WHEAT

Combining Field Surveys, Remote Sensing, and Regression Trees to Understand Yield Variations in an Irrigated Wheat Landscape

David B. Lobell,* J. Ivan Ortiz-Monasterio, Gregory P. Asner, Rosamond L. Naylor, and Walter P. Falcon

ABSTRACT

Improved understanding of the factors that limit crop yields in farmers' fields will play an important role in increasing regional food production while minimizing environmental impacts. However, causes of spatial variability in crop yields are poorly known in many regions because of limited data availability and analysis methods. In this study, we assessed sources of between-field wheat (*Triticum aestivum* L.) yield variability for two growing seasons in the Yaqui Valley, Mexico. Field surveys conducted in 2001 and 2003 provided data on management practices for 68 and 80 wheat fields throughout the Valley, respectively, while yields on these fields were estimated using concurrent Landsat satellite imagery. Management–yield relationships were analyzed with *t* tests, linear regression, and regression trees, all of which revealed significant but year-dependent impacts of management on yields. In 2001, an unusually cool year that favored high yields, N fertilizer was the most important source of between-field variability. In 2003, a warmer year with reduced irrigation water allocations, the timing of the first postplanting irrigation was found to be the most important control. Management explained at least 50% of spatial yield variability in both years. Regression tree models, which were able to capture important nonlinearities and interactions, were more appropriate for analyzing yield controls than traditional linear models. The results of this study indicate that adjustments in management can significantly improve wheat production in the Yaqui Valley but that the relevant controls change from year to year.

Variability in crop yields between fields is a ubiquitous feature of agricultural landscapes and often manifests itself in a significant gap between average yields and those achieved on the highest-yielding lands. Narrowing this yield gap will play a critical role in raising food production in step with continued growth in demand, especially as the genetic yield potential ceiling for many major crops fails to increase at historical rates (Cassman, 1999). Improved understanding of which factors most limit yields in farmers' fields (and, as importantly, those that do not) is also needed to reduce environmental impacts of agriculture, such as those resulting from overapplication of fertilizers, and to identify opportunities for improving farmer income.

Despite the importance and prevalence of the yield gap, its precise causes in many regions are not well known, owing in part to a lack of data on spatial variations in crop yields and yield-controlling factors (White et al., 2002). Surveys of farmer practices, supplemented by measurements of soil properties and crop performance, have provided a valuable means of assessing yield constraints in farmers' fields (e.g., Calvino and Sadras, 2002; Sadras et al., 2002). However, the time required to conduct a comprehensive survey, and in particular to collect accurate soil and crop measurements, can limit the number and extent of surveys. This is particularly true in regions with limited resources devoted to agricultural research, such as throughout the developing world. In addition, surveys are often motivated by specific questions and, as a result, fail to measure the full suite of variables needed to analyze yield variation (Wiese, 1982).

Recent developments in remote sensing have shown great promise for quantifying yield variations both within and between fields (Maas, 1988; Moulin et al., 1998; Shanahan et al., 2001; Baez-Gonzalez et al., 2002; Lobell et al., 2003). However, while many studies have employed remote sensing in precision agriculture to analyze variations within individual fields (e.g., Wiegand et al., 1994; Plant, 2001), few have addressed between-field yield variations across the landscape. In the context of crop surveys, yield remote sensing potentially provides three unique advantages over ground-based approaches. First, the ability to bypass field measurements of yield allows more time for other survey activities, which can result in increased sample sizes. Second, remote sensing allows yield estimates at a range of spatial scales, whereas field measurements are typically obtained from a limited number of small plots within fields and are therefore prone to sampling errors associated with within-field variability. Third, crop yields can be assessed for previous growing seasons using archived imagery, enabling analysis of past surveys that may not have measured yield.

Remote sensing thus offers a chance to increase the quantity and quality of survey data needed to identify on-farm yield constraints. Another important factor for understanding yield constraints is the type of model used to analyze the data. Multiple linear regression modeling, for example, is a commonly used approach but can lead to inaccurate and unstable solutions when applied to data sets with certain characteristics, such as a large number of insignificant predictor variables or the pres-

D.B. Lobell and G.P. Asner, Dep. of Global Ecol., Carnegie Inst. of Washington, Stanford, CA 94305, and Dep. of Geol. and Environ. Sci., Stanford Univ., Stanford, CA 94305; J. Ivan Ortiz-Monasterio, Int. Maize and Wheat Improvement Cent. (CIMMYT), Wheat Progr., Apdo. Postal 6-641, 06600 Mexico D.F., Mexico; and R.L. Naylor and W.P. Falcon, Cent. for Environ. Sci. and Policy, Inst. for Int. Studies, Stanford Univ., Stanford, CA 94305. Received 4 Mar. 2004. *Corresponding author (dlobell@stanford.edu).

Published in Agron. J. 97:241–249 (2005).

677 S. Segoe Rd., Madison, WI 53711 USA

Abbreviations: CC, compacted clay; DC, deep clay; OLS, ordinary least squares.

Table 1. Remotely sensed estimates of wheat area and yield in the Yaqui Valley for survey years compared with official statistics.

Year	Estimated area	Official area	Difference	Estimated yield	Official yield	Difference
	ha		%	t ha^{-1}		%
2001	151 642	152 439	−0.5	6.08	5.98	1.7
2003	174 270	179 542	−2.9	4.92	5.00	−1.6

ence of strong interactions between variables (Hastie et al., 2001). Yield survey data, which often exhibit both of these characteristics, may therefore be poorly modeled with linear regression. Various alternatives to linear models have been developed in recent years that take advantage of the greater computing power available today. One such technique is regression tree modeling (Breiman et al., 1984), which is a conceptually simple yet powerful analysis tool that has been increasingly applied in ecological and agricultural sciences (e.g., Plant et al., 1999; De'ath and Fabricius, 2000; Lapen et al., 2001). Important features of regression trees related to survey data are (i) automated variable selection, (ii) a structure that highlights interactions between variables, (iii) ease of interpretation, and (iv) an ability to handle missing data (Hastie et al., 2001).

This study investigates sources of between-field yield variability in the Yaqui Valley, an irrigated region comprising 225 000 ha in Sonora, Mexico. Average yields of wheat, the main crop in the Valley, increased from roughly 2.0 t ha^{-1} in 1960 to 5.0 t ha^{-1} in 1980 and have since remained near this level. Yet experimental trials and several farmers in the region regularly attain yields of 7.5 to 8 t ha^{-1}, indicating a yield gap of roughly 2.5 t ha^{-1} that represents a significant opportunity for increasing regional production. Periodic surveys have been conducted in the Valley since 1981, revealing considerable variability in farmer practices (Flores et al., 2001). However, only one survey directly measured crop yields, and in this case, the factors underlying variability were not clearly resolved, due in part to limited yield variability among the 52 samples (Meisner et al., 1992).

Here we used Landsat Enhanced Thematic Mapper Plus (ETM+) data, with 30-m spatial resolution, to estimate wheat yields in the Yaqui Valley for the 2001 and 2003 harvest seasons. These yield estimates were combined with data on management practices from coincident field surveys to identify factors contributing to yield variations in farmers' fields. Both linear regression and regression trees were used to analyze management–yield relationships, providing a means to assess the relative performance of each technique in the context of explaining the yield gap.

MATERIALS AND METHODS

Site Description

The Yaqui Valley is an intensive agricultural region situated along the Gulf of California coast, with agroclimatic conditions similar to that of 40% of developing world wheat production (Pingali and Rajaram, 1999). Fields in this region average roughly 20 ha in size, with up to 85% of cultivated land planted with wheat each winter season (November–April). Daily temperatures during the wheat growing season average 9.8 and 27.1°C for nighttime and daytime, respectively. Soils in this region are predominantly vertisols, with elevation varying very gradually from 0 m on the western coast to a peak of 60 m in the eastern edge of the district.

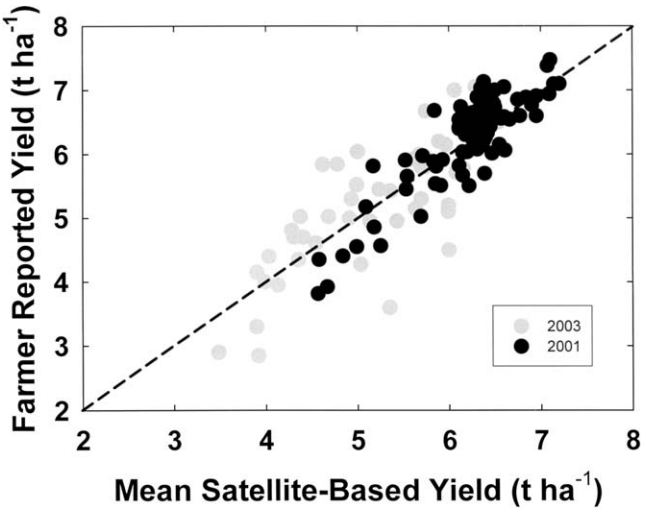

Fig. 1. Comparison of yield estimates derived from Landsat with farmer-reported values in 2001 and 2003. Regression statistics for 2001: n = 80, root mean square error (RMSE) = 0.37 t ha^{-1}, and R^2 = 0.78. For 2003: n = 47, RMSE = 0.64 t ha^{-1}, and R^2 = 0.60. For all data: n = 127, RMSE = 0.49 t ha^{-1}, and R^2 = 0.76.

Remotely Sensed Yield Estimates

Landsat ETM+ images of the Yaqui Valley were acquired on 11 Jan. and 16 Mar. of 2001 and 1 Jan., 6 Mar., and 22 Mar. of 2003. These images were used to estimate wheat yields following the approach described in detail by Lobell et al. (2003). Briefly, this approach uses instantaneous estimates of canopy light absorption from the satellite images to adjust a locally calibrated model of wheat growth, which then provides an estimate of wheat yield for each pixel determined, based on a multitemporal classification, to contain wheat. The total area and average yield estimates for the two growing seasons were within 3% of values reported for the agricultural district (Table 1). In addition, yields for individual fields provided by local farmers were compared with the average of remote-sensing estimates for pixels completely contained within their fields. This field-level evaluation resulted in a close agreement between ground and remote-sensing–based estimates (Fig. 1), demonstrating the ability of remote sensing to capture spatial variability of yields across the landscape.

Field Surveys

Field surveys were conducted from late February to late March of each year to obtain information on management practices. The measured variables (a subset of which are defined in Table 2) included methods of soil preparation; choice of wheat cultivar; date and method of planting; type, timing, and amount of fertilizer applications; timing and number of irrigations; type and amount of herbicide, fungicide, and insecticide applications; and residue management techniques. The farmer estimated the date of final irrigation in most cases because the last irrigation typically occurs in late March, after the surveys were completed. Importantly, we measured all management variables believed to have a potential impact on yield, to avoid misinterpretation arising from the existence of important latent variables. Information on selected socioeco-

Table 2. Statistical summary of selected management variables and yield estimates for survey fields.

		2001				2003			
Variable	Description	Mean	SD†	Median	IQR‡	Mean	SD	Median	IQR
DTPL	planting date, d after 1 Nov.	41.8	11.0	43.0	16.0	34.1	13.0	35.5	21.3
QSEED	seeding density, kg ha^{-1}	146.3	27.3	150.0	40.5	146.6	23.5	150.0	31.0
BEDSP	bed spacing, cm	79.3	5.4	80.0	0.0	79.1	10.2	80.0	0.0
ROWSP	row spacing, cm	17.7	4.4	17.0	5.0	16.5	3.7	15.0	4.0
NOAP	number of fertilizer applications	2.8	0.8	3.0	1.0	2.2	0.6	2.0	0.3
N	total applied N, kg N ha^{-1}	263.3	43.5	257.0	67.5	250.8	50.3	251.5	52.0
P	total applied P, kg P ha^{-1}	37.3	22.5	46.0	27.5	50.7	15.0	52.0	6.0
KGHAN1	N applied in first application, kg N ha^{-1}	160.6	47.5	150.0	38.0	158.4	43.3	149.0	45.3
KGHAN2	N applied in second application, kg N ha^{-1}	69.5	27.1	70.0	29.5	83.6	37.8	82.0	34.0
HERBICID	herbicide applied (no = 0, yes = 1)	0.6	0.5	1.0	1.0	0.7	0.5	1.0	1.0
INSECT	insecticide applied (no = 0, yes = 1)	0.7	0.4	1.0	1.0	1.0	0.2	1.0	0.0
FUNGICID	fungicide applied (no = 0, yes = 1)	0.0	0.2	0.0	0.0	0.7	0.5	1.0	1.0
NOIRIG	number of irrigations	4.1	0.2	4.0	0.0	3.7	0.5	4.0	1.0
IRR0	preplant irrigation, d before planting	18.1	6.2	17.0	6.0	20.8	9.1	20.5	10.0
IRR1	first auxiliary irrigation, d after planting	50.1	4.6	50.0	7.0	54.7	8.8	55.0	10.0
IRR2	second auxiliary irrigation, days after planting	79.4	6.0	80.0	8.5	85.4	9.4	85.0	10.5
IRR3	third auxiliary irrigation, d after planting	100.3	6.1	100.0	7.0	102.4	7.0	103.0	6.0
YIELD	image yield estimate, t ha^{-1}	6.4	0.6	6.4	0.8	5.2	0.8	5.3	1.2

† SD, standard deviation.
‡ IQR, interquartile range.

nomic factors, such as level of education, type of land tenure, and source of credit, were also collected.

In 2001, a total of 68 wheat fields (approximately 1% of all wheat fields) were randomly selected throughout the Valley. This survey was originally designed solely to update information on management practices within the Valley. The 2003 survey was specifically aimed at understanding yield variations. We therefore employed a stratified random sampling design in 2003 as follows. The two early-season images (1 January and 6 March) were used at the beginning of the survey period to generate an image of preliminary yield predictions. [These yield predictions are distinguished from the final estimates in two ways: They do not reflect changes in canopy condition observed in the 22 March image, and they use expected weather for the remainder of the growing season (multiyear averages) rather than actual weather.] The yield image was then combined with a GIS layer of the two main soil types in the Valley, namely deep clay (DC) and compacted clay (CC), to identify four classes of fields:

1. Predicted yield $>$ 5.5 t ha^{-1} on DC soils.
2. Predicted yield $>$ 5.5 t ha^{-1} on CC soils.
3. Predicted yield $<$ 5.0 t ha^{-1} on DC soils.
4. Predicted yield $<$ 5.0 t ha^{-1} on CC soils.

A random sample of 20 fields was selected from each of the four classes, resulting in a total of 80 fields. The main goal of this stratified design was to ensure sufficient contrast in yields between fields for the statistical analysis. A secondary goal was to evaluate soil type–management interactions.

Soil properties were not directly measured in the surveys, for several reasons. First, the 2001 survey was originally focused on understanding farmer practices and not specifically on sources of yield variability. Therefore, soil properties were not of direct interest in the original context of the 2001 survey. Second, the required time and expense for soil collection and analysis made soil testing for each field within the survey unfeasible. Third, an existing map of soil types within the Valley obtained from the National Institute of Forestry, Agricultural and Animal Research (INIFAP) enabled at least a general description of soils on each field. Fourth, and most importantly, a previous study of spatial patterns in remotely sensed yields indicated that the majority of yield variability occurred over short distances, suggesting that between-field variations in management practices were a more important contributor than soil properties to yield variability (Lobell et al., 2002). Thus, the limited scope and resources of the surveys, prior knowledge of general soil conditions, and indications that soil properties were not a major source of yield variability resulted in the absence of detailed soil measurements. In addition, meteorological conditions were not measured on each field but were assumed equal to conditions measured at the central meteorological station because of the close proximity of the fields and the minimal change in elevation. The implications of the missing soil and weather information are discussed below.

Data Analysis

Three approaches were used to assess causes of yield variation. In the first analysis, the data were split into two subsets: one containing fields with the highest 20 yields and the second with the lowest 20 yields. A *t* test was then performed for each survey variable to test the hypothesis that its average value was the same for the lowest- and highest-yielding fields (Meisner et al., 1992). The Mann–Whitney (or Wilcoxon) test, which is the nonparametric equivalent to the *t* test, was also used to ensure the results were not influenced by non-Gaussian distributions in the management variables (Conover, 1999). However, the results were very similar to the *t* test and are therefore not presented.

The second analysis employed multiple linear regression, with forward stepwise variable selection used to identify the relevant predictor variables (Hastie et al., 2001). The Akaike Information Criterion (AIC) was used to determine the stopping point (i.e., number of variables included):

$$\mathrm{AIC} = n \log(\mathrm{RSS}/n) + 2p$$

where n is the number of observations, RSS is the model residual sum of squares, and p is the number of parameters. The minimum of the AIC is commonly used, as in this case, to identify a parsimonious model that has both low error and few parameters.

Finally, the survey and yield data sets were analyzed with regression trees (Breiman et al., 1984). In this method, the response variable (i.e., yield) is modeled as a piece-wise constant function. The data are first split into two subsets based on the predictor variable and value of that variable that results in the greatest increase in explained variance of the response variable. Each subset, or daughter node, is then analyzed independently using the same binary partitioning procedure, with

Table 3. Comparison of yields and selected management variables for 20 lowest- and 20 highest-yielding fields in two survey years. Variable definitions and units are given in Table 2.

Year	Variable	20 lowest-yielding fields				20 highest-yielding fields				*p* value of *t* test
		Mean	SD†	Median	IQR‡	Mean	SD	Median	IQR	
2001	YIELD	5.7	0.3	5.7	0.5	7.3	0.3	7.2	0.3	0.00
	DTPL	37.7	9.3	37.5	8.8	45.2	11.7	45.0	15.3	0.03
	N	240.8	55.6	234.5	60.5	269.0	32.3	257.0	48.8	0.06
	P	33.7	24.9	38.0	34.8	40.1	22.4	49.0	25.3	0.40
	NOIRIG	4.2	0.4	4.0	0.0	4.0	0.0	4.0	0.0	0.08
	IRR0	18.5	5.9	17.0	5.0	17.3	5.7	15.5	7.5	0.52
	IRR1	50.0	4.5	50.0	4.8	50.2	4.3	50.0	5.3	0.91
	IRR2	78.4	6.2	79.5	8.0	79.2	5.2	80.0	8.5	0.66
	INSECT	0.7	0.5	1.0	1.0	0.8	0.4	1.0	0.0	0.48
	HERBICID	0.6	0.5	1.0	1.0	0.4	0.5	0.0	1.0	0.22
	FUNGICID	0.0	0.0	0.0	0.0	0.1	0.3	0.0	0.0	0.16
2003	YIELD	4.1	0.3	4.1	0.4	6.1	0.2	6.0	0.3	0.00
	DTPL	34.8	13.1	31.0	19.0	38.0	9.4	42.0	10.3	0.38
	N	256.9	45.6	265.0	51.5	259.0	41.4	260.0	51.8	0.88
	P	49.6	14.5	52.0	6.0	52.9	15.2	52.0	3.0	0.53
	NOIRIG	3.5	0.5	3.5	1.0	3.8	0.5	4.0	0.3	0.07
	IRR0	20.1	5.2	21.0	2.8	25.4	9.0	25.5	12.0	0.03
	IRR1	60.4	6.7	60.0	6.8	51.0	9.0	51.0	8.8	0.00
	IRR2	91.4	8.6	90.0	9.0	82.7	9.6	83.0	10.3	0.00
	INSECT	1.0	0.2	1.0	0.0	1.0	0.2	1.0	0.0	1.00
	HERBICID	0.6	0.5	1.0	1.0	0.8	0.4	1.0	0.0	0.18
	FUNGICID	0.7	0.5	1.0	1.0	0.8	0.4	1.0	0.3	0.50

† SD, standard deviation.
‡ IQR, interquartile range.

a split performed only if the resulting model exceeds a predefined threshold of improvement. The result of this recursive binary partitioning is a model whose structure can be displayed as a tree-like graph, with each split in the tree labeled according the threshold used to define the split. All analyses described above were implemented in the software package R (Ihaka and Gentleman, 1996).

A potential problem when applying ordinary least-squares (OLS) regression models to spatial data is that errors may be spatially correlated (i.e., not independent), which violates a basic assumption of OLS methods and may introduce bias into model interpretation (Long, 1998; Haining, 2003). In particular, one is prone to underestimate the uncertainties associated with model parameters and, thus, the corresponding *p* values. To assess the influence of spatially correlated errors, model residuals were tested for spatial correlation using the Moran I and Geary *c* tests. Both tests indicated a significant level of spatial correlation in model errors for both the linear regression and regression tree models ($p < 0.05$). We therefore repeated the linear regression analysis using a maximum likelihood approach to simultaneously solve for model parameters and spatial error correlation, as implemented by the package "spdep" in R. The coefficients of each model variable were within 6% of the original estimates, indicating that explicit consideration of spatial correlation of errors did not substantially change model estimates. We therefore present only the OLS estimates while acknowledging that spatial autocorrelation may contribute to underestimation of parameter uncertainties.

RESULTS AND DISCUSSION

The survey answers for selected management variables are summarized in Table 2. While a detailed analysis of management variability, in itself, is beyond the scope of this paper, we note that the distributions of management practices were generally similar between the two years. Exceptions to this are seen in the planting date, which averaged a week later in 2001 than 2003; timing of first and second auxiliary (postplanting) irrigations, which averaged 5 d later in 2003 than 2001; and the increased prevalence of fungicide application in 2003, resulting from more widespread infestation of leaf rust.

A comparison of the highest- and lowest-yielding fields (Table 3) revealed that no management variable was significantly different (i.e., $p < 0.05$) between the two yield classes in both years. For example, planting date (DTPL) and N rate (N) appeared as important factors in 2001 but not 2003, whereas irrigation timing appeared important in 2003 but not 2001.

These differences can be understood in the context of the different climatic conditions for the two years, as well as the different management regimes. Figure 2 shows the cumulative average daily temperature for the two growing seasons, along with the 20-yr average. While 2003 was representative of average temperatures in the past two decades, 2001 was an unusually cool year. In this region, cooler temperatures favor enhanced wheat

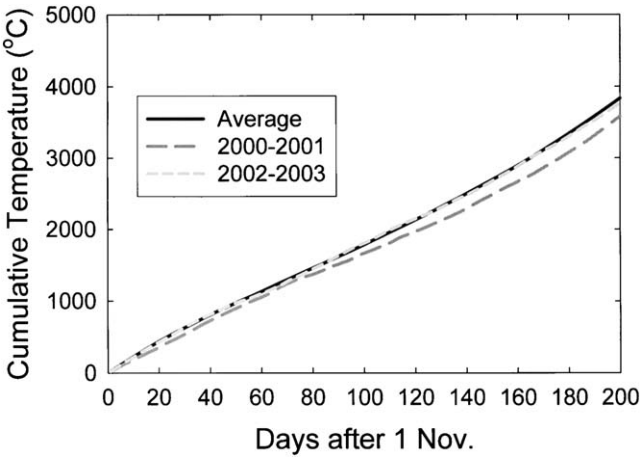

Fig. 2. Cumulative average daily temperature for 2001 and 2003 during the wheat growing season in the Yaqui Valley. Also shown is the average for 1984–2003.

yield potential (Fischer, 1985), as reflected in the higher yields achieved in 2001 (Table 1). Higher yield potential, in turn, increases demand for fertilizer N because the N requirement of wheat increases nonlinearly with grain yield (Ortiz-Monasterio, 2002).

In 2003, on the other hand, temperatures were less favorable to wheat growth. In addition, farmers were allocated an average of only 43 cm of irrigation water in 2003 compared with 53 cm in 2001, resulting in lower total irrigations and delayed application of the first irrigation by an average of 5 d (Table 2). The combination of reduced irrigation and warmer temperatures, which increase water demand, help to explain the increased importance of water in 2003.

The fact that sources of spatial yield variability changed significantly between years reflects the importance of climate–management interactions at the regional scale. As this study spanned only 2 yr, it is difficult to say whether most years in the Valley are similar to one of these years or whether each year presents a unique set of factors that dominate spatial yield variability. In either case, it is clear that management recommendations should account for climatic conditions when possible and that surveys conducted in individual years must be interpreted with caution when applied to new situations.

Stepwise Linear Regression

The *t* tests presented above provide useful comparisons of the relationship between individual factors and crop yields, but multivariate models are needed to assess the combined impact of different variables taking into account their covariations. For 2001, stepwise linear regression selected a model with nine variables (Table 4). In this model, insecticide application, N rates, planting date, P rates, and field ploughing were deemed positively related to yield, whereas negative relationships were inferred for bed reformation, number of irrigations, days between preplant irrigation and planting, and leveling of canals. The magnitude and sign of the regression coefficients should be interpreted with care, particularly for those variables with high standard errors since correlation between predictor variables can impact the regression estimates. In this case, predictor variables were not highly correlated, with only bed reformation and leveling of canals exhibiting a correlation with absolute magnitude greater than 0.35 (not shown). Nonetheless, of interest here is the explanatory power of the model, which equaled 51% with nine variables.

In 2003, only three variables were selected for the regression model (Table 5), with days between planting and first irrigation alone explaining 18% of yield variability. The two land-leveling operations LPLANE and LEVEL exhibited similar effects on yield but in opposite directions. LPLANE is done with a land plane, which is a large piece of equipment that does a much better job of leveling the land than LEVEL, which is the use of a plank to correct minor leveling problems. The primary objective of LEVEL is often solely to create a flat surface for machinery to drive on the field (e.g., for fertilizer application) and can only correct minor leveling problems. Because most farmers perform only one of the two operations, the two variables are negatively correlated. This may explain the apparent negative impact of LEVEL, which is simply a reflection of anticorrelation with effective land leveling (LPLANE).

Regression Trees

The regression tree models for 2001 and 2003 are shown in Fig. 3 and 4, respectively. In these figures, all observations that satisfy the criterion at a given split fall to the left-hand daughter node while those not meeting the criterion continue to the right. The number of fields and their average yield, which is equal to the model estimate, is shown for each terminal node. In 2001 for example, there were 16 fields that received greater than 240.5 kg N ha^{-1}, were planted before 13 December, and received their third irrigation less than 98.5 d after planting, with an average yield of 6.59 t ha^{-1}. Also shown (Fig. 3b) is a plot of observed vs. modeled yield, which indicates the spread associated with each terminal node.

Results of the regression tree model indicated that N rate was the most important variable determining yield in 2001, with fields that received more than 240.5 kg N ha^{-1} achieving higher yields than those that did not (Fig. 3). Planting date was of secondary importance and impacted only those fields with high N. The fact that no additional splits were performed on fields with low N rates indicates that N was the main constraint to yield for these fields. The overall model explained roughly

Table 4. Regression statistics for 2001 survey, with variables listed in order of selection by stepwise forward procedure.

Significant variables	Description	Estimate	SE†	*t* value	$P > \|t\|$	Cumulative R^2	AIC‡
Constant	model intercept	7.21	1.07	6.76	0.00	–	−61.67
INSECT	insecticide applied (no = 0, yes = 1)	0.58	0.15	3.79	0.00	0.09	−66.11
N	total applied N (kg N ha^{-1})	0.005	0.001	3.47	0.00	0.17	−70.02
REFORM	bed destruction and reformation, used for weed control (no. times)	−0.39	0.13	−2.96	0.00	0.25	−75.19
DTPL	planting date (days after 1 Nov.)	0.02	0.01	3.18	0.00	0.31	−78.84
NOIRIG	number of irrigations	−0.71	0.26	−2.70	0.01	0.37	−82.86
IRR0	preplant irrigation (days before planting)	−0.02	0.01	−2.21	0.03	0.41	−85.72
BORRADO	leveling of canals for machine access	−0.45	0.26	−1.71	0.09	0.46	−89.19
P	total applied P (kg P ha^{-1})	0.01	0.003	2.09	0.04	0.48	−90.59
PLOUGH	disk plowing to a depth of 22 to 28 cm	0.23	0.13	1.82	0.07	0.51	−92.37

† SE, standard error.
‡ AIC, Akaike Information Criterion.

Table 5. Regression statistics for 2003 survey, with variables listed in order of selection by stepwise forward procedure.

Significant variables	Description	Estimate	SE†	*t* value	$P > \|t\|$	Cumulative R^2	AIC‡
Constant	model intercept	7.47	0.51	14.59	0.00	–	−43.77
IRR1	first irrigation (days after planting)	−0.04	0.01	−4.43	0.00	0.18	−55.92
LEVEL	minor land-leveling operation using plank (no. times)	−0.29	0.14	−2.07	0.04	0.26	−60.85
LPLANE	major land-leveling operation using land plane (no. times)	0.25	0.14	1.84	0.07	0.29	−62.35

† SE, standard error.
‡ AIC, Akaike Information Criterion.

44% of yield variability using three variables, whereas the linear model used nine variables to explain 51%.

In 2003, time between planting and first irrigation was the most important variable, with fields irrigated more than 56 d after planting experiencing yield reductions (Fig. 4). In those fields that were irrigated in time, the amount of N received at first application was an important determinant of yield. In contrast, the most important variable for fields that were not irrigated in time was land leveling. These differences reflect the interaction between management factors; i.e., N levels were only important if the plant had sufficient water to make use of the N. Interestingly, not one of the 13 fields that received sufficient water and fertilizer fell below 5.5 t ha^{-1} while fields that were irrigated more than 56 d after planting and experienced one or more leveling were all below this level. The overall model explained roughly 52% of yield variability using three variables, which is a substantial improvement over the linear model (29% with three variables).

To evaluate the interaction between management and soil type, the regression tree model was applied separately to the fields on each soil type. The results indicated that timing of irrigation was the most important variable on both soils but that the critical threshold was 3.5 d earlier on the CC soil, which has lower water-holding capacity (Fig. 5). In addition, days between preplant irrigation and planting were deemed important on the CC while N rates were selected on the DC. These results are consistent with the fact that CC soils hold less water, and thus yields are more sensitive to water management.

A comparison of the linear regression and regression tree results reveals that the two models identified the same major management variable for each year: N in 2001 and days to first irrigation in 2003. However, regression trees were able to capture important nonlinear effects as well as interactions between management variables. In the context of identifying causes of the yield gap and potential management approaches, models that provide both a simple and accurate (i.e., parsimonious) representation of yield controls are desired (Lark, 2001). The regression tree models were therefore deemed more appropriate than the linear models for both survey years. In 2001, the regression tree model explained a similar amount of yield variability but with far fewer variables. In 2003, the regression tree explained a much greater amount of variability with the same number of variables. In both years, the regression tree models explained roughly half of yield variability using only three management factors.

Unmeasured Sources of Variability

The fraction of yield variability not explained by the statistical models above (roughly 50% in both years)

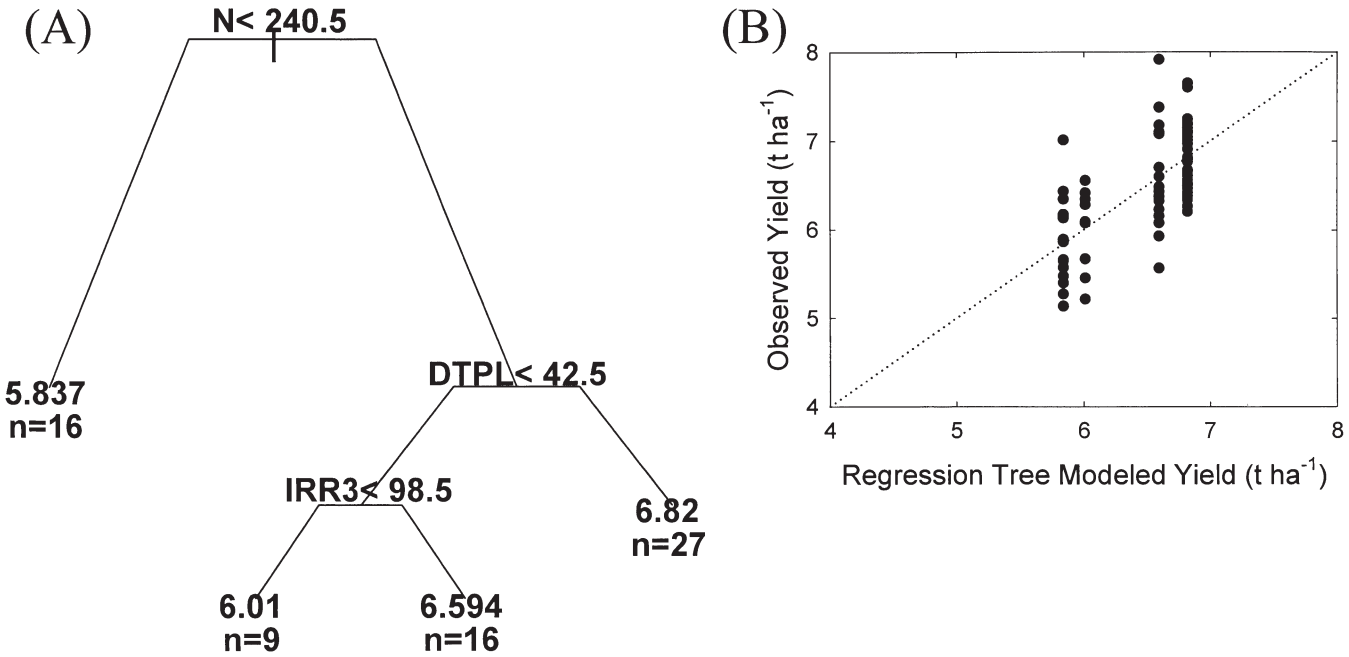

Fig. 3. (A) Regression tree model for 2001 survey. (B) Comparison of yield estimates with regression tree model predictions (R^2 = 0.44).

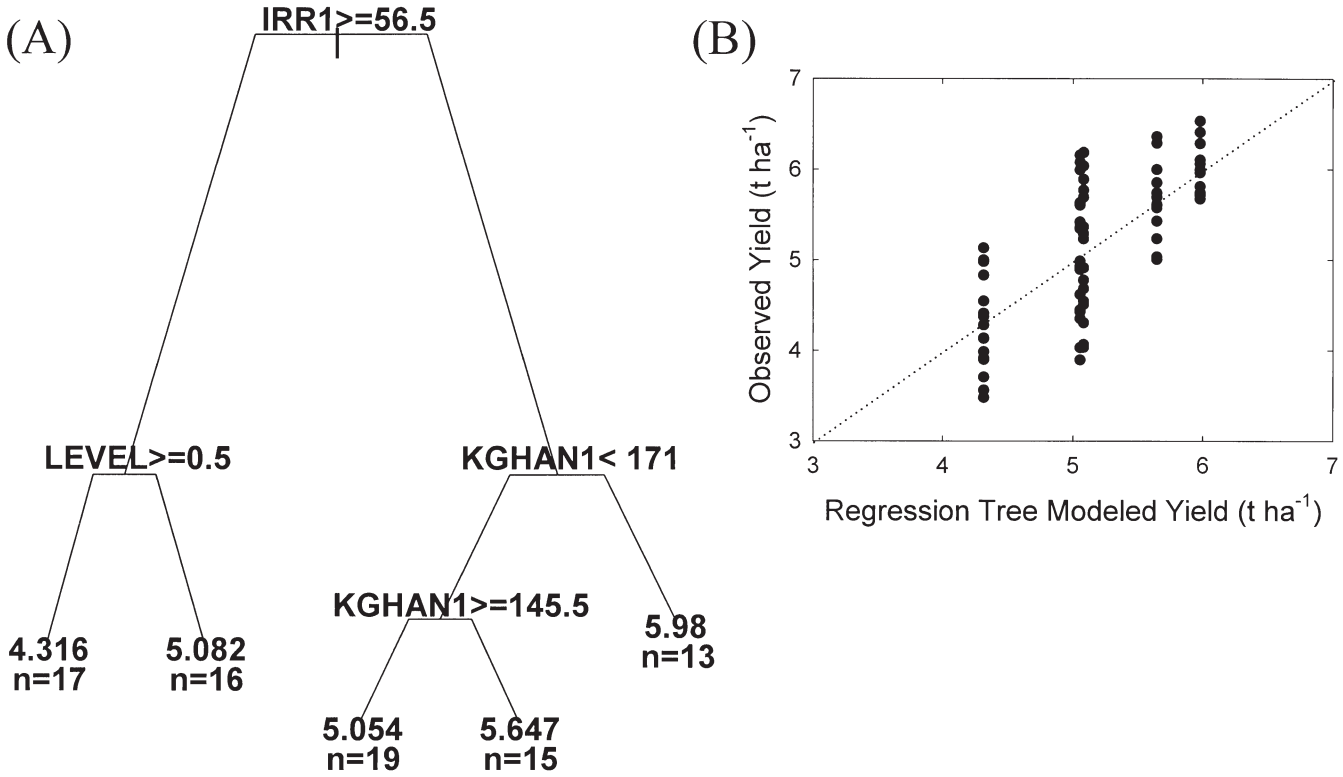

Fig. 4. (A) Regression tree model for 2003 survey. (B) Comparison of yield estimates with regression tree model predictions (R^2 = 0.52).

can generally be attributed to three factors. First is the presence of measurement error, both for management variables and yield estimates. While farmers' answers to survey questions represent the best available information, errors may result from imperfect farmer memory of practices for a specific land parcel. These errors were not quantified in this study, as doing so would require independent sources of management information. Errors in the Landsat yield estimates may also contribute to model error. Based on the observed correlation be-

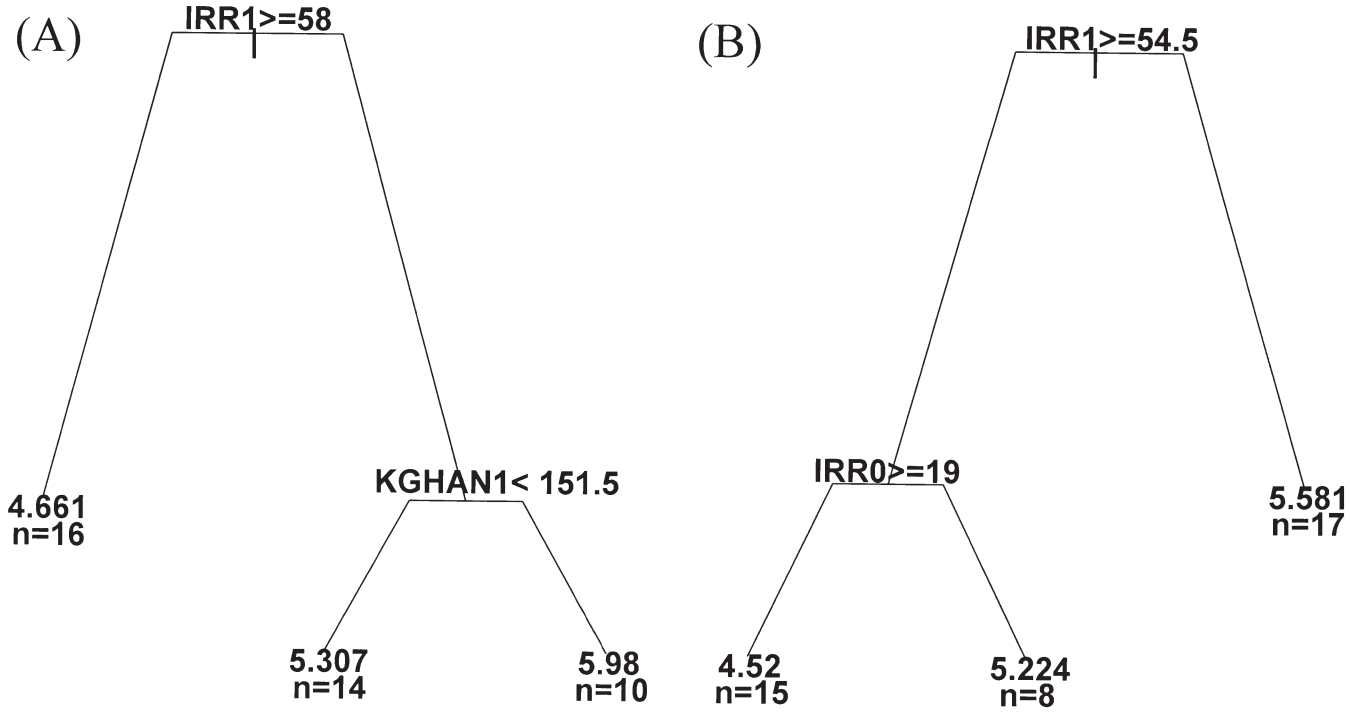

Fig. 5. Regression tree model for 2003 survey for fields in (A) deep clay (DC) and (B) compacted clay (CC) soils. R^2 = 0.40 on DC and 0.45 on CC.

tween actual and estimated yields (Fig. 1), yield errors can explain a maximum of roughly 25% of model error.

A second potential source of unexplained yield variance is an inability of the statistical models to capture the true relationship between management and yield. The improvement of decision trees over linear models, for instance, reflects the increase in explanatory power possible with more appropriate models. It is possible that the use of process-based crop models would improve the agreement between modeled and measured yields.

Finally, model error may reflect the absence of important explanatory variables, such as management variables that were not measured in the survey (e.g., planting depth), differences in pest populations, soil properties, or spatial variations in weather. The existence of spatial correlation in model errors suggests that at least part of the unexplained variance in yields was attributable to factors that exhibit spatial autocorrelation across the landscape. The management variables that were recorded in this study did not generally exhibit spatial autocorrelation (Moran's I test, $p > 0.1$); therefore, one would expect nonmanagement variables to explain part of the model residuals. Moreover, the spatial patterns of these residuals did not correspond to existing maps of soil type (not shown). Therefore, we suspect that at least part of the model error is due to factors such as soil characteristics other than type, weather conditions, and spatially dependent biological processes such as weed competition or rust infestation. Future work is needed to explore these factors in more detail. However, our results suggest that such processes are likely to explain a small fraction of yield variability relative to the major management factors.

As with all empirical models, the interpretation of the results above must be qualified with two caveats. First, it is always possible that an unmeasured, latent variable has introduced bias into model results, even though we were careful to measure all factors that we considered potential explanatory variables. Second, one must recognize the possibility that the inferred importance of a measured variable is not due solely to a direct effect of that variable on the response but in part to the effect of another variable that covaries with the first.

For these reasons, it is often suggested that empirical models should be used only for prediction of unobserved quantities and not for modeling response of systems to change (i.e., extrapolation). However, in situations where direct experimental manipulation is impractical, empirical models play an important if not complete role in uncovering cause–effect relationships. In particular, empirical models of spatial crop yield variability can provide valuable information on the relative importance (or unimportance) of known mechanisms at the field or regional scale (Landau et al., 2000; Corwin et al., 2003). In these cases, an important distinction should be made between correlation and causation, and model interpretation should be guided by whether model variables and their coefficients are physically reasonable, as they were in this study.

CONCLUSIONS

We used remote sensing to generate spatially extensive estimates of wheat yield at 30-m resolution, which were then combined with 2 yr of survey data to assess important sources of spatial variability in yields. The results indicate that, for the Yaqui Valley, causes of yield variability differ considerably between years. In 2001, which was an unusually cool year, N fertilizer was the most important yield determinant despite very high average rates of N application ($>$260 kg N ha^{-1}). In 2003, a more typical year climatically, the timing of water application represented the most important source of yield variability.

Overall, it appears that management variations, as opposed to soil or climatic constraints, drive the majority of yield variability in the Yaqui Valley. This conclusion is consistent with previous interpretations based on analysis of spatial yield patterns (Lobell et al., 2002) and has the important implication that the yield gap can be significantly reduced through management changes. For example, the average yield in the highest management class was 0.84 t ha^{-1} higher than the Valley average in 2001 and 0.98 t ha^{-1} higher in 2003.

However, these results also indicate that strategies to improve yields through management must consider the role of climate variability. For example, N availability may constrain yields in one year, but increasing application rates may not make sense in the context of interannual climate variability (Lobell et al., 2004). Conversely, the climatic dependence of management impacts implies that seasonal weather forecasts would be useful for a wide range of management decisions, including those related to fertilizer, irrigation, planting date, soil preparation, and pest control.

The results presented here imply that improved fertilizer and water use are the most pressing management needs for increasing yields. It is interesting to note that timing of irrigation was more important than number of irrigations, suggesting that the efficient use of water is more important than total amount of water applied for yields in this region. Similarly, previous studies in this region have documented the improved N use efficiency attainable through better timing of N application (Matson et al., 1998). For both water and N, it therefore appears that efficiency of resource use plays as important a role, if not more, than total input amounts. Consideration of the economic and environmental costs associated with increased inputs places an even greater emphasis on the need for more efficient resource use (Matson et al., 1998; Cassman, 1999).

In general, an understanding of biophysical constraints to yield is only a first step toward improved management because food production is only one objective of agricultural activity. Farmers, for example, are most concerned with profit, and yield gains from higher fertilizer rates may not justify the associated increase in costs. An eventual goal of this research is thus to quantify all of the agronomic, economic, and environmental trade-offs associated with management changes. A quantitative understanding of yield controls is a critical step in this direction.

Beyond the Yaqui Valley, the results presented here have important implications for the design and interpre-

tation of studies aimed at understanding regional yield variations. For example, it is important to conduct surveys in multiple years to ensure that results are not specific to the (potentially unusual) climatic conditions of the survey year. The need for multiple surveys places a premium on cost-effective approaches to conducting yield surveys, such as achieved by replacing intensive field measurements of yield with remote-sensing estimates. In addition, commonly used linear regression models have important limitations when used to assess yield variations. Regression trees, which provide a simple means of capturing nonlinear relationships and variable interactions, appear to be a valuable tool for identifying regionally significant yield constraints.

Given the growing demand for food, the limited prospects for increasing yield potential, and the environmental consequences associated with overapplication of inputs, improved understanding of spatial variations in crop yields is greatly needed. Remotely sensed estimates of crop production provide a unique perspective that, when combined with field surveys, should enhance our ability to identify management priorities for improving regional production and/or reducing environmental impacts.

ACKNOWLEDGMENTS

We are indebted to Dagoberto Flores for conducting the field surveys and three anonymous reviewers for their helpful comments. This work was supported by a National Science Foundation Graduate Research Fellowship, NASA New Investigator Program grant NAG5-8709, and the Packard Foundation. This is CIW Department of Global Ecology Publication 83.

REFERENCES

Baez-Gonzalez, A.D., P. Chen, M. Tiscareno-Lopez, and R. Srinivasan. 2002. Using satellite and field data with crop growth modeling to monitor and estimate corn yield in Mexico. Crop Sci. 42:1943–1949.

Breiman, L., J. Friedman, R. Olshen, and C. Stone. 1984. Classification and regression trees. Wadsworth, Belmont, CA.

Calvino, P., and V. Sadras. 2002. On-farm assessment of constraints to wheat yield in the south-eastern Pampas. Field Crops Res. 74:1–11.

Cassman, K.G. 1999. Ecological intensification of cereal production systems: Yield potential, soil quality, and precision agriculture. Proc. Natl. Acad. Sci. USA 96:5952–5959.

Conover, W.J. 1999. Practical nonparametric statistics. John Wiley & Sons, New York.

Corwin, D.L., S.M. Lesch, P.J. Shouse, R. Soppe, and J.E. Ayars. 2003. Identifying soil properties that influence cotton yield using soil sampling directed by apparent soil electrical conductivity. Agron. J. 95:352–364.

De'ath, G., and K.E. Fabricius. 2000. Classification and regression trees: A powerful yet simple technique for ecological data analysis. Ecology 81:3178–3192.

Fischer, R.A. 1985. Number of kernels in wheat crops and the influence of solar-radiation and temperature. J. Agric. Sci. 105:447–461.

Flores, D., F. Carrión, and P.R. Aquino. 2001. El cultivo del trigo en el Valle del Yaqui: Cambios en los factores tecnológicos y socioeconómicos [Online]. Available at http://economics.cimmyt.org/Yaqui/Memoria/index.htm (verified 17 Sept. 2004). CIMMYT, Mexico, D.F.

Haining, R. 2003. Spatial data analysis: Theory and practice. Cambridge Univ. Press, Cambridge, UK.

Hastie, T., R. Tibshirani, and J. Friedman. 2001. The elements of statistical learning: Data mining, inference, and prediction. Springer, New York.

Ihaka, R., and R. Gentleman. 1996. R: A language for data analysis and graphics. J. Comput. Graph. Stat. 5:299–314.

Landau, S., R.A.C. Mitchell, V. Barnett, J.J. Colls, J. Craigon, and R.W. Payne. 2000. A parsimonious, multiple-regression model of wheat yield response to environment. Agric. For. Meteorol. 101: 151–166.

Lapen, D.R., G.C. Topp, E.G. Gregorich, H.N. Hayhoe, and W.E. Curnoe. 2001. Divisive field-scale associations between corn yields, management, and soil information. Soil Tillage Res. 58:3–4.

Lark, R.M. 2001. Some tools for parsimonious modelling and interpretation of within-field variation of soil and crop systems. Soil Tillage Res. 58:99–111.

Lobell, D.B., G.P. Asner, J.I. Ortiz-Monasterio, and T.L. Benning. 2003. Remote sensing of regional crop production in the Yaqui Valley, Mexico: Estimates and uncertainties. Agric. Ecosyst. Environ. 94:205–220.

Lobell, D., J. Ortiz-Monasterio, C. Addams, and G. Asner. 2002. Soil, climate, and management impacts on regional wheat productivity in Mexico from remote sensing. Agric. For. Meteorol. 114:31–43.

Lobell, D.B., J.I. Ortiz-Monasterio, and G.P. Asner. 2004. Relative importance of soil and climate variability for nitrogen management in irrigated wheat. Field Crops Res. 87:155–165.

Long, D.S. 1998. Spatial autoregression modeling of site-specific wheat yield. Geoderma 85:181–197.

Maas, S.J. 1988. Using satellite data to improve model estimates of crop yield. Agron. J. 80:655–662.

Matson, P.A., R. Naylor, and I. Ortiz-Monasterio. 1998. Integration of environmental, agronomic, and economic aspects of fertilizer management. Science 280:112–114.

Meisner, C.A., E. Acevedo, D. Flores, K. Sayre, I. Ortiz-Monastero, D. Byerlee, and A. Limon. 1992. Wheat production and grower practices in the Yaqui Valley, Sonora, Mexico. CIMMYT, Mexico, D.F.

Moulin, S., A. Bondeau, and R. Delecolle. 1998. Combining agricultural crop models and satellite observations: From field to regional scales. Int. J. Remote Sens. 19:1021–1036.

Ortiz-Monasterio, J.I. 2002. Nitrogen management in irrigated spring wheat. p. 433–452. *In* B. Curtis et al. (ed.) Bread wheat improvement and production. FAO, Rome.

Pingali, P.L., and S. Rajaram. 1999. Global wheat research in a changing world: Options and sustaining growth in wheat productivity. *In* P.L. Pingali (ed.) CIMMYT 1998–1999 world wheat facts and trends. CIMMYT, Mexico, D.F.

Plant, R.E. 2001. Site-specific management: The application of information technology to crop production. Comput. Electron. Agric. 30:9–29.

Plant, R.E., A. Mermer, G.S. Pettyfrove, M.P. Vayssieres, J.A. Young, R.O. Miller, L.F. Jackson, R.F. Denison, and K. Phelps. 1999. Factors underlying grain yield spatial variability in three irrigated wheat fields. Trans. ASAE 42:1187–1202.

Sadras, V., D. Roget, and G. O'Leary. 2002. On-farm assessment of environmental and management constraints to wheat yield and efficiency in the use of rainfall in the Mallee. Aust. J. Agric. Res. 53:587–598.

Shanahan, J.F., J.S. Schepers, D.D. Francis, G.E. Varvel, W.W. Wilhelm, J.M. Tringe, M.R. Schlemmer, and D.J. Major. 2001. Use of remote-sensing imagery to estimate corn grain yield. Agron. J. 93:583–589.

White, J.W., J.D. Corbett, and A. Dobermann. 2002. Insufficient geographic characterization and analysis in the planning, execution and dissemination of agronomic research? Field Crops Res. 76:45–54.

Wiegand, C.L., J.D. Rhoades, D.E. Escobar, and J.H. Everitt. 1994. Photographic and videographic observations for determining and mapping the response of cotton to soil-salinity. Remote Sens. Environ. 49:212–223.

Wiese, M.V. 1982. Crop management by comprehensive appraisal of yield determining variables. Annu. Rev. Phytopathol. 20:419–432.