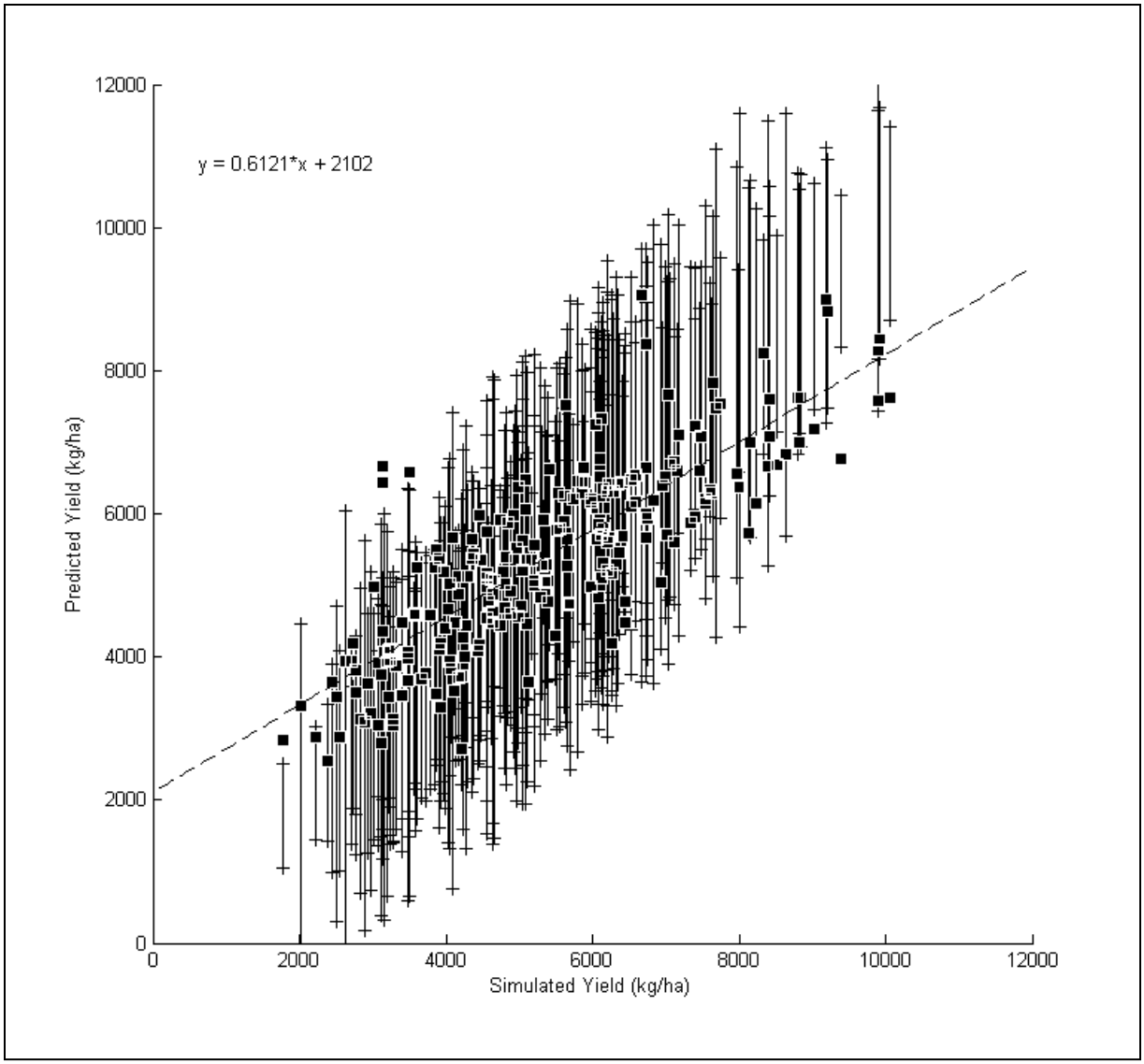

Figure 2.3 – Final ANN model yield predictions vs. target simulated yield for the development data set. Error bars show +/- one standard deviation for all counties in a district for a given year, average standard deviation of 2032 kg/ha.

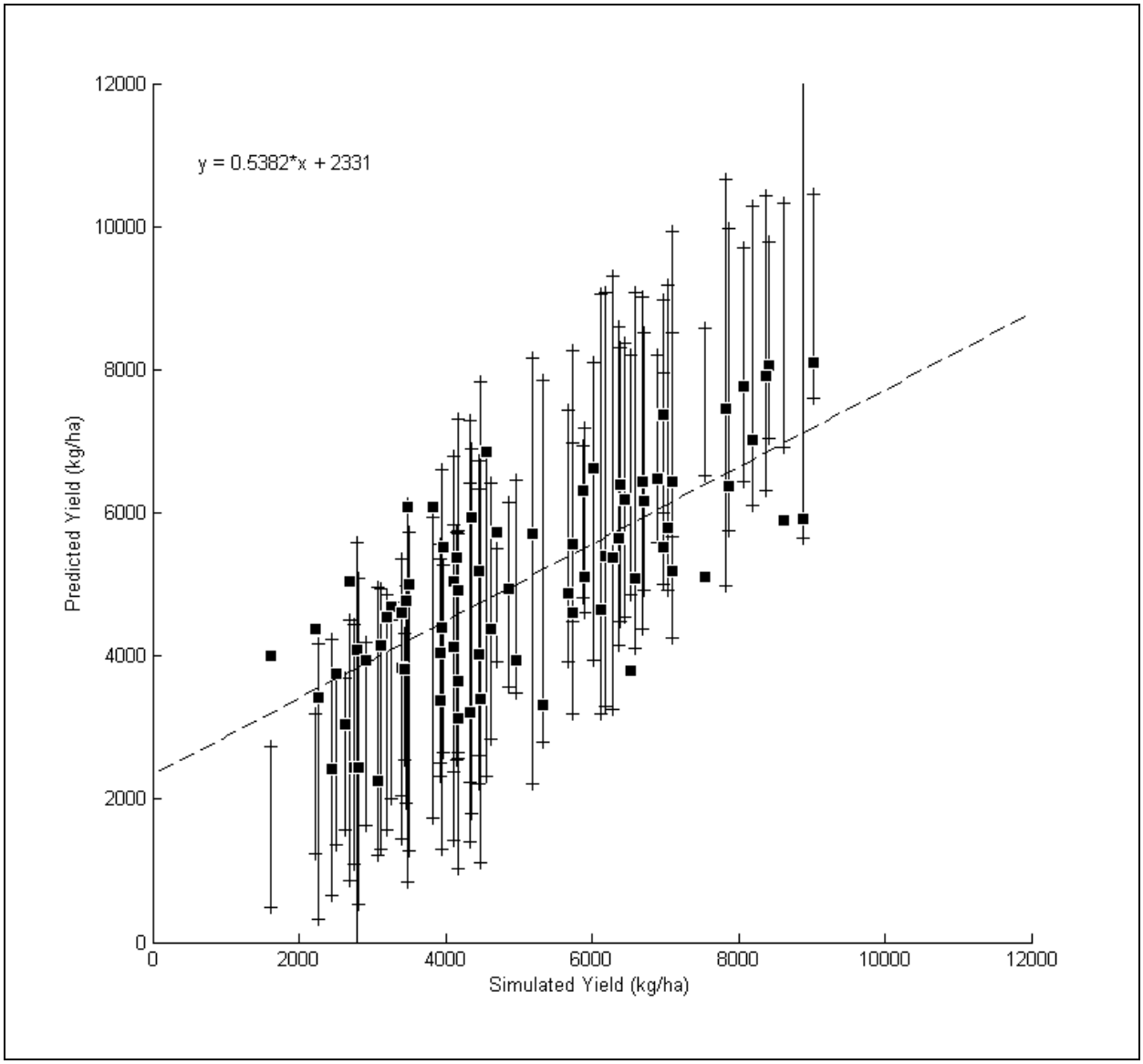

Figure 2.4 – Final ANN model yield predictions vs. target simulated yield for the selection data set. Error bars show +/- one standard deviation for all counties in a district for a given year, average standard deviation of 1920 kg/ha.

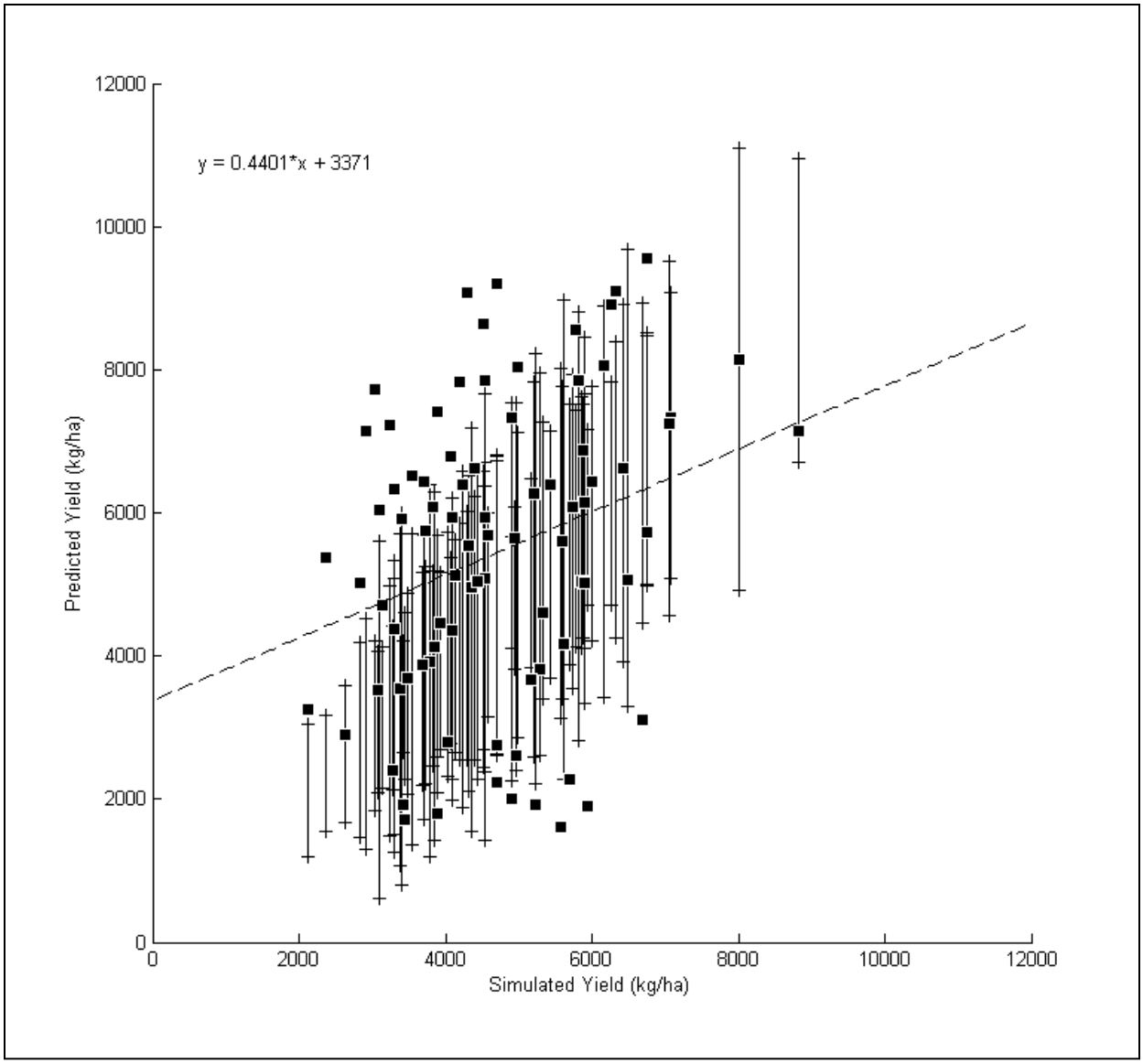

Figure 2.5 – Final ANN model yield predictions vs. target simulated yield for the evaluation data set. Error bars show +/- one standard deviation for all counties in a district for a given year, average standard deviation of 1923 kg/ha.

CHAPTER 3

THE EFFECTS OF VARYING PREDICTION DATE ON A MODEL FOR PREDICTING CROP YIELD BASED ON SEA SURFACE TEMPERATURE[2]

[2] Martin, C. M., G. Hoogenboom, R. W. McClendon, J. Paz, and T. Persson. To be submitted to *Computers and Electronics in Agriculture.*

ABSTRACT

Variations in climate can affect crop production, thus data describing such climate variations may be of value for predicting potential impact on crop yield. The goal of this research was to develop Artificial Neural Network (ANN) models to predict maize yield for the southeastern US using four monthly climate indices as inputs. Specific objectives included determining the preferred inputs to these ANN models using a Genetic Algorithm (GA) search, and determining how the accuracy of its model and the selection of its inputs were affected by moving the prediction date earlier in the year. The indices used for this research were based on three meteorological phenomena: the El Niño-Southern Oscillation, North Atlantic Oscillation, and Pacific-North American pattern. Weather data and crop conditions from seven crop reporting districts in Georgia, Alabama, and Florida were included in the study. Maize yield data were simulated using the Cropping System Model for the years 1950 to 2006. These yield values were then averaged across crop reporting districts, resulting in a data set with a high level of variability, as shown by the overall standard deviation of 1923 kg/ha. In order to determine which inputs could be used to most accurately predict crop yield, the space of potential model inputs was searched using a GA. In order to determine the effect of earlier prediction dates on model accuracy, genetic algorithm searches were conducted using prediction dates ranging from January to July for the year of the current growing season. The inputs selected by the genetic algorithm search for each prediction date were then compared. Moving the prediction date earlier in the year reduced the degree to which the GA search was able to minimize model error, as well as affecting the GA's selection of inputs, causing some inputs to be selected more or less frequently. The final model for the January prediction date had the lowest error for the independent evaluation set, with a mean absolute error of 1498 kg/ha, lower than the overall data

set's standard deviation value. The three earliest prediction dates achieved the lowest MAE values on the model development and evaluation data sets, indicating that accurate predictions could be made earlier in the growing season, a beneficial fact for farm managers. Further research is necessary to determine whether other types of computational models could successfully be applied to the same problem.

INTRODUCTION

Climate variability can have a major impact on agricultural production. For example, fluctuations in the amount of annual precipitation have been associated with the majority of crop failures within the United States (Ibarra and Hewitt, 1999). Climate variability, therefore, can pose a risk to the overall operation and economic return of the agricultural industry. However, the potential risks caused by meteorological variations can partially be minimized if farm were provided with climate forecasts (Baigorria et al., 2008). By utilizing crop yield predictions based on climate data, managers can modify their strategies to reduce the possible harmful effects of climate conditions, and take advantage of potentially advantageous circumstances (Martinez et al., 2009).

Large-scale meteorological phenomena, such as the El Niño-Southern Oscillation (ENSO), are measured in specific areas of the world, but are often linked to climate variation in other areas of the world through signals known as teleconnections (Izaurralde et al., 1999). Research has shown that regional climate conditions can be influenced by ENSO conditions (Philander, 1990; Trenberth, 1997). Past studies have focused on the impact of ENSO on crop yield within the southeastern US (Handler, 1990; Hansen et al., 1998, 1999; Baigorria et al., 2008). However, recent research has indicated that climate indices beyond ENSO may also have a strong impact on local climate variation (Travasso et al., 2008). Additional studies have

explored the influence of other climate phenomena such as the Pacific-North American pattern (Leathers et al., 1991) and indices such as tropical North Atlantic sea surface temperature (Enfield, 1996) on local meteorological conditions. In a recent study, Martinez et al. (2009) used data relating to multiple meteorological phenomena to model maize yield for the southeastern US. Their analysis used values of the tropical North Atlantic sea surface temperature, an index of the Pacific-North American pattern, the Bermuda high index, and the Japan Meteorological Agency's ENSO index. Following principal component analysis to summarize their data, Martinez et al. (2009) performed a linear correlation analysis. They then used the results of this analysis to create principal component regression models. These models were based both on lagged (prior to planting) and concurrent values of climate indices. One model used both lagged and concurrent values, and another model used only lagged indices. However, the predictive accuracy of these models was not evaluated for an independent evaluation data set. Other studies exploring the relation between meteorological phenomena and crop yield have used other statistical methods such as canonical correlation analysis (Baigorria et al., 2008), Monte Carlo procedures (Everingham et al., 2003), regression models (D'Arrigo and Wilson, 2008), and linear correlation (Travasso et al., 2008).

Artificial neural networks (ANNs) are computational tools that can be used to model complex relations in order to classify patterns and make predictions (Bose and Liang, 1996; Haykin, 1999). ANNs consist of a number of interconnected nodes whose connections have specific numerical weights. These weights define a mapping of the input nodes to output values. ANNs have been used for a number of meteorological and agricultural applications, including frost prediction (Jain et al. 2003; 2006), air temperature prediction (Smith et al. 2007; 2009), dewpoint temperature prediction (Shank et al. 2008a, 2008b), and prediction of aflotoxin

contamination in peanuts (Henderson et al. 1998). In some applications, ANNs are deployed in tandem with other computational tools such as the genetic algorithm (GA), which emulates the process of natural selection for search and optimization purposes (Holland, 1975). The GA can be used to select ANN architectures (Dasgupta and McGregor 1992), choose parameters for the ANN learning process (Henderson et al. 1998), train the networks directly (Whitley, 1995), or select appropriate and relevant inputs for the ANN from a larger set of possible inputs (Guo and Uhrig, 1992).

In Chapter 2, ANN models were developed to predict simulated maize yield using the values of sea surface temperature-based climatological indices. These ANN models were developed using values of monthly indices of the El Niño-Southern Oscillation, North Atlantic Oscillation, and Pacific-North American pattern as inputs. Data were used ranging through June of the year of the growing season, for a prediction date of July. A series of tests was conducted to determine the preferred ANN architecture and associated parameters. Subsequently, a GA was used to search all available inputs and determine the subset of inputs which would minimize the ANN model error when predicting crop yield. The final ANN model based on this GA-selected set of inputs was then applied to an independent evaluation set, resulting in a mean absolute error of 1840 kg/ha.

The goal of the research presented herein was to determine how the accuracy of an ANN model for predicting maize yield using indices of meteorological phenomena would be affected by moving its prediction date progressively earlier within the year. Objectives included: 1) to determine the preferred meteorological indices through the use of a GA search, 2) to determine the variation in index selection and model accuracy as a function of changing the prediction date

over time, and 3) to compare these results with previous research for predicting crop yield using teleconnections.

METHODOLOGY

A. Data

The data used for this study consisted of maize yield values used for model targets and climate index values used for model development. Simulated crop yield data were used for the target values rather than observed yield. This was done to avoid the influence of technological advancements and changes in crop management practices, such as the introduction of irrigation management, which occurred during the time period of the study. Data used in this study was taken from the years 1951-2006, as 1950 was the first year in which data for some indices were available. Maize yield was simulated with the Cropping System Model (CSM; Jones et al., 2003) of the Decision Support System for Agrotechnology Transfer (DSSAT; Hoogenboom et al., 2004), as described by Persson et al. (2009a, 2009b). These simulations were based on the maize cultivar Pioneer 31G98, which is a widely used maize hybrid throughout the southeastern US (Persson et al., 2009a, 2009b). For all input variables that were used for crop yield simulations, representative crop management practices and environmental conditions were used. Persson et al. (2009a, 2009b) simulated yield for three planting dates and various nitrogen fertilizer levels. However, in our study, only the middle planting date (the 76th day of the year, i.e., March 17 (March 16 during leap years)) was used. Additionally, only the simulated yield data based on the highest available level of fertilizer were used. All simulations in our study were also based on rainfed crop growth conditions, since the effect of irrigation could reduce the sensitivity of crops to climate variability, especially rainfall variability (Martinez et al. 2009). The simulations were conducted for each county in Georgia, Florida, and Alabama where corn is a major agronomic

crop. The simulations were run for each county using multiple soil types (Persson et al., 2009a, 2009b). However, only the simulations based upon the most common soil type in each county were considered for the study described herein.

Simulated yield values for 129 counties with significant maize production in Georgia, Alabama, and Florida were selected from the data generated by Persson et al. (2009a, 2009b). These counties were chosen to match the data used by Martinez et al. (2009), although their research was based on observed yield, not simulated values. The simulated yield data was then partitioned by crop reporting district. Those districts containing a small number of counties were dropped from the data set. The resulting data set contained simulated yield values for 94 counties within seven crop reporting districts, including two each in Georgia and Florida and three in Alabama. The simulated yield data were then averaged across all counties within each of the seven districts, resulting in one averaged annual yield value for each crop reporting district. These averages were used as the target values for the models in this study. This averaging introduced variance into the data, with an average standard deviation of 1959 kg/ha for all of the averaged yield values.

Values of four indices of climatic variability were used for model inputs: the Japan Meteorological Agency (JMA) index (Center for Ocean-Atmospheric Prediction Studies, 2009), the North Atlantic Oscillation (NAO) index (National Oceanic and Atmospheric Administration, 2009), the Pacific-North American (PNA) teleconnection pattern (National Oceanic and Atmospheric Administration, 2009), and the Oceanic Niño Index (ONI) (National Oceanic and Atmospheric Administration, 2009). Both the JMA and ONI are indicators of the ENSO phenomenon. Martinez et al. (2009) found the JMA to be the ENSO index that was most strongly correlated with maize yield. The PNA (Leathers et al. 1991) and NAO (Hurrell and Deser 2009)

are large-scale meteorological phenomena similar to ENSO. Each index consists of one value per month. Each ONI value represents a three-month moving average, e.g. the value for March would be a composite of data from February, March, and April. Maize is normally harvested during the summer in the southeastern US (Martinez et al. 2009). Thus, the latest data for each year came from June, with data for that year beginning in the previous July. The coordinates of the centroid of each crop reporting district were also included as model inputs.

The inputs and corresponding yield output were grouped into individual patterns. The target output value for each pattern was the average simulated yield value for that district. Each pattern had 50 potential inputs consisting of 12 monthly values of four indices, plus the latitude and longitude of the district centroid. These patterns were partitioned into three data sets: 60 % for model development, for use in ANN training, 15 % for model selection to prevent overfitting and to determine model settings, and 25 % for final evaluation of the model. The data were partitioned into these sets by year in order to retain the proportions of the overall data set with regards to ENSO status (El Niño, La Niña, or neutral). The results of the partitioning are shown in Table 3.1. The model development data set contained 33 years of data, the selection data set contained 9 years of data, and the evaluation data set contained 14 years of data.

B. Model Development

The ANN models of this study were developed as described by Martin et al. (2010). The models utilized the Ward network architecture (Ward Systems Group, 1993), as implemented by Smith et al. (2007). Like most ANNs, Ward networks are trained using the error backpropagation (EBP) algorithm (Haykin 1999). Networks trained using EBP are initialized with random weights, which are then modified based on the network error when predicting target values. Ward networks differ from more common EBP networks in that different nodes within the

network's hidden layer use different activation functions to transform their input, as opposed to all nodes in the hidden layer using the same activation function. Models developed for this research had 50 inputs available to them, and models differed from one another according to which of these inputs were included. Separate instantiations of a specific model could differ from one another due to the fact that EBP networks begin the training process with random initial weights.

A GA was utilized to search the space of possible inputs and to determine the subset of index values which would minimize the ANN model error. A GA is a population search method which selects individuals to be propagated to future generations based on a measure of fitness. For the purposes of this research, an individual within the GA's population consisted of a list of inputs specifying a particular model. An individual's fitness was determined by developing three instantiations of this ANN model, and averaging these instantiations' mean absolute error (MAE) on the development data set after training. Once two individuals were selected based on this fitness measure, their genetic information (the list of inputs included in model development) was combined, and the resulting individuals were placed in the population of the next generation. Once each generation was complete, the individual with the best fitness in the population, i.e., the lowest error for the development data set, was tested on the selection set to ensure that the GA was not overfitting to the development set. The MAE values for the selection set were saved for each generation, and when there had been no improvement in selection set error for 10 generations, the GA search was terminated. The individual with the best fitness value in the final population was saved for each run. The final solutions of multiple separate GA searches were compared in order to determine which inputs were important to model accuracy. Each input's

importance was determined by the frequency with which it was selected by the GA to be included in the final set of inputs.

After using the GA to search all available climate indices as described in Martin et al. (2010), the set of available inputs was reduced by moving the prediction date progressively earlier in the year. This was done to determine how the accuracy of the ANN model would be affected by making predictions at an earlier point in the year with less data and a longer prediction period. There were 48 climate index values available to the models, consisting of four monthly indices, from July of the prior year to June of the year of harvest. Initial GAs searched all 48 index values with a prediction point of July, i.e., as soon as the June climate data became available. Subsequent searches were run to limit this data and push back the date of prediction. Each time the prediction date was moved one month earlier, the number of inputs available to the ANN was reduced by four, i.e., one month of data for each climate index. Searches were run for prediction dates as early as January of the year of harvest. Consequently, seven prediction dates were tested. For each of these seven prediction dates, 20 GA searches were completed. Their final solutions were compared to determine the frequency with which each available input was selected by the GA. These proportions of inclusion were then compared across prediction dates to determine the effect on the GA's selection of inputs.

For each set of 20 GA searches relating to each prediction date, the proportions of inclusion specified by averaging all 20 solutions were used to determine a set of inputs for a final model. Threshold values were implemented to determine final model inputs. Only the inputs whose proportion of inclusion was higher than the threshold would be included in the final model. Threshold values of 0.2, 0.4, 0.6, and 0.8 were tested to determine which threshold value produced the model with the lowest error for the selection set. The specific subset of inputs used

by this model was then selected as the template for all subsequent models developed for that particular prediction date. Ten instantiations of this model were developed for each of the seven prediction dates. For each prediction date, the instantiation with the lowest error for the selection set was selected as the final model, which was then applied to an independent evaluation set to determine its predictive accuracy.

RESULTS AND DISCUSSION

For each prediction date, 20 GA runs were conducted. The average development set error of the population averaged across all 20 runs for each of the seven prediction dates can be seen in Figure 3.1. For the latest three prediction dates, i.e., July, June, and May, the GA minimized the average development set error for each run at comparable rates. For each of these three prediction dates, the average development set error in the final population after 20 generations was near 1050 kg/ha. These average development set error values from 20 generations into the GA searches can be seen in Table 3.2. Moving the prediction date back to April increased the development set MAE at 20 generations to 1119 kg/ha, and a prediction date of March resulted in a similar population error of 1117 kg/ha. Moving the prediction date further back to February and January increased the average development set MAE of the population further to 1165 kg/ha for February and 1193 kg/ha for January. Overall, moving the prediction date earlier so that the GA had less data available to it for model development caused the average error of the models in the population to increase. However, there was still a substantial reduction in average population development set error for GA runs using a January prediction date. Figure 3.2 shows the average population development set error for each of the 20 GA runs using a July prediction date. This figure shows that the duration of the individual GA runs varied considerably, with some runs continuing past 50 generations. The final average development set error of these runs varied

considerably as well, with some runs having a final error value below 1000 kg/ha, and other runs terminating closer to a final error of 1100 kg/ha. By contrast, Figure 3.3 shows the individual GA runs for a prediction date of January. There was considerably less variation within these runs, especially in terms of their final error. Almost all of the 20 runs shown in the figure had a final average development set error of approximately 1200 kg/ha.

For each of the prediction dates, the final solutions from the 20 GA runs were averaged to determine the frequency with which each of the climate inputs was selected by the GA. Table 3.3 shows these frequency values for the 12 months of the JMA index. Completely random selection of an input would result in a frequency value of 0.50. Thus, a frequency value higher than 0.50 indicates that the GA found the inclusion of that input to be beneficial to the model development process. Index values which were selected with a frequency greater than or equal to 0.70 are shown in bold in the tables displaying these results. For the JMA index, the prediction date seems to have had little impact on the GA's preference for individual months, with most months being selected with similar frequency across prediction dates. In their attempt to predict maize yield using climate indices, Martinez et al. (2009) indicated that values of the JMA index from July to September of the year before harvest showed maximum correlation to crop yield. However, the results presented in Table 3.3 show that the GA runs conducted for this study did not show a preference for these values. The only months of the JMA index selected by the GA consistently when they were available were January and February of the year of harvest, as well as November of the year prior to harvest.

The proportion of inclusion for values of the NAO index is shown in Table 3.4. Moving the prediction date earlier in the year affected the frequency with which the GA selected particular months. For instance, the value for the February NAO input was only selected 10 % of

the time in GA runs using the July prediction date, but was active 65 % of the time in GA runs for a March prediction date. For this earlier prediction date, the value for February was the latest one available. It is possible that this information was not useful in reducing model error when more information was available to the GA, but was selected more frequently when this information was removed. Conversely, the value for November NAO was selected 25 % of the time for the prediction dates of July, June, and May, but only 5 % of the time for earlier prediction dates of March, February, and January. This could indicate that the presence of extraneous or irrelevant information produced greater error in models with less available inputs, and consequently was selected less frequently. Martinez et al. (2009) found that the north Atlantic sea surface temperature from March to May was most highly correlated with crop yield of the current growing season. It can be seen in Table 3.4 that the GA runs did commonly select data from March and April for its models when such data was available. The GA showed a strong preference for April values in particular, never selecting them in less than 95 % of final solutions.

The frequency with which the GA runs selected inputs from the PNA index is shown in Table 3.5. The effects of moving the prediction date earlier are most pronounced for the December data, which increased from being selected in 30 % of GA runs with a July prediction date to being selected in 80 % of GA runs with a January prediction date. Martinez et al. (2009) identified December to February as the period during which PNA values showed the greatest correlation with crop yield. However, the months for which the GA most commonly chose PNA inputs to be active in this study were September and October of the year prior to the maize growing season, along with March and April from the year of the growing season.

The proportion of GA runs for which inputs of the ONI index were selected is shown in Table 3.6. The ONI, like the JMA, is a measure of the ENSO phenomenon, which Martinez et al. (2009) found to be most highly correlated with crop yield from July to September of the year prior to the maize growing season. Although the GA selected the July ONI input to be active in more than 50% of runs for all prediction dates other than July, it did not select the August or September inputs in more than 50 % of cases. In general, the GA did not show a strong preference either positive or negative for any of the inputs from August to November of the year prior to harvest. Instead, the months most frequently selected by the GA were January and February of the year of the growing season. ONI values for most months had proportions of inclusion close to 0.50, with 50 of the 63 available inputs being selected in between 35 and 65 % of runs. This high amount of proportions close to 0.50 indicates that the inclusion or exclusion of these inputs was more random than for the other indices tested.

Threshold values of 0.2, 0.4, 0.6, and 0.8 were applied to these proportions of inclusion to determine model inputs, with ten model instantiations developed for each threshold value. The results of applying these models to the development set for each of the prediction dates is shown in Table 3.7. There was no clear trend in error according to threshold value, with all but one of the averaged errors having a value between 900 and 1100 kg/ha. The threshold values to determine the inputs for the final models were selected based on the average selection set error (Table 3.8). Accordingly, the final models for the prediction dates of July, June, and April used a threshold value of 0.6, the prediction dates of March and January used a threshold value of 0.4, and the prediction dates of May and February used a threshold value of 0.2.

Out of the ten instantiations developed based on these threshold values, a single final model was selected for each prediction date based on the MAE of the selection set. The MAE

values for each of these final models when applied to the development set, selection set, and final evaluation set are shown in Table 3.9. The prediction date whose final model produced the lowest error for the development set was February, with an MAE of 755 kg/ha. The final model for the March prediction date resulted in the lowest MAE value for the selection set (1019 kg/ha). The lowest error for the evaluation set was achieved by the final model for the January prediction date (1498 kg/ha). In each case the final error for the evaluation set was lower than the overall standard deviation value of the average yield (1959 kg/ha), as in the case of the final model developed in Chapter 2 of this thesis. Interestingly, the models which performed best for all three data sets were those for January, February, and March, despite the fact that these were the three earliest prediction dates. Also important is the fact that the second lowest evaluation set error was achieved by the final model for the February prediction date. This indicates that models based on the earliest prediction dates had the lowest MAE on the evaluation set, even though these models had the least data available to them for model development.

Figure 3.4 contains scatterplots showing the performance of the final model for each of the seven prediction dates as they were applied to all three data sets. Each plot also contains error bars showing the standard deviation resulting from the averaging of the simulated yield values within each crop reporting district for each year of data. As can be seen from these plots, the final models predicted most simulated yield values for the development and selection data sets within one standard deviation. While the evaluation set predictions were generally less accurate, they were still often within one standard deviation of the simulated yield values. The plots show no major differences between the later prediction dates of July (Figure 3.4 a, b, c), June (Figure 3.4 d, e, f), May (Figure 3.4 g, h, i), April (Figure 3.4 j, k, l) or March (Figure 3.4 m, n, o), but the plots for the earliest prediction dates of February (Figure 3.4 p, q, r) and January (Figure 3.4

s, t, u) show these models' improved accuracy, particularly on the evaluation set. These plots confirm that these two models had the most accurate prediction dates for the evaluation set, as indicated by their error values. In comparison with the performance of other final models on the evaluation set, the February and January models had less of a tendency to overpredict.

SUMMARY AND CONCLUSIONS

GA searches were conducted to explore the space of available inputs for an ANN model for predicting maize yield based on climate indices. Trials were run for seven different prediction dates to determine how model accuracy would be impacted by predicting crop yield earlier in the year. The results of these GA searches were compared to determine how modifying the prediction date affected model accuracy and which inputs were selected by the GA most frequently. Although modifying the prediction date had no consistent effect on which climate data was selected by the GA across the four climate indices tested, it did affect the GA's overall ability to minimize the error of the models it developed. For each prediction date, a final model was developed and tested with an independent evaluation set. In each case, the MAE of the evaluation set was comparable to the level of variation within the data. Additionally, the models with the lowest error for the independent evaluation set were those using the earliest prediction dates, i.e., January and February. This increases the potential usefulness of such models for farm managers, since these models are able to predict yield earlier in the year when more action can be taken to modify and adapt crop management based on model predictions. Further research into this area using other computational techniques could shed more light on the effects of earlier prediction dates on both input selection and model error. The potential of combining non-ANN modeling techniques with the GA search also has yet to be explored. Research into either area could reveal more about the relationships between large-scale climate indices and crop yield.

REFERENCES

Baigorria, G. A., J. W. Hansen, N. Ward, J. W. Jones, and J. J. O'Brien, 2008. Assessing predictability of cotton yields in the southeastern United States based on regional atmospheric circulation and surface temperatures. Journal of Applied Meteorology and Climatology 47: 76-91.

Bose, N. K. and P. Liang, 1996. Neural network fundamentals with graphs, algorithms, and applications. In McGraw-Hill Series in Electrical and Computer Engineering, ed. S. W. Director, New York, NY: McGraw-Hill.

Center for Ocean-Atmospheric Prediction Studies, 2009. Monthly JMA Index. Florida State University. ftp://www.coaps.fsu.edu/pub/JMA_SST_Index/. Accessed on June 10, 2009.

D'Arrigo, R., and R. Wilson, 2008. El Niño and Indian Ocean influences on Indonesian drought: implications for forecasting rainfall and crop productivity. International Journal of Climatology 28(5): 611-616.

Dasgupta, D., and D. R. McGregor, 1992. Designing application-specific neural networks using the structured genetic algorithm. COGANN-92: International Workshop on Combinations of Genetic Algorithms and Neural Networks, Baltimore, MD: 87-96.

Enfield, D. B., 1996. Relationships of inter-American rainfall to tropical Atlantic and Pacific SST variability. Geophysical Research Letters 23(23): 3305-3308.

Everingham, Y. L., R. C. Muchow, R. C. Stone, and D. H. Coomans, 2003. Using southern oscillation index phases to forecast sugarcane yields: a case study for Northeastern Australia. International Journal of Climatology 23(10): 1211-1218.

Guo, Z., and R. E. Uhrig, 1992. Using genetic algorithms to select inputs for neural networks. COGANN-92: International Workshop on Combinations of Genetic Algorithms and Neural Networks, Baltimore, MD: 223-234.

Handler, P, 1990. USA corn yields, the El Niño and agricultural drought: 1867-1988. International Journal of Climatology 10(8): 819-828.

Hansen, J. W., A. W. Hodges, and J. W. Jones, 1998. ENSO Influences on agriculture in the southeastern United States. Journal of Climate 11(3): 404-411.

Hansen, J. W., J. W. Jones, C. F. Kiker, A. W. Hodges, 1999. El Niño-Southern Oscillation impacts on winter vegetable production in Florida. Journal of Climate 12(1): 92-102.

Haykin, S, 1999. Neural Networks: A Comprehensive Foundation (Second Edition). Upper Saddle River, NJ: Prentice Hall.

Henderson, C. E., W. D. Potter, R. W. McClendon, and G. Hoogenboom, 1998. Using a genetic algorithm to select parameters for a neural network that predicts aflatoxin contamination

in peanuts. In Methodology and Tools in Knowledge-Based Systems, by Tim Hendtlass, et al., 460-469. Berlin: Springer.

Holland, J. H, 1975. Adaptation in Neural and Artificial Systems. Ann Arbor, MI: University of Michigan Press.

Hoogenboom, G., J.W. Jones, P.W. Wilkens, C.H. Porter, W.D. Batchelor, L.A. Hunt, K.J. Boote, U. Singh, O. Uryasev, W.T. Bowen, A.J. Gijsman, A. du Toit, J.W. White, and G.Y. Tsuji, 2004. Decision Support System for Agrotechnology Transfer Version 4.0 [CD-ROM]. Honolulu, HI: University of Hawaii.

Hurrell, J.W. and C. Deser, 2009. North Atlantic climate variability: The role of the North Atlantic Oscillation. Journal of Marine Systems 78(1): 28-41.

Ibarra, R., and T. Hewitt, 1999. Utilizing crop insurance to reduce production risk. Institute of Food and Agricultural Sciences FE-198, Florida Cooperative Extension Service.

Izaurralde, R. C., N. J. Rosenberg, R. A. Brown, D. M. Legler, M. T. Lopez, R. Srinivasan, 1999. Modeled effects of moderate and strong 'Los Niños' on crop productivity in North America. Agricultural and Forest Meteorology 94(3): 259-268.

Jain, A., R. W. McClendon, G. Hoogenboom, and R. Ramyaa, 2003. Prediction of frost for fruit protection using artificial neural networks. American Society of Agricultural Engineers, St. Joseph, MI, ASAE Paper 03-3075.

Jain, A., R. W. McClendon, and G. Hoogenboom, 2006. Freeze prediction for specific locations using artificial neural networks. Transactions of the ASABE 49(6): 1955-1962.

Jones, J. W., G. Hoogenboom, C. H. Porter, K. J. Boote, W. D. Batchelor, L. A Hunt, P. W. Wilkens, U. Singh, A. J. Gijsman, and J. T. Ritchie, 2003. The DSSAT cropping system model. European Journal of Agronomy 18(3): 235-265.

Leathers, D. J., B. Yarnal, M. A. Palecki, 1991. The Pacific/North American pattern and United States climate. Journal of Climate 4(5): 517-528.

Martinez, C. J., G. A. Baigorria, and J. W. Jones, 2009. Use of climate indices to predict corn yields in southeast USA. International Journal of Climatology 20(11): 1680-1691.

Martin, C. M., R. W. McClendon, J. Paz, and G. Hoogenboom, 2010. A genetic algorithm & neural network hybrid for predicting crop yields based on sea surface temperatures. Expert Systems With Applications (Submitted for publication).

National Oceanic and Atmospheric Adminstration. 2009a. Monthly mean North Atlantic Oscillation index. NOAA Climate Prediction Center. http://www.cpc.ncep.noaa.gov/products/precip/CWlink/pna/nao.shtml. Accessed on June 10, 2009; Verified on November 16, 2009.

National Oceanic and Atmospheric Adminstration. 2009b. Monthly mean Pacific-North American Pattern index. NOAA Climate Prediction Center. http://www.cpc.ncep.noaa.gov/products/precip/CWlink/pna/pna.shtml. Accessed on June 10, 2009; Verified on November 16, 2009.

National Oceanic and Atmospheric Adminstration. 2009c. Monthly Oceanic Niño Index. NOAA Climate Prediction Center. http://www.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ensoyears.shtml. Accessed on June 10, 2009; Verified on November 16, 2009.

Persson, T., A. Garcia y Garcia, J. O. Paz, J. W. Jones, and G. Hoogenboom, 2009a. Net energy value of maize ethanol as a response to different climate and soil conditions in the southeastern USA. Biomass & Bioenergy 33(8): 1-10.

Persson, T., A. Garcia y Garcia, J. Paz, J. Jones, and G. Hoogenboom, 2009b. Maize ethanol feedstock production and net energy value as affected by climate variability and crop management practices. Agricultural Systems 100(1): 11-21.

Philander, S. G, 1990. El Niño, La Niña, and The Southern Oscillation. San Diego, CA: Academic Press.

Shank, D. B., G. Hoogenboom, and R. W. McClendon, 2008a. Dewpoint temperature prediction using artificial neural networks. Journal of Applied Meteorology & Climatology 47(6): 1757-1769.

Shank, D. B., R. W. McClendon, J. Paz, and G. Hoogenboom, 2008b. Ensemble artificial neural networks for prediction of dew point temperature. Applied Artificial Intelligence 22(6): 523-542.

Smith, B. A., R. W. McClendon, and G. Hoogenboom, 2007. Improving air temperature prediction with artificial neural networks. International Journal of Computational Intelligence 3(3): 179-186.

Smith. B. A., G. Hoogenboom, and R. W. McClendon, 2009. Artificial neural networks for automated year-round temperature prediction. Computers and Electronics in Agriculture 68(1): 52-61.

Travasso, M. I., G. O. Magrin, M. O. Grondona, and G. R. Rodriguez, 2009. The use of SST and SOI anomalies as indicators of crop yield variability. International Journal of Climatology 29: 23-29.

Trenberth, K. E, 1997. The definition of El Niño. Bulletin of the American Meteorological Society 78(12}: 2771-2777.

Ward Systems Group, 1993. Manual of Neuroshell 2. Frederick, MD.

Whitley, D, 1995. Genetic algorithms and neural networks. In Genetic Algorithms in Engineering and Computer Science, edited by J. Periaux and G. Winter, 191-201. John Wiley & Sons Ltd.

Table 3.1: Partitioning and corresponding ENSO status.

Data Set	Neutral	El Nino	La Nina
Model Development	1951, 1961, 1963, 1969, 1980, 1982, 1985, 1986, 1991, 1993, 1994, 1995, 1996, 1997, 2002, 2005, 2006	1958, 1964, 1970, 1973, 1983, 1987, 1992, 2003	1956, 1957, 1965, 1971, 1972, 1975, 1976, 1989
Selection	1954, 1962, 1979, 1984, 2004	1966, 1998	1974, 1999
Evaluation	1953, 1959, 1960, 1967, 1978, 1981, 1990, 2001	1952, 1977, 1988	1955, 1968, 2000

Table 3.2: Development set Mean Absolute Error after 20 generations averaged across 20 GA runs for seven prediction dates.

Prediction Date	Development Set MAE (kg/ha)
July	1054
June	1029
May	1039
April	1119
March	1117
February	1165
January	1193

Table 3.3: Proportions of inclusion for each month of the Japan Meteorological Agency (JMA) index, seven prediction dates. Inputs with a proportion greater than or equal to 0.70 in bold.

Prediction Date	Month											
	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May	Jun.
July	0.35	0.55	0.25	0.50	0.65	0.65	**0.75**	**0.85**	0.60	0.25	0.60	0.30
June	0.60	0.50	0.40	0.55	**0.70**	0.35	0.50	**0.95**	0.55	0.10	0.40	0
May	0.45	0.50	0.50	0.40	**0.80**	0.40	0.45	**0.90**	**0.75**	0.30	0	0
April	0.35	0.35	0.40	0.25	0.60	0.60	**0.85**	**0.90**	0.55	0	0	0
March	0.25	0.30	0.35	0.40	0.20	0.55	**0.85**	**1.00**	0	0	0	0
February	**0.80**	0.20	0.40	0.50	**0.85**	0.55	0.45	0	0	0	0	0
January	0.50	0.60	0.45	0.40	**0.90**	0.65	0	0	0	0	0	0

Table 3.4: Proportions of inclusion for each month of the North Atlantic Oscillation (NAO) index, seven prediction dates. Inputs with a proportion greater than or equal to 0.70 in bold.

	Month											
Prediction Date	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May	Jun.
July	0.25	0.65	0.55	0.35	0.25	0.60	**0.80**	0.10	**0.95**	**0.95**	0.20	0.45
June	0.65	0.65	0.55	0.25	0.25	**0.85**	**0.75**	0.05	**0.75**	**0.95**	0.45	0
May	0.55	**0.70**	0.50	0.30	0.25	0.60	0.60	0.40	0.60	**1.00**	0	0
April	0.30	0.35	0.55	0.20	0.15	**0.75**	0.65	0.10	0.65	0	0	0
March	0.35	0.45	0.35	0.15	0.05	0.45	0.45	0.65	0	0	0	0
February	0.45	0.55	**0.95**	0.20	0.05	0.60	**1.00**	0	0	0	0	0
January	0.50	0.60	**0.75**	0.25	0.05	0.55	0	0	0	0	0	0

Table 3.5: Proportions of inclusion for each month of the Pacific-North American (PNA) index, seven prediction dates. Inputs with a proportion greater than or equal to 0.70 in bold.

Prediction Date	Month											
	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May	Jun.
July	**0.70**	0.60	**0.90**	**0.95**	0.00	0.30	0.45	0.45	**0.95**	**0.85**	0.55	0.40
June	0.65	0.50	**0.85**	**1.00**	0.15	0.50	0.30	0.50	**1.00**	**1.00**	0.55	0
May	**0.75**	0.60	**0.95**	**0.90**	0.00	0.60	0.50	0.45	**1.00**	**1.00**	0	0
April	**0.80**	0.45	**0.80**	**0.95**	0.05	0.60	0.45	0.50	**0.80**	0	0	0
March	0.50	0.60	**0.70**	0.60	0.10	0.40	0.60	0.50	0	0	0	0
February	**0.85**	0.50	0.55	**1.00**	0.35	**0.85**	0.30	0	0	0	0	0
January	0.55	**0.70**	0.60	**0.90**	0.10	**0.80**	0	0	0	0	0	0

Table 3.6: Proportions of inclusion for each month of the Oceanic Niño Index (ONI), seven prediction dates. Inputs with a proportion greater than or equal to 0.70 in bold.

	Month											
Prediction Date	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May	Jun.
July	0.40	0.60	0.55	0.35	0.25	**0.75**	**0.75**	**0.85**	0.60	0.40	0.45	0.60
June	0.60	0.35	0.25	0.50	0.45	0.25	0.60	0.40	0.45	0.35	0.55	0
May	0.60	0.40	0.35	0.40	0.50	0.60	0.65	0.65	0.40	0.55	0	0
April	0.60	0.45	0.60	0.40	0.40	**0.70**	0.65	0.55	0.45	0	0	0
March	**0.75**	0.55	0.45	0.50	0.35	0.60	**0.70**	**0.70**	0	0	0	0
February	0.60	0.55	0.50	0.50	0.35	0.50	**0.75**	0	0	0	0	0
January	0.65	0.55	0.45	0.55	0.35	**0.95**	0	0	0	0	0	0

Table 3.7: Development set mean absolute error for models developed using 4 threshold values based on the GA results for all seven prediction dates. Each value is an average from 10 instantiations of the specified model.

	Mean Absolute Error (kg/ha)			
	Threshold Value			
Prediction date	0.2	0.4	0.6	0.8
July	1034	1020	**927**	1082
June	**961**	1000	994	999
May	995	**942**	981	1020
April	**867**	969	948	1020
March	961	941	**908**	950
February	**937**	1025	981	1085
January	1043	**946**	1016	987

Table 3.8: Selection set mean absolute error for models developed using 4 threshold values based on the GA results for all seven prediction dates. Each value is an average from 10 instantiations of the specified model.

	Mean Absolute Error (kg/ha)			
	Threshold Value			
Prediction date	0.2	0.4	0.6	0.8
July	1452	1391	**1346**	1488
June	1363	1405	**1362**	1364
May	**1343**	1345	1366	1373
April	1285	1378	**1259**	1444
March	1366	**1287**	1332	1337
February	**1320**	1388	1382	1395
January	1346	**1300**	1377	1344

Table 3.9: Mean absolute error for each data set from the final model for each prediction date.

	Mean Absolute Error (kg/ha)		
Prediction date	Development	Selection	Evaluation
July	792	1045	1840
June	1048	1129	1751
May	869	1052	1804
April	776	1076	1598
March	844	1019	1879
February	755	1030	1554
January	800	1103	1498

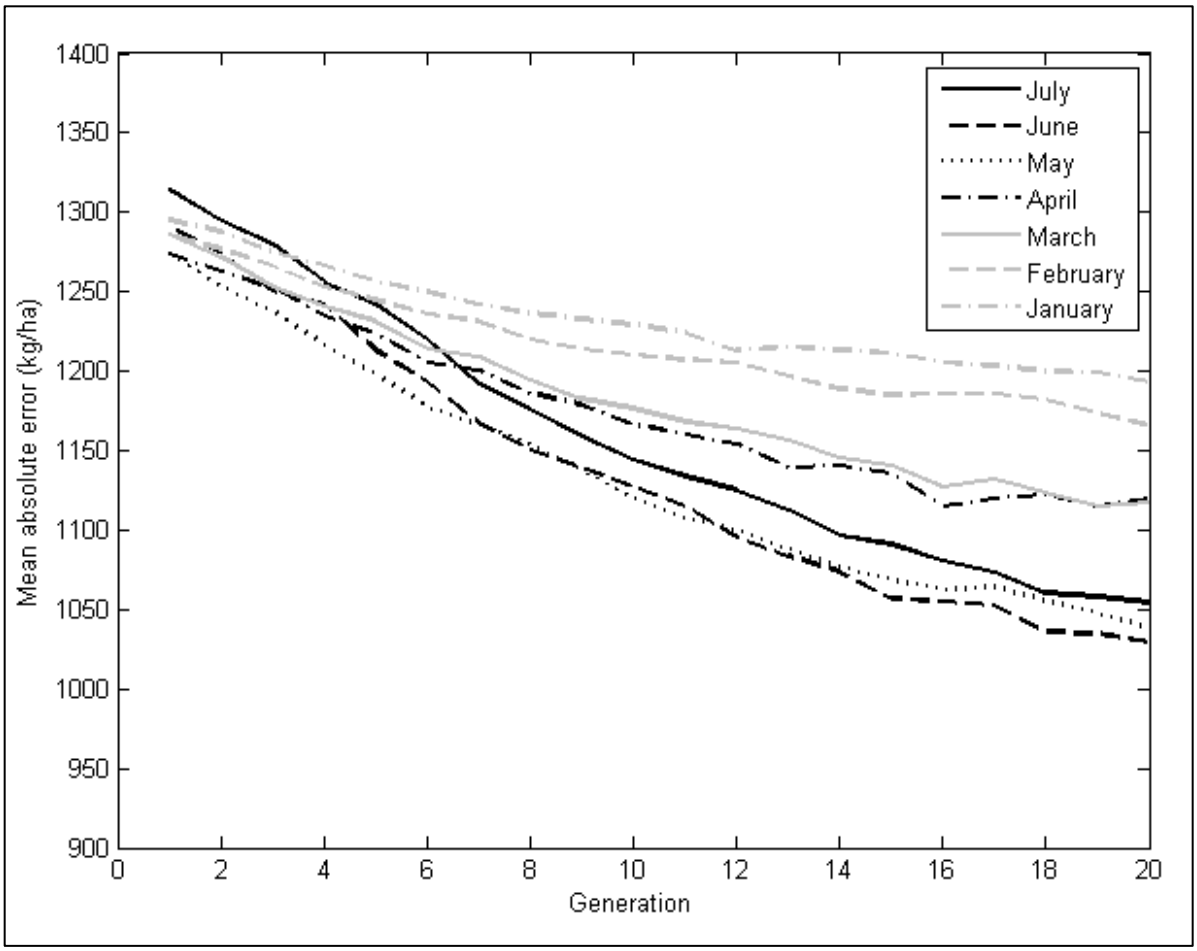

Figure 3.1: Average development set mean absolute error for seven prediction dates. Each line is averaged from 20 runs of the genetic algorithm.

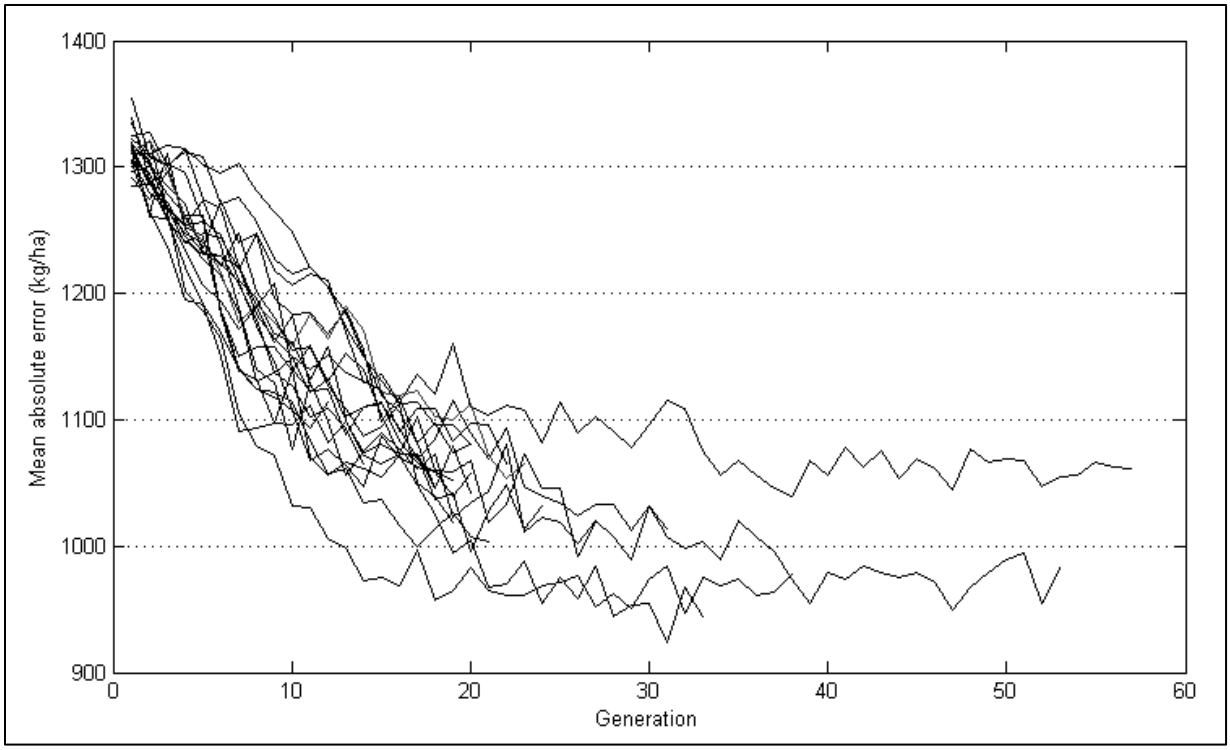

Figure 3.2: Average fitness expressed as mean absolute error on the development set for each generation of the 20 genetic algorithm runs for the July prediction date.

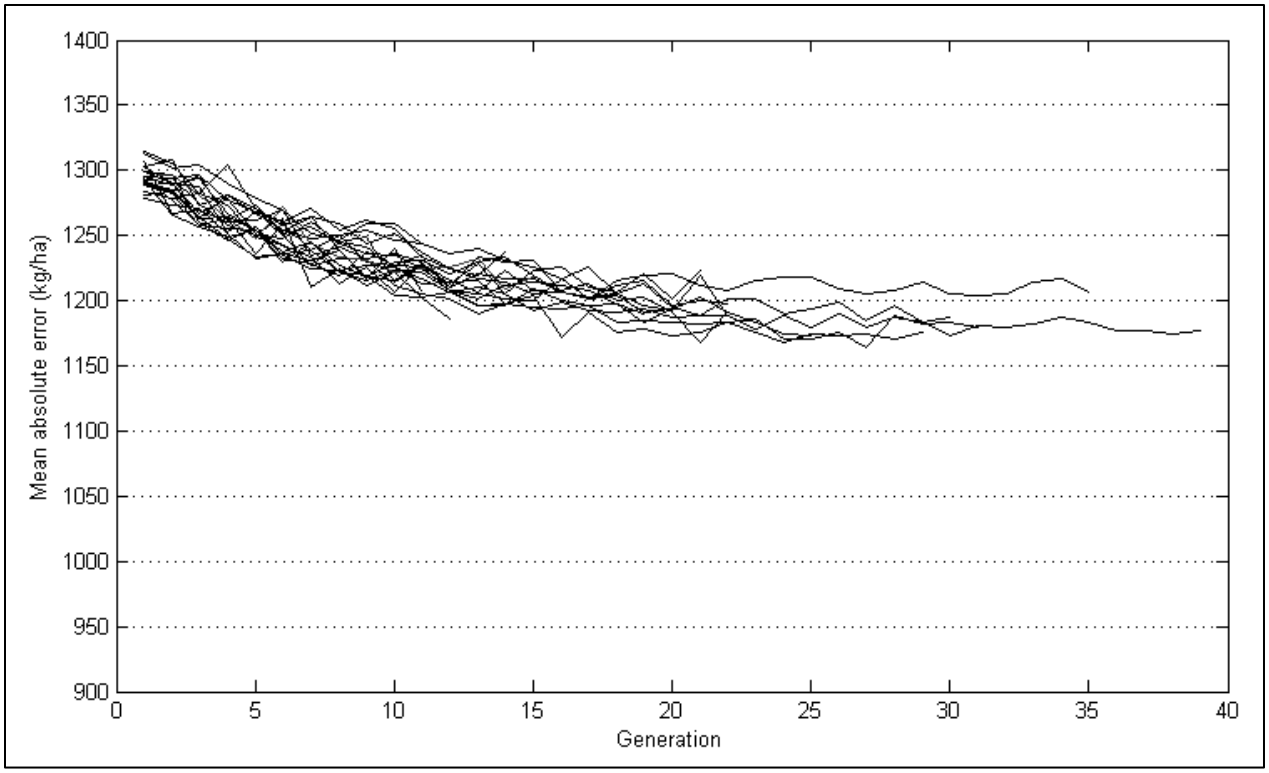

Figure 3.3: Average Mean Absolute Error on the development set for each generation of the 20 genetic algorithm runs for the January prediction date.

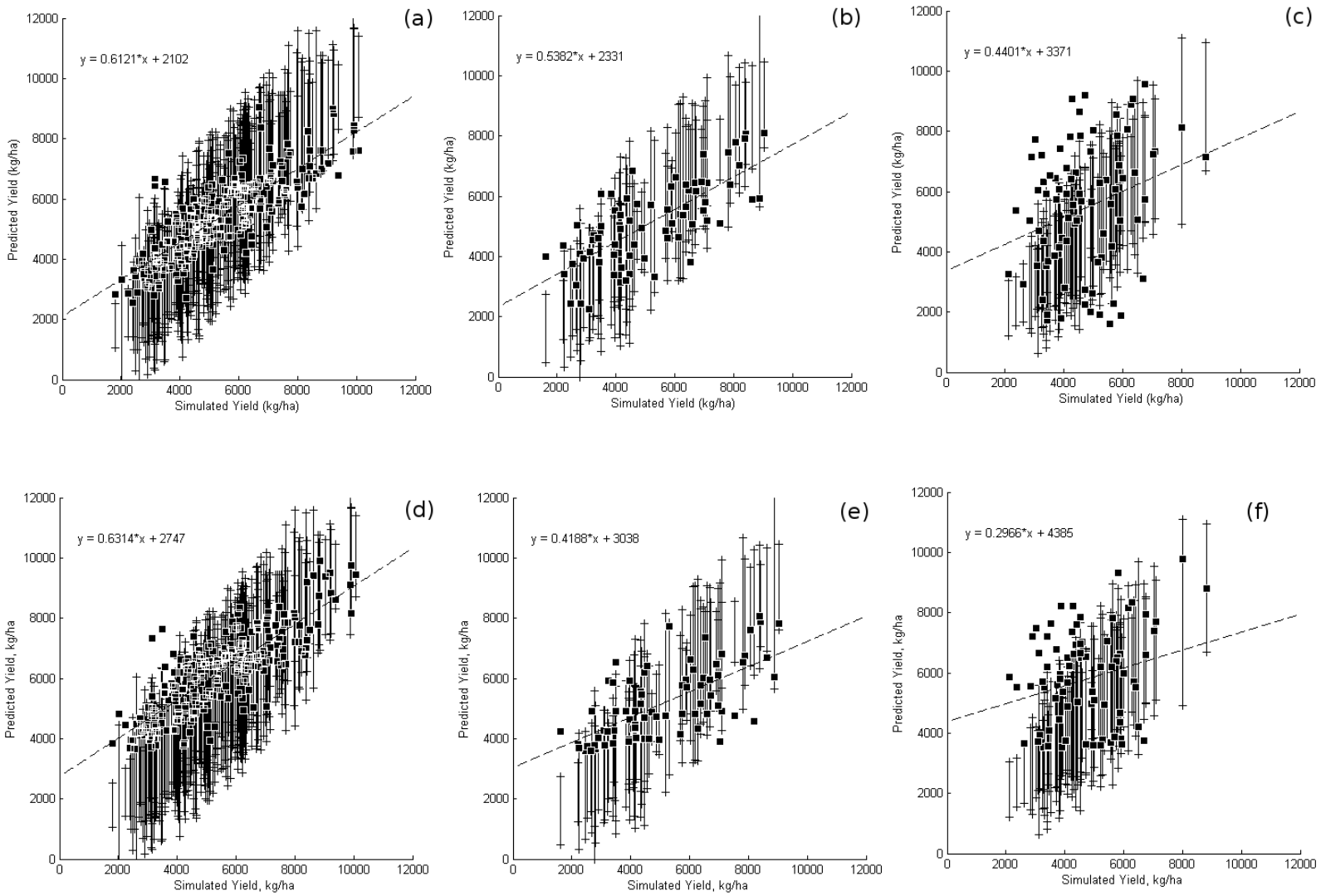

Figure 3.4: Yield predictions vs. target simulated yield for July development (a), selection (b), and evaluation (c) sets, and for the June development (d), selection (e), and evaluation (f) sets.

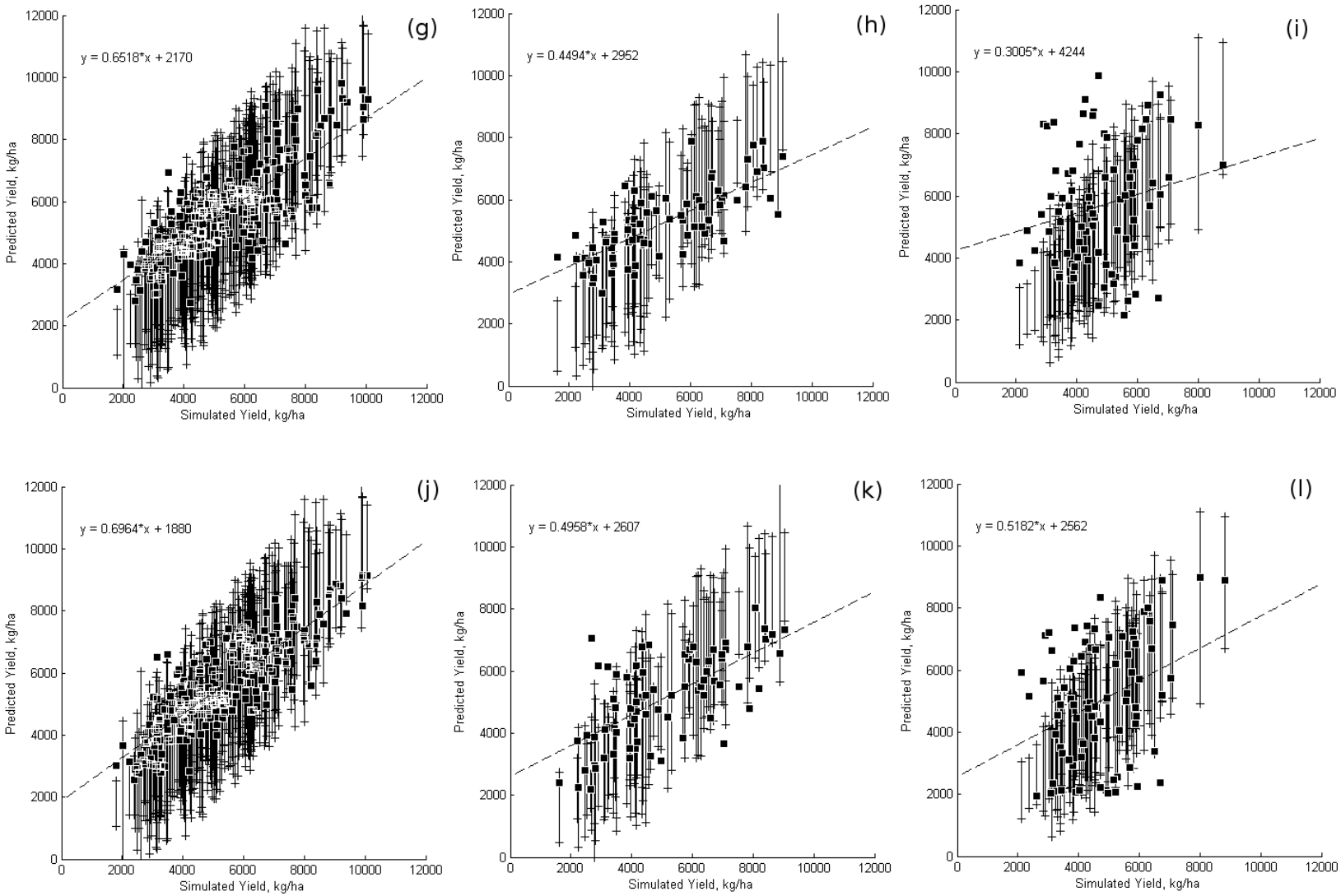

Figure 3.4 (cont.): Yield predictions vs. target simulated yield for May development (g), selection (h), and evaluation (i) sets, and for the April development (j), selection (k), and evaluation (l) sets.

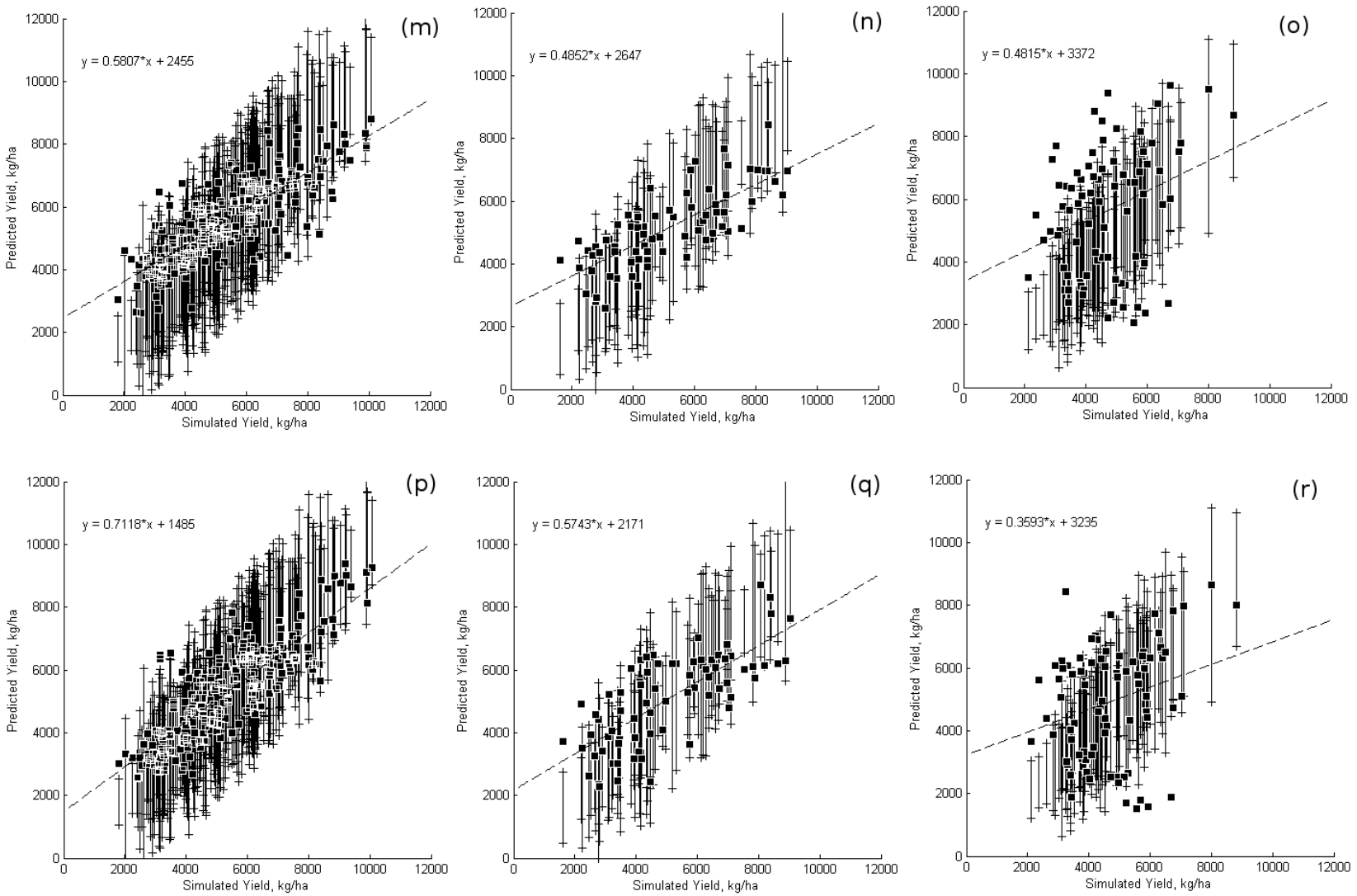

Figure 3.4 (cont.): Yield predictions vs. target simulated yield for March development (m), selection (n), and evaluation (o) sets, and for the February development (p), selection (q), and evaluation (r) sets.

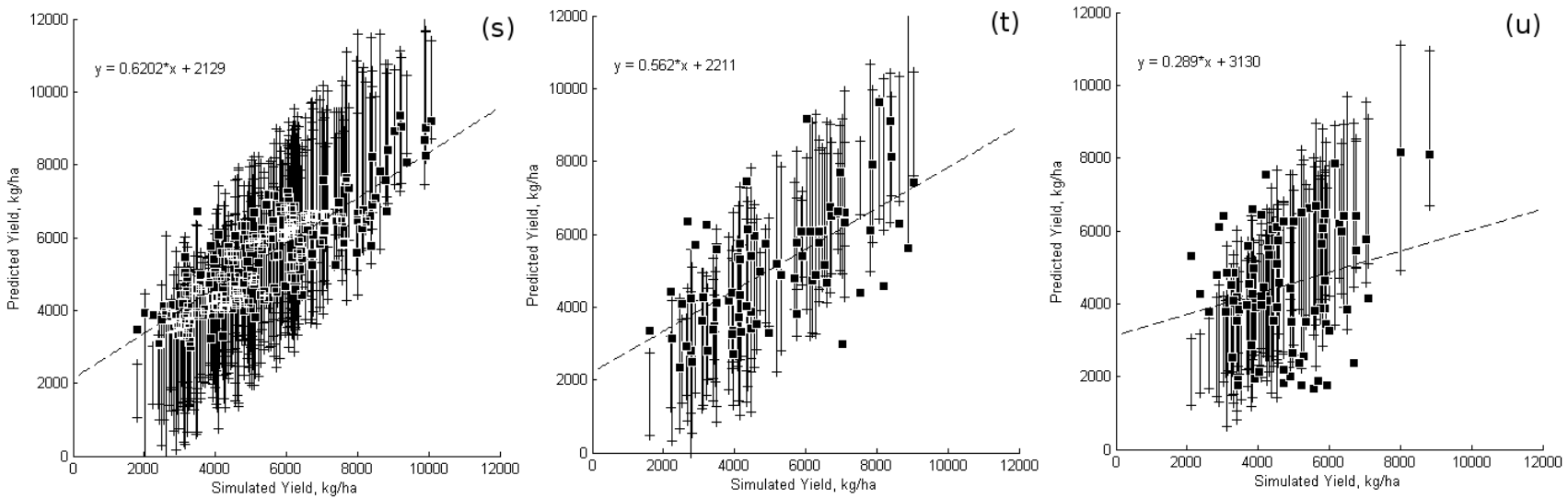

Figure 3.4 (cont.): Yield predictions vs. target simulated yield for January development (s), selection (t), and evaluation (u) sets.

CHAPTER 4

SUMMARY AND CONCLUSIONS

The goal of the research presented in this thesis was to predict southeastern US maize yields with ANN models using GA-selected inputs. Models were developed for seven prediction dates throughout the year, ranging from January to July of the year of harvest. In Chapter 2, parameters for ANN model development were selected based upon trials using all available inputs. All ANN models developed in Chapter 2 used data ranging through June of the year of harvest, with a prediction date of July. A GA search was then conducted to select a subset of inputs to produce a model with minimal error. Each input was assigned a proportion of inclusion based upon the percentage of GA searches in which it was included as an element of the final solution. Threshold values were applied to these proportions in order to select a final set of inputs. A final model was developed using these inputs and tested on an independent evaluation set. This final model achieved lower error values on both the model development and selection sets than the initial model using all available inputs, and had a mean absolute error of 1840 kg/ha on the evaluation set.

In Chapter 3, six more final models were developed for prediction dates earlier in the year, ranging from January to June. These were compared to one another, as well as the final model from Chapter 2. Moving the prediction date earlier in the year reduced the GA's overall ability to minimize error through input selection. Changing the prediction date also affected the frequency with which certain inputs were selected by the GA. When the performance of the final

model for each prediction date was compared, the models for the earliest prediction dates (January and February) were found to have the lowest error on the evaluation set.

This research demonstrates that indices of large-scale meteorological phenomena can be successfully used for modeling regional crop yields. Additionally, the improvement in model accuracy when using GA-selected subsets of inputs indicates that GA/ANN hybrids are useful tools for meteorological and agricultural applications. Further work could explore other enhancements to boost the accuracy of the predictions made by the ANN models presented in this thesis, such as searching a larger set of inputs or using other search and optimization methods apart from GAs. Future research could also explore the weights assigned to separate inputs by the GA searches, and how these align with existing research into the underlying meteorological phenomena. Additionally, the methods used in this study could be applied to other geographical regions or other types of crops, thus increasing the general applicability of the techniques presented in this thesis.

REFERENCES

Baigorria, G. A., J. W. Hansen, N. Ward, J. W. Jones, and J. J. O'Brien, 2008. Assessing predictability of cotton yields in the southeastern United States based on regional atmospheric circulation and surface temperatures. Journal of Applied Meteorology and Climatology 47: 76-91.

Bose, N. K. and P. Liang, 1996. Neural network fundamentals with graphs, algorithms, and applications. In McGraw-Hill Series in Electrical and Computer Engineering, ed. S. W. Director. New York, NY: McGraw-Hill.

Center for Ocean-Atmospheric Prediction Studies. 2009. Monthly JMA Index. Florida State University. ftp://www.coaps.fsu.edu/pub/JMA_SST_Index/. Accessed on June 10, 2009; Verified on November 16, 2009.

D'Arrigo, R., and R. Wilson, 2008. El Niño and Indian Ocean influences on Indonesian drought: implications for forecasting rainfall and crop productivity. International Journal of Climatology 28(5): 611-616.

Dasgupta, D., and D. R. McGregor, 1992. Designing application-specific neural networks using the structured genetic algorithm. COGANN-92: International Workshop on Combinations of Genetic Algorithms and Neural Networks, Baltimore, MD: 87-96.

Enfield, D. B., 1996. Relationships of inter-American rainfall to tropical Atlantic and Pacific SST variability. Geophysical Research Letters 23(23): 3305-3308.

Everingham, Y. L., R. C. Muchow, R. C. Stone, and D. H. Coomans, 2003. Using southern oscillation index phases to forecast sugarcane yields: a case study for Northeastern Australia. International Journal of Climatology 23(10): 1211-1218.

Guo, Z., and R. E. Uhrig, 1992. Using genetic algorithms to select inputs for neural networks. COGANN-92: International Workshop on Combinations of Genetic Algorithms and Neural Networks, Baltimore, MD: 223-234.

Handler, P, 1990. USA corn yields, the El Niño and agricultural drought: 1867-1988. International Journal of Climatology 10(8): 819-828.

Hansen, J. W., A. W. Hodges, and J. W. Jones, 1998. ENSO Influences on agriculture in the southeastern United States. Journal of Climate 11(3): 404-411.

Hansen, J. W., J. W. Jones, C. F. Kiker, A. W. Hodges, 1999. El Niño-Southern Oscillation impacts on winter vegetable production in Florida. Journal of Climate 12(1): 92-102.

Haykin, S, 1999. Neural Networks: A Comprehensive Foundation (Second Edition). Upper Saddle River, NJ: Prentice Hall.

Henderson, C. E., W. D. Potter, R. W. McClendon, and G. Hoogenboom, 1998. Using a genetic algorithm to select parameters for a neural network that predicts aflatoxin contamination in peanuts. In Methodology and Tools in Knowledge-Based Systems, by Tim Hendtlass, et al., 460-469. Berlin: Springer.

Holland, J. H, 1975. Adaptation in Neural and Artificial Systems. Ann Arbor, MI: University of Michigan Press.

Hoogenboom, G., J.W. Jones, P.W. Wilkens, C.H. Porter, W.D. Batchelor, L.A. Hunt, K.J. Boote, U. Singh, O. Uryasev, W.T. Bowen, A.J. Gijsman, A. du Toit, J.W. White, and G.Y. Tsuji, 2004. Decision Support System for Agrotechnology Transfer Version 4.0 [CD-ROM]. Honolulu, HI: University of Hawaii.

Hurrell, J.W. and C. Deser, 2009. North Atlantic climate variability: The role of the North Atlantic Oscillation. Journal of Marine Systems 78(1): 28-41.

Ibarra, R., and T. Hewitt, 1999. Utilizing crop insurance to reduce production risk. Institute of Food and Agricultural Sciences FE-198, Florida Cooperative Extension Service.

Izaurralde, R. C., N. J. Rosenberg, R. A. Brown, D. M. Legler, M. T. Lopez, R. Srinivasan, 1999. Modeled effects of moderate and strong 'Los Niños' on crop productivity in North America. Agricultural and Forest Meteorology 94(3): 259-268.

Jain, A., R. W. McClendon, G. Hoogenboom, and R. Ramyaa, 2003. Prediction of frost for fruit protection using artificial neural networks. American Society of Agricultural Engineers, St. Joseph, MI, ASAE Paper 03-3075.

Jain, A., R. W. McClendon, and G. Hoogenboom, 2006. Freeze prediction for specific locations using artificial neural networks. Transactions of the ASABE 49(6): 1955-1962.

Jones, J. W., G. Hoogenboom, C. H. Porter, K. J. Boote, W. D. Batchelor, L. A Hunt, P. W. Wilkens, U. Singh, A. J. Gijsman, and J. T. Ritchie, 2003. The DSSAT cropping system model. European Journal of Agronomy 18(3): 235-265.

Julstrom, B. A, 1994. Seeding the population: improved performance in a genetic algorithm for the rectilinear Steiner problem. Proceedings of the 1994 ACM Symposium on Applied Computing. Phoenix, AZ: 222-226.

Leathers, D. J., B. Yarnal, M. A. Palecki, 1991. The Pacific/North American pattern and United States climate. Journal of Climate 4(5): 517-528.

Martinez, C. J., G. A. Baigorria, and J. W. Jones, 2009. Use of climate indices to predict corn yields in southeast USA. International Journal of Climatology 20(11): 1680-1691.

National Oceanic and Atmospheric Adminstration, 2009a. Monthly mean North Atlantic Oscillation index. NOAA Climate Prediction Center. http://www.cpc.ncep.noaa.gov/products/precip/CWlink/pna/nao.shtml. Accessed on June 10, 2009; Verified on November 16, 2009.

National Oceanic and Atmospheric Adminstration, 2009b. Monthly mean Pacific-North American Pattern index. NOAA Climate Prediction Center. http://www.cpc.ncep.noaa.gov/products/precip/CWlink/pna/pna.shtml. Accessed on June 10, 2009; Verified on November 16, 2009.

National Oceanic and Atmospheric Adminstration, 2009c. Monthly Oceanic Niño Index. NOAA Climate Prediction Center. http://www.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ensoyears.shtml. Accessed on June 10, 2009; Verified on November 16, 2009.

Persson, T., A. Garcia y Garcia, J. O. Paz, J. W. Jones, and G. Hoogenboom, 2009a. Net energy value of maize ethanol as a response to different climate and soil conditions in the southeastern USA. Biomass & Bioenergy 33(8): 1-10.

Persson, T., A. Garcia y Garcia, J. Paz, J. Jones, and G. Hoogenboom, 2009b. Maize ethanol feedstock production and net energy value as affected by climate variability and crop management practices. Agricultural Systems 100(1): 11-21.

Philander, S. G, 1990. El Niño, La Niña, and The Southern Oscillation. San Diego, CA: Academic Press.

Shank, D. B., G. Hoogenboom, and R. W. McClendon, 2008a. Dewpoint temperature prediction using artificial neural networks. Journal of Applied Meteorology & Climatology 47(6): 1757-1769.

Shank, D. B., R. W. McClendon, J. Paz, and G. Hoogenboom, 2008b. Ensemble artificial neural networks for prediction of dew point temperature. Applied Artificial Intelligence 22(6): 523-542.

Smith, B. A., R. W. McClendon, and G. Hoogenboom, 2007. Improving air temperature prediction with artificial neural networks. International Journal of Computational Intelligence 3(3): 179-186.

Smith. B. A., G. Hoogenboom, and R. W. McClendon, 2009. Artificial neural networks for automated year-round temperature prediction. Computers and Electronics in Agriculture 68(1): 52-61.

Travasso, M. I., G. O. Magrin, M. O. Grondona, and G. R. Rodriguez, 2009. The use of SST and SOI anomalies as indicators of crop yield variability. International Journal of Climatology 29: 23-29.

Trenberth, K. E, 1997. The definition of El Niño. Bulletin of the American Meteorological Society 78(12}: 2771-2777.

Ward Systems Group, 1993. Manual of Neuroshell 2. Frederick, MD.

Whitley, D, 1995. Genetic algorithms and neural networks. In Genetic Algorithms in Engineering and Computer Science, edited by J. Periaux and G. Winter, 191-201. John Wiley & Sons Ltd.

Yang, J., and V. Honavar, 1998. Feature subset selection using a genetic algorithm. IEEE Intelligent Systems 13(2): 44-49.