Field Crops Research 221 (2018) 130–141

Contents lists available at ScienceDirect

Field Crops Research

journal homepage: www.elsevier.com/locate/fcr

Sifting and winnowing: Analysis of farmer field data for soybean in the US North-Central region

Spyridon Mourtzinis[a,*], Juan I. Rattalino Edreira[b], Patricio Grassini[b], Adam C. Roth[a], Shaun N. Casteel[c], Ignacio A. Ciampitti[d], Hans J. Kandel[e], Peter M. Kyveryga[f], Mark A. Licht[g], Laura E. Lindsey[h], Daren S. Mueller[i], Emerson D. Nafziger[j], Seth L. Naeve[k], Jordan Stanley[e], Michael J. Staton[l], Shawn P. Conley[a]

[a] Department of Agronomy, University of Wisconsin-Madison, Madison, WI 53706, USA
[b] Department of Agronomy and Horticulture, University of Nebraska-Lincoln, Lincoln, NE 68583-0915, USA
[c] Department of Agronomy, Purdue University, West Lafayette, IN 47907, USA
[d] Department of Agronomy, Kansas State University, Manhattan, KS 66506, USA
[e] Department of Plant Sciences, North Dakota State University, Fargo, ND 58108-6050, USA
[f] Iowa Soybean Association, Ankeny, IA 50023, USA
[g] Department of Agronomy, Iowa State University, Ames, IA 50011-1010, USA
[h] Department of Horticulture and Crop Science, The Ohio State University, Columbus, OH 43210, USA
[i] Department of Plant Pathology and Microbiology, Iowa State University, Ames, IA 50011, USA
[j] Department of Crop Sciences, University of Illinois, Urbana, IL 61801, USA
[k] Department of Agronomy and Plant Genetics, University of Minnesota, St. Paul, MN 55108, USA
[l] Department of Crop and Soil Sciences, Michigan State University, East Lansing, MI 48824, USA

ARTICLE INFO

Keywords:
Soybean
Survey data
On-farm yield
Statistical technique
Conditional inference tree

ABSTRACT

Field trials are commonly used to estimate the effects of different factors on crop yields. In the present study, we followed an alternative approach to identify factors that explain field-to-field yield variation, which consisted of farmer survey data, a spatial framework, and multiple statistical procedures. This approach was used to identify management factors with strongest association with on-farm soybean yield variation in the US North Central (NC) region. Field survey data, including yield and management information, were collected over two crop growing seasons (2014 and 2015) from rainfed and irrigated soybean fields (total of 3568 field-year observations). Fields were grouped into technology extrapolation domains (TEDs) that accounted for soil and climate variation and 9 TEDs were selected based on the number of fields needed to detect yield differences due to management as determined using power analysis. Average yield ranged from 2.5 to 5 Mg ha^{-1} across TEDs, with field yield distributions in half of the domains having a distributional peak that was close to maximum yields. Conditional inference trees analysis was chosen among 26 statistical procedures as the approach that best combines ability to detect and rank factors (and their interactions) with greatest influence on on-farm yield and relatively easy interpretation of results. Survey data from *ca.* 150 fields in each of the nine TEDs allowed us to identify key management factors influencing yields for an agricultural area that includes *ca.* 7 million ha sown with soybean. In five of the nine TEDs, highest yields were observed in early-sown fields. Other factors explaining on-farm yield variation were maturity group, and in-season foliar fungicide and/or insecticide application, but, in some cases, their influence on yield depended upon sowing date and water regime. While the approach proposed here cannot establish cause-effect relationships conclusively, it can certainly provide a focus to replicated field experiments in relation to which management factors to investigate. We believe that future agronomic studies based on farmer survey data can greatly benefit from *ex-ante* identification of most important TEDs (relative to crop area and production) as well as determination of minimum number of farmer survey data that needs to be collected from each of them based on expected yield differences and variability. The approach is generic enough to be applied in other crop producing regions as long as farmer data and associated climate and soil databases are available.

* Corresponding author.
E-mail address: mourtzinis@wisc.edu (S. Mourtzinis).

https://doi.org/10.1016/j.fcr.2018.02.024
Received 18 November 2017; Received in revised form 22 February 2018; Accepted 23 February 2018

1. Introduction

Average crop yields will need to increase substantially during the next 33 years to meet expected food demand increase while avoiding massive expansion of cropland area (Tilman et al., 2011; Alexandratos and Bruinsma, 2012; Grassini et al., 2013). This challenge can be achieved by increasing the rate at which best management practices are identified and adopted for a particular soil-climate context. Replicated field experiments are used in agricultural research to test new technologies and management practices. In these experiments, researchers selectively manipulate a production factor and, by comparing final yield against the yield of a "control" treatment, the magnitude of the yield response and its economic profitability are assessed. A limitation of this approach is that it often examines the effect of management practices at a small number of sites and years due to practical constraints (*e.g.*, costs, logistics, *etc.*). Hence, extrapolation of their findings is typically confined to a narrow range of environments. Likewise, field experiments cannot test the effect of a large number of production factors (and their interactions) on yield due to the large number of plots that would be needed. And, finally, the management selected as "background" for these experiments (*e.g.*, sowing date, tillage method) will also influence crop responses to a given technology or management. Given these limitations, it is relevant to search for alternative, cost-effective approaches that provide an indication of the management practices that perform best for a given climate-soil context.

Farmer survey data can be utilized as a cost-effective source of information to identify yield constraints and fine-tune management practices so that these yield limitations can be ameliorated or eliminated (*e.g.*, Calvino and Sadras, 2002; Sadras et al., 2002; Lobell et al., 2005; Tittonell et al., 2008). An advantage of using farmer data is that it allows examination of opportunities for yield increase within the range of current management practices that are both cost-effective and logistically feasible in farmer fields. Another advantage of using farmer data is that, if surveyed fields are properly contextualized relative to their biophysical environment, it is possible to explore and quantify management × environment interactions (Rattalino Edreira et al., 2017). Such assessment would allow identification of suites of management practices that perform best for a given environment and provide a focus to traditional, costly field experiments so that they can target those management practices with the most likely impact on crop productivity and input-use efficiency.

Statistical analysis of farmer self-reported data poses challenges that need to be addressed to make meaningful and unbiased inferences. For example, in field experiments, different levels of a given management or input are assigned to experimental units. These experimental units are carefully selected based on their similarity, in order to avoid confounding factors influencing yield and to minimize the error variance. Each treatment level is applied to several experimental units ('replicates') to obtain an estimate of average yield and its variation. In contrast, farmer data do not follow an experimental design and lack random allocation of experimental units and replication. Variation in soil, weather, and management practices across fields results in minimal control over error variance. Several management practices (or inputs) may be applied simultaneously, leading to multi-collinearity, making interpretation of results more challenging (Hastie et al., 2001). Additionally, it may be the case that a given management practice does not appear to be significantly associated with yield simply because that practice has already been widely adopted across fields (*e.g.*, cultivars with herbicide-resistance traits). Despite all these limitations, farmer data have the potential to give an indication of the most important yield-limiting factors in a given region, which can, in turn, then be tested in more detailed field trials to experimentally confirm cause-effect relationships.

We argue here that proper analysis of farmer field data, when evaluating the influence of management factors on yield, requires: (i) a biophysical spatial framework to cluster fields into groups with relatively similar climate and soil, (ii) use of appropriate statistical methods that can handle the nuances associated with the structure of farmer survey data and to identify management interactions, and (iii) a deep agronomic knowledge and understanding of the cropping system context to interpret results and translate them into practical recommendations. Application of a spatial framework to identify causes of yield gaps has been addressed in a previous study (Rattalino Edreira et al., 2017). A major limitation of this previous study, as well as other studies looking into the causes of yield gaps (*e.g.*, Mercau et al., 2001, Sadras et al., 2002; Grassini et al., 2011, 2015; Silva et al., 2016), is that the analysis was limited to a comparison of management practices between high- versus low-yield fields or regressions between yield and individual or multiple management practices for a given climate-soil domain, without an explicit attempt to rank the importance of each management practice based on its influence on yield and to identify interactions.

In the present study, we addressed the second requirement listed above, that is, the use of a proper statistical technique to identify and rank management factors (and their interactions) influencing soybean yield in farmer fields. We focused on soybean fields in the North Central US region, which accounts for *ca.* 85% of US soybean production and *ca.* 30% of global production (FAOSTAT, 2016; USDA-NASS, 2016). The objective of this study was to utilize self-reported farmer data and multiple statistical techniques, together with a spatial framework, to identify the management practices with greatest influence on rainfed and irrigated soybean yields across diverse climate and soil conditions.

2. Materials and methods

2.1. Database description

Soybean yield and management practices data were collected from 3568 fields sown with soybean in 2014 and 2015 across 10 states in the US NC region: Iowa (IA), Illinois (IL), Indiana (IN), Kansas (KS), Michigan (MI), Minnesota (MN), Ohio (OH), North Dakota (ND), Nebraska (NE), and Wisconsin (WI) (Fig. 1). Detailed description of the database is provided elsewhere (Rattalino Edreira et al., 2017). The majority of surveyed fields were non-irrigated, except in Nebraska, where there were both rainfed (34%) and irrigated fields (66%) located within the same region. Maize was the predominant prior crop (88% of total fields). Average regional yield represents *ca.* 22 (rainfed) and 13% (irrigated) of the estimated yield potential, indicating a relatively small (but still exploitable) room for increasing farmer yields through fine tuning of current management practices (Rattalino Edreira et al., 2017).

Farmers reported data on field location, average yield (adjusted to 13% moisture content), and management practices, including sowing date, seeding rate, row spacing, variety name, tillage method, drainage system, total irrigation amount (for irrigated crops), seed treatment, fertilizer inputs, lime, manure, and pesticides (Table 1). Farmers also reported incidence of other field adversities such as pests, diseases, weeds, iron deficiency chlorosis, hail, waterlogging, and frost. Data were subjected to quality control to remove erroneous entries. Likewise, fields subjected to unmanageable field adversities (*e.g.*, hail, frost, flooding) leading to substantial yield losses were excluded from the analysis. To do this, fields reported as affected by any of the aforementioned adversities were grouped within regions with similar soil and climate (see Section 2.2), and we excluded those that fall below the 25th percentile of the yield data distribution within each region-year. To summarize, we excluded data from fields affected by unmanageable adversities *and* that fell below the 25th percentile of the yield distribution in each climate-soil domain; these data were excluded from all the statistical analyses, as well as tables and figures presented here. We did not exclude fields that suffered from drought, heat stress, temporary waterlogging, or disease, insect or weed pressure. After quality control, the database contained data from a total of 3216 fields sown with soybean in 2014 and 2015 (92% of total surveyed fields). Fields were

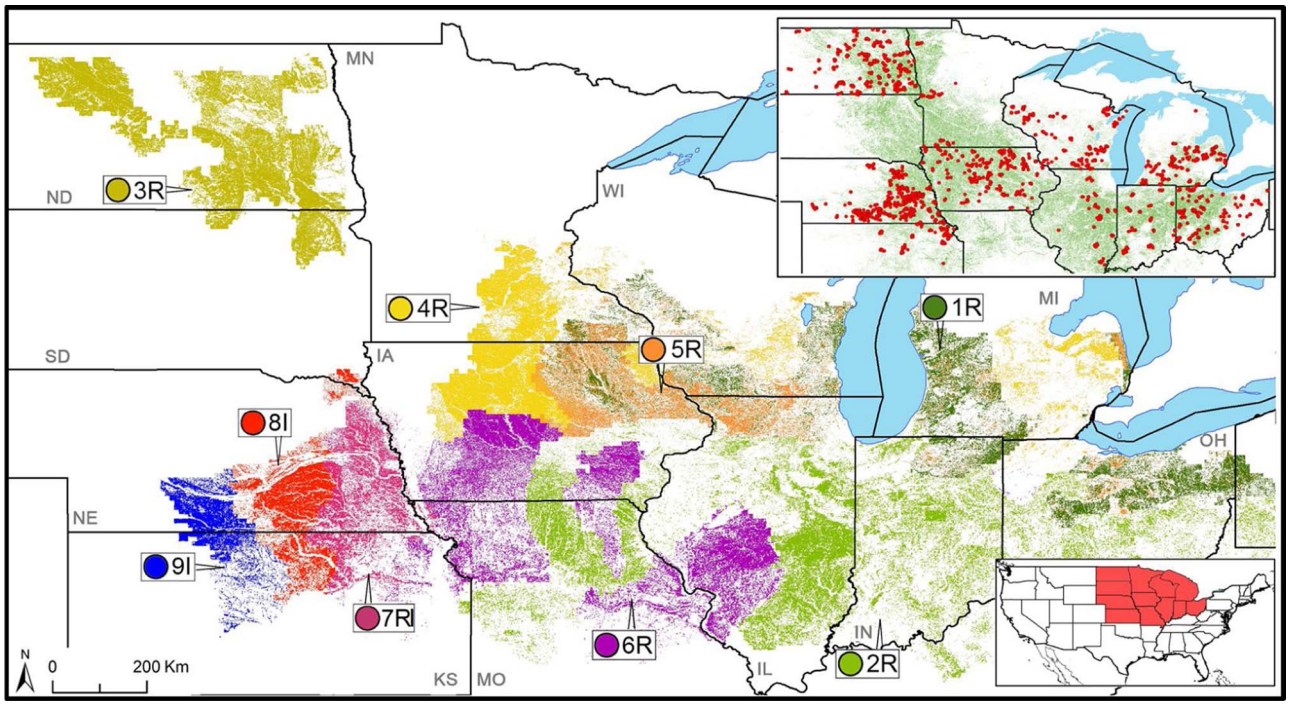

Fig. 1. Map of the surveyed region showing nine technology extrapolation domains (TEDs). Each TED is shown with a different color. Upper inset: soybean harvested area in 2015 shown in green; (USDA-NASS, 2016) and location of 3568 surveyed soybean fields (red dots). Bottom inset: location of US NC region within the conterminous US. Note: R = rainfed fields; I = irrigated fields; RI = rainfed and irrigated fields within the TED. Taken from Rattalino Edreira et al., 2017. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
List of variables included in the statistical analyses.

Variable	Units (or classes)
Sowing date	Julian day
Cultivar maturity group	unitless
Foliar fungicide	yes/no
Foliar insecticide	yes/no
Seed treatment	yes/no
Fungicide seed treatment	yes/no
Insecticide seed treatment	yes/no
Tillage method	no-till/reduced/conventional
Seeding rate	seeds m^{-2}
Starter fertilizer	yes/no
Residue management [a]	none/grazed/harvested
Row spacing [b]	narrow/intermediate/wide
Potassium fertilizer	kg K_2O ha^{-1}
Phosphorous fertilizer	kg P_2O_5 ha^{-1}
Lime	yes/no
Manure	yes/no
Topsoil pH (0–30 cm)	unitless
Subsoil pH (30–150 cm)	unitless
Topography wetness index	unitless
Artificial drainage	yes/no
Soybean cyst nematodes	yes/no/unknown
Iron chlorosis deficiency	yes/no

[a] Plant residue left after harvest of previous crop.
[b] Narrow (≈18 cm), intermediate (≈38 cm), and wide (≈76 cm) row spacing.

grouped into narrow (≈18 cm), intermediate (≈38 cm), and wide (≈76 cm) row spacing (Table 1). Fields were classified based upon tillage method as (i) conventional (chisel and disk), (ii) reduced (strip-till, ridge-till, cultivator), and (iii) no-till. Fields were classified depending upon seed treatment (ST) as untreated, fungicide-ST, and insecticide-ST. Because a substantial number of surveys did not indicate if ST included fungicide, or insecticide or both, we also used a generic ST class for the statistical analysis. Fields were also classified according to presence or absence of artificial drainage system such as new systematic tiles, old clay tiles, *etc*.

Mean pH was calculated for the topsoil (0–30 cm) and subsoil (30–150 cm) in each field from the SSURGO database (https://websoilsurvey.sc.egov.usda.gov/App/WebSoilSurvey.aspx) (Table 1). Mean pH for a given field was derived from the pixel pH distribution within each field (*ca.* 9 pixels per field). Using the mode instead of the mean would have resulted in a negligible change in the calculated field-level pH (< 1%). To account for differences in slope and terrain across a field, which could influence the crop water balance and final seed yield, we calculated the topography wetness index (TWI) for each field (Table 1). TWI has been used to characterize the potential for surface run-off and run-on in landscapes; hence, it can be used indirectly to assess the influence of field topography on crop productivity (Moore et al., 1993). High values are associated with flat terrain whereas smaller values are associated with more uneven fields (*e.g.*, fields with slopes). TWI is usually correlated with other soil attributes, including soil organic matter, soil texture, and phosphorous content; hence, higher TWI values are generally associated with more productive soils. TWI was calculated using the rsaga.wetness.index package in R (R development Core team, 2016) using the 30-m resolution National Elevation Dataset (USDA:NRCS:Geospatial Data Gateway; https://datagateway.nrcs.usda.gov/). Mean TWI for a given field was derived from the pixel TWI distribution within each field.

2.2. Field clustering

Fields were aggregated in clusters based on their biophysical properties using a technology extrapolation domain (TED) spatial framework (Rattalino Edreira et al., 2017; http://www.yieldgap.org/web/guest/cz-ted). Briefly, TED framework delineates regions based on: (i) annual total growing degree-days (10 classes), (ii) aridity index (10 classes), (iii) annual temperature seasonality (3 classes), and (iv) plant-available water holding capacity in the rootable soil depth (10 classes; 50-mm class interval). Each TED corresponds to a specific combination of the four aforementioned parameters. For our analysis, we selected

fields located within nine TEDs (Fig. 1). These TEDs included only rainfed fields (1R, 2R, 3R, 4R, 5R, 6R), only irrigated fields (8I and 9I), and both (7R-I). These TEDs portray well the range of climates and soils within the US NC region, including 7 million ha annually sown with soybean, which represent 21% of US soybean area. Detailed description of the field clustering by TEDs is provided elsewhere (Rattalino Edreira et al., 2017). Because water supply and management is very different between irrigated and rainfed fields, we treated 7R and 7I separately for the descriptive analysis. However, for the statistical analysis, we pooled the data from irrigated and rainfed fields to identify interactions between water regime and management practices. Because not all of the 3568 surveyed fields were located within one of the selected TEDs, our analysis used data from a subset of 1373 fields. There were more than 98 fields per TED, with an average of 153 fields per TED. This number of fields per TED represented a good compromise between maximizing the number of TEDs and having a reasonable number of fields per TED to detect yield differences due to management practices (see Sections 2.3 and 3.1).

2.3. Statistical analysis

2.3.1. Power analysis to establish number of fields required to detect yield differences

Ex-ante power analysis was used to determine the number of fields (sample size) needed to detect statistically significant yield differences due to management practices with a reasonable high level of confidence. Here, sample size was evaluated by power analysis based on realistic estimates of expected yield differences and yield variability for rainfed and irrigated soybean fields using SAS v.9.4 software (SAS Institute Inc., 2016). Different scenarios of expected yield difference between levels of binary variables and standard deviation of yield were explored. Our power analysis was constrained to agronomically-relevant ranges of seed yield variation and yield responses, which were chosen based on the range of yield variability across TEDs, measured with the standard deviation (SD), and yield differences due to management practices reported in the literature for soybean. For every yield difference and standard deviation combination, 500 random samples with different sample size (from 10 to 800 fields) and normally distributed data were created. Then, each sample was evaluated at 5% significance level using one-way analysis of variance. The power of each sample size for every combination of input parameters was the proportion of times that a given yield difference was detected at 5% significance level.

2.3.2. Statistical analysis to identify drivers of on-farm yield variation

The second part of the statistical analysis involved the use of multiple statistical procedures to identify the management and soil variables with the strongest influence on yield within each TED (Table 2). A total of 26 statistical procedures were used. These procedures utilize variable selection features and are commonly used in studies with unstructured data (*e.g.*, observational studies) that contain multiple independent variables. Additionally, regression procedures, such as LASSO (least squared shrinkage operator) and elastic net, have desirable properties that can mitigate data multi-collinearity issues (Zou and Hastie, 2005; Dormann et al., 2013).

For each TED, 22 independent variables (Table 1) were ranked in descending order based on the frequency in which each variable was identified as statistically significant across the 26 statistical models. Hence, if a given variable was detected as statistically significant across all fitted models, the frequency would sum up to 26. Ranking the variables using a weighted sum based on their similarity (*e.g.,* stepwise, backward elimination and forward selection) would have resulted in very similar ranks to those obtained with our simple frequency sum.

Regression trees analysis has been used in previous studies to identify yield constraints in farmer fields located within small geographic regions (*e.g.*, Lobell et al., 2005; Tittonell et al., 2008; Ferraro et al., 2009). This method does not have assumptions relative to data distribution, with appealing features for survey data analysis, including automatic variable selection, interpretability of interactions between variables, and ability to handle missing data (Hastie et al., 2001). It can handle categorical and continuous explanatory variables without statistical distribution assumptions, it is robust in the presence of outliers, multicollinearity, and heteroskedasticity, and can reveal interactions among factors. Nevertheless, this approach has been criticized for lack of concept of statistical significance (Mingers, 1987), data overfitting, and selection bias towards covariates with many levels and many missing observations (Hothorn et al., 2006). Conditional inference trees, which is the focus of the third part of our analysis, have been proposed as an alternative to regression trees as the former overcome the bias and overfitting issues by utilizing the distributional properties of the data (Hothorn et al., 2006). This method estimates a relationship among several variables by binary recursive partitioning in a conditional inference framework using distributional properties of variables (Hothorn et al., 2006).

The conditional inference tree analysis was performed using the *partykit* package in R (R development Core team, 2016). Application of conditional inference trees to analyze combined data from multiple experiments has been described in Mourtzinis et al. (2018). Briefly, the algorithm tests the null hypothesis of independence between the response variable (*i.e.*, yield) and any of the input variables (*i.e.*, management and field variables; see Table 1). The algorithm selects the input variable with strongest association, measured by a *p-value*, with the response variable. Then, a binary split is implemented in the

Table 2
List of the 26 statistical methods, and associated criteria, used to identify soil and management practices with greatest influence on farmer soybean yields within each technology extrapolation domain.

Method	Selection criteria
Stepwise	AIC
	AIC-1000 model average-80% random sampling
	as previous with 20% partition for validation
Backward elimination	AIC
	AIC-1000 model average-80% random sampling
	As previous with 20% partition for validation
Forward selection	AIC
	AIC-1000 model average-80% random sampling
	As previous with 20% partition for validation
Least angle regression (LAR)	AIC
	AIC-1000 model average-80% random sampling
	As previous with 20% partition for validation
Least squared shrinkage operator (LASSO)	AIC
	AIC-1000 model average-80% random sampling
	As previous with 20% partition for validation
Group LASSO	AIC
	AIC-1000 model average-80% random sampling
	as previous with 20% partition for validation
Adaptive LASSO	AIC
	AIC-1000 model average-80% random sampling
	As previous with 20% partition for validation
Elastic net	AIC
	AIC-1000 model average-80% random sampling
	As previous with 20% partition for validation
Random Forest regression	Number of trees = 1000, number of permutations = 1000
Conditional inference trees	See Section 2.3.2.

AIC: Akaike information criterion.

selected input variable (node) and all steps are recursively repeated. The terminal node accounts for the final subset of fields. The result of this procedure is a graph that looks like a tree. The sizes of intermediate and terminal nodes are defined according to pre-specified criteria. In this analysis, the criterion for the independence test was based on univariate *p-values* (alpha = 0.05). To ensure adequate power, besides the *p-value*, we ensured that each intermediate node account for a minimum of 33% of total observations, and a terminal node should contain a minimum of 11% (one third of observations in an intermediate node). All these criteria must be met at every step of the algorithm so that a variable can qualify for a split. To avoid overfitting and enhance interpretability, the maximum tree depth was set to 10 nodes. Explanatory power of the conditional inference tree was calculated with the coefficient of determination (R^2) and root mean square error (RMSE). Sensitivity of the results due to the chosen criteria was assessed by repeating the analysis with different combinations of minimum number of fields per intermediate node (20–40% of total observations), and tree depth up to 20 nodes. To avoid overfitting and development of low-power models, terminal nodes were not allowed to contain less than 11% of total fields as this would result in nodes with low number of fields (< 5). In all models, regardless of the chosen criteria, the identified primary and secondary important variables and their thresholds were identical. Only the tertiary and least important variables varied in a few larger trees (> 10 nodes) and the goodness of fit of these expanded models was not substantially improved (< 5%).

To identify putative factors with greatest influence on on-farm soybean yield, and their interactions, all management practices reported by farmers, as well as pH and TWI values, were included in the conditional inference tree analysis. The analysis was performed separately for each TED. Data were pooled across years for the analysis because TEDs explained 31× more of the variation in farmer yields than year or TED × year interaction (Rattalino Edreira et al., 2017). TEDs accounted for variation in plant-available water holding capacity in the rootable soil depth, which is inherently correlated with other soil parameters such as soil texture, soil organic matter, and soil depth. In contrast, plant-available water holding capacity is not necessarily correlated with pH and TWI, which justifies inclusion of these two parameters as independent variables in the analysis. Furthermore, these two parameters are potentially manageable in a given field through pH correction and artificial drainage. Soybean varieties were described relative to their maturity group (MG), using the latter as one of the independent variables in our models. Maturity groups are usually designated using triple zero, double zero, zero and Roman numerals from I to X for very short- and long-season varieties, respectively. We did not attempt to quantify the influence of specific varieties on yield given the multitude of varieties (*ca.* 2000) sown across farmer fields (which will would leave us with very few observations per variety and per TED) and the rapid varietal turnover over time, which would make any inference about variety become obsolete very quickly.

3. Results

3.1. Required number of fields per TED to detect yield differences due to management

Sample size needed to reach power = 0.8 for different expected yield differences and yield variability is shown in Fig. 2. High and low SD lines corresponded to hypothetical environments with respective high and low yield variation. For example, the SD for selected TEDs in our study ranged from 510 (irrigated and favorable rainfed environments) to 790 kg ha^{-1} (rainfed environments), with an average of 600 kg ha^{-1}. The magnitude of the yield difference reflects the expected yield response to a management factor or applied input. For example, a previous study indicated that, on average, foliar fungicide and/or insecticide application increased yield by *ca.* 300 kg ha^{-1}, while a 4-week delay in sowing after end of April would reduce yield *ca.*

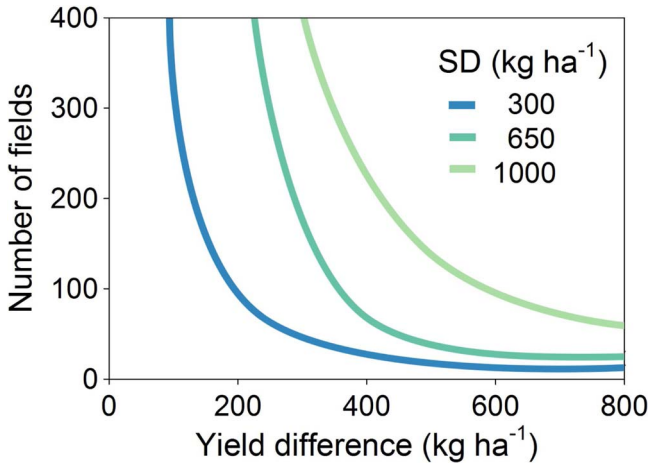

Fig 2. Number of surveyed farmer fields needed to identify a given yield difference as statistically significant (power = 0.8) for three scenarios of yield variation, the latter quantified with the standard deviation (SD). Our number of fields per TED ranged from 98 to 201, with SD ranging from 510 to 790 kg ha^{-1}.

600 kg ha^{-1} (Rattalino Edreira et al., 2017). As indicated previously, each of our selected TEDs included > 98 fields, with an average of 153 fields per TED. Such a sample size seems adequate to detect a significant yield difference with power ≥0.8 for most of the variables considered in the analysis that are expected to influence farmer yields, especially in TEDs with low yield variation (Fig. 2). In contrast, our analysis will have less power for testing yield differences on their statistical significance in environments with high yield variation. Fortunately, our selected TEDs corresponded to the first category of environments, as indicated by the small within-TED coefficients of variation for farmer yield (see Fig. 4).

3.2. Soybean management in the US NC region

Descriptive analysis for soybean management practices in each TED was summarized in Table 3 and Fig. 3. The study region was characterized by diversity of soil types, weather, and management practices. Except for the alkaline subsoil in TEDs 3 and 9 (pH ≈ 8), average pH in the topsoil and subsoil ranged between 6 and 7.5, with the subsoil exhibiting slightly higher pH (Fig. 3 C-D). Higher TWI values in fields in TEDs 2, 7I, 8 and 9 indicated a smaller run-off potential and favorable soils compared to fields in other TEDs (Fig. 3H). Topsoil and subsoil pH and TWI varied greatly across fields within some of the TEDs (*e.g.*, TEDs 1 and 6), which further justified their inclusion as independent variables in the stastistical analysis. Average sowing date varied by up to 2 weeks among TEDs, from early-May to late-May in the southern (TED 2, 9) and northern (TED 3) regions, respectively (Fig. 3A). Most varieties sown in farmer fields belong to MGs 2 and 3, except for fields located in the north-west region (TED 3; MGs 0 and 1) (Fig. 3B). Narrow (≈18 cm) and intermediate (≈38 cm) row spacing prevailed across TEDs located in rainfed production environments; in contrast, wider row spacing (≈76 cm) was dominant in irrigated fields (Table 3). Seeding rates ranged from 35 to 45 seeds m^{-2} (Fig. 3G), which, given a typical emergence rate of ca. 85–90% in soybean (Gaspar et al., 2017), indicate that seeding rates used by farmers are much higher than those required to achieve a plant density that maximize yield (27–32 plants m^{-2}; De Bruin and Pedersen, 2008). Higher seeding rates (*ca.* 10%) were observed in the eastern (TEDs 1 and 2) and western fringes (TEDs 3, 8 and 9) of the US NC region.

Applied fertilizer amounts (in fields that received fertilizer) ranged from 5 to 245 kg ha^{-1} (P_2O_5) and from 10 to 340 kg ha^{-1} (K_2O), respectively, with rates increasing following a west-east gradient (Fig. 3E–F). Starter N fertilizer (*i.e.*, a small N fertilizer application at sowing) was rarely applied in fields located in the central and eastern

Table 3
Description of management practices across technology extrapolation domains.

Production factor (% fields)	Technology Extrapolation Domains (TEDs)									
	1R	2R	3R	4R	5R	6R	7R	7I	8I	9I
Inputs										
Seed treatment	83	95	95	92	89	92	86	98	90	81
Foliar fungicide	20	38	11	40	47	39	20	24	20	19
Foliar insecticide	19	36	40	43	40	24	18	19	16	18
Starter N fertilizer	7	0	39	5	6	3	14	10	11	18
Lime	10	23	0	15	4	16	16	10	3	0
Manure	12	12	0	10	11	16	4	12	0	0
Field & crop management										
Artificial drainage	69	73	36	88	88	83	20	12	4	18
Residue management:										
Grazed	0	1	1	1	0	7	22	20	24	34
Harvested	6	23	1	3	2	0	15	17	16	19
Tillage method:										
No-till	60	44	20	48	59	52	72	67	50	90
Reduced till	17	19	25	25	19	20	14	13	17	5
Conventional till	23	37	55	27	22	28	14	20	33	5
Row spacing:										
Narrow (≈18 cm)	18	31	25	14	2	13	2	10	14	14
Intermediate (≈38 cm)	60	61	49	35	64	47	53	29	22	22
Wide (≈76 cm)	22	8	26	51	34	40	45	61	64	64
Adversities										
Iron chlorosis deficiency	26	0	20	28	2	25	0	0	1	4
Soybean cyst nematode:										
Yes	16	15	7	28	26	11	7	19	13	7
Unknown	38	50	41	38	34	62	40	22	19	10

parts of the US NC regions (< 10% of fields) (Table 3). About 10–20% fields in the western fringe of the region received N starter (TEDs 7, 8, 9), with this frequency increasing up to *ca.* 40% in the TED located in the north-west region (TED 3). This TED also has the largest frequency of tilled fields (55%). In contrast, no-till was the most common tillage method across the rest of the TEDs. Frequency of fields with artificial drainage followed the east-west gradient in seasonal precipitation, increasing dramatically from < 30% fields with artificial drainage systems in the western fringe of the US NC region to > 70% fields with drainage systems in the central and eastern regions (Table 3). Harvest and/or grazing of the residue left by previous maize crop were rarely practiced, except for 35–50% of fields located in western TEDs (TEDs 7, 8, and 9). Lime and manure were applied in < 20% of fields across TEDs, with most of these fields located in the central and eastern regions (Table 3).

Use of a seed treatment, which usually includes fungicide and/or insecticide, was a widespread practice across all TEDs, with seed being treated in > 80% of fields (Table 3). The frequency of fields that received foliar fungicide and/or insecticide applications ranged from 20 to 50% across TEDs and number of fungicide- and insecticide-treated fields were similar, in part because farmers tended to apply fungicide and insecticide together. A notable exception was the north-west TED (TED 3) where frequency of fields only treated with insecticides was much higher in relation with fungicide-treated fields (40 *versus* 11%). On average, 15% of surveyed fields reported incidence of soybean cyst nematode (SCN, *Heterodera glycines Ichinoche*); however, it was remarkable that *ca.* 35% of the farmers did not know (because of lack of soil testing) about the incidence of this pest in their soybean fields.

Examination of TED 7 allowed assessing differences in management practices between rainfed and irrigated fields within the same climate-soil context (Fig. 3, Table 3). For example, irrigated fields were sown (*ca.* 7 days) earlier and with earlier maturing varieties (0.5 MG difference) than rainfed crops. Likewise, a greater frequency of irrigated fields were tilled, received seed treatment and foliar fungicide, and used wider row spacing relative to rainfed fields located within the same TEDs. Higher TWI in irrigated *versus* rainfed fields indicated that the former were located in positions of the landscape with smaller surface runoff potential.

3.3. Yield variation among and within TEDs

Average soybean yield ranged from *ca.* 2.5 Mg ha^{-1} in short-season rainfed environments (TED 3) to *ca.* 5 Mg ha^{-1} in favorable irrigated areas (TEDs 8 and 9) (Fig. 4). Field-to-field variation within TEDs (quantified using the coefficient of variation [CV]) decreased with increasing average TED yield ($R^2 = 0.75$, $P < 0.05$). Yield variability within TEDs (range: 11–23%) was similar to the yield variation among TEDs (average CV = 16%), indicating that substantial field-to-field variation in yield remained after farmer fields were clustered based upon their TEDs. The high within-TED yield variation reflected the influence of other factors not accounted by the TEDs such as management practices, pH, and TWI.

For all TEDs, the distribution of soybean farmer yields was negatively skewed, although this pattern was more evident (skewness < −0.10) in half of the domains (Fig. 4). In other words, soybean field yields tended to have a distributional peak that was close to maximum yields, with skewing attributable to some low-yielding fields. Negatively skewed yield distributions have also been reported for other high-yield crop systems such as irrigated wheat grown in good soils in Yaqui Valley, northwestern Mexico (Lobell et al., 2005) and irrigated maize and soybean in Nebraska, USA (Grassini et al., 2011, 2014a). The degree of skewness was negatively correlated with the average TED yield ($R^2 = 0.35$, $P = 0.1$). To summarize, high-yield production environments exhibited smaller field-to-field yield variation, with a higher proportion of fields closer to maximum values compared with environments with lower yield.

3.4. Identification of candidate management factors with strongest influence on yield

Analysis of the soybean data with 26 different models helped us identify and rank the most important variables influencing soybean yield in each TED (Table 4). For example, row spacing, use of seed treatment, sowing date, topsoil pH, and TWI were the top-five ranked variables in TED 1. There were variables that consistently explained yield variation in a large number of TEDs. For example, sowing date

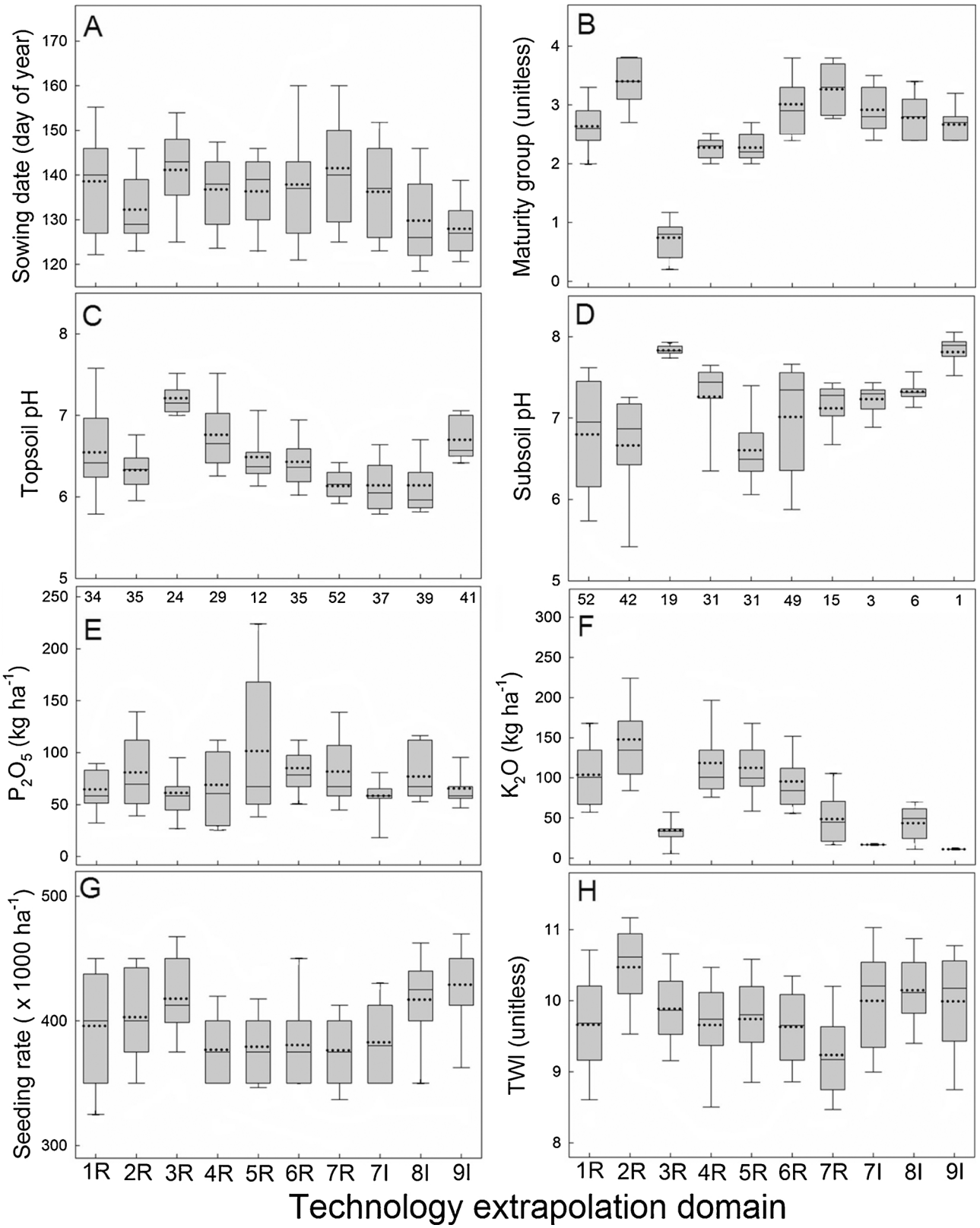

Fig 3. Description of rainfed (R) and irrigated (I) soybean farmer fields across technology extrapolation domains (TED). Variables include: (A) sowing date, (B), maturity group, (C) topsoil (0–30 cm) pH, (D) subsoil (30–150 cm) pH, (E) P_2O_5 fertilizer rate, (F) K_2O fertilizer rate, (G) seeding rate, and (H) topography wetness index (TWI). Boxes delimit first and third quartiles. Solid and dotted lines inside the box indicate median and mean, respectively. The upper and lower whiskers represent the maximum and minimum values, respectively. Values inside (E) and (F) indicate percentage of fields that received fertilizer application in each TED.

ranked amongst the top-fifteen variables in all TEDs and within the top-five variables in 6 of the 9 TEDs. Other variables that appeared as significant in most TEDs included the use of foliar fungicide and/or insecticide, seed treatment, TWI and pH, which appeared listed within the top-five variables in, at least, 4 of the 9 TEDs. Although there were similarities in the ranking among TEDs, there were also many

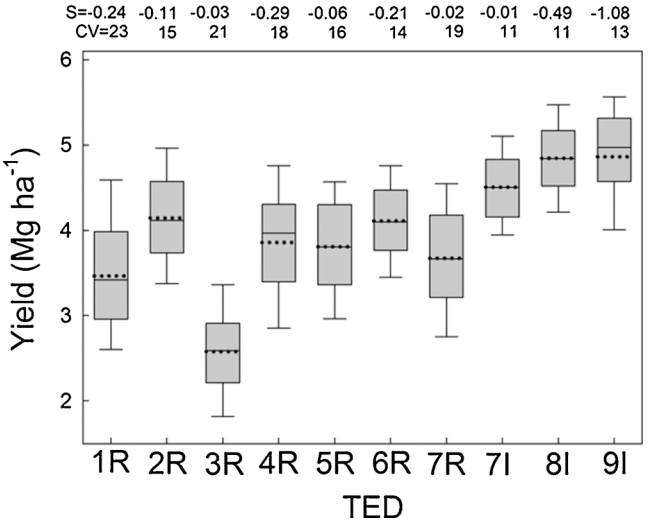

Fig. 4. Box plots for farmer soybean rainfed (R) and irrigated (I) yields across 10 technology extrapolation domains (TED). Boxes delimit first and third quartiles. Solid and dotted lines inside the box indicate median and mean, respectively. Upper and lower whiskers represent maximum and minimum values, respectively. Skewness (S) and coefficient of variation (CV) are shown.

differences. For example, sowing date ranked first in TED 4R, but the same variable ranked last in TED 3R. Such an abrupt difference in rank is consistent with the remarkable difference in yield response to sowing date reported for these same TEDs (-33 *versus* $-1\,kg\,ha^{-1}\,d^{-1}$, respectively) by Rattalino Edreira et al. (2017), which was attributed to differences in water balance during the pod setting (R3-R5) phase.

Although the models in Table 2 identified and ranked variables in terms of importance (see Table 3), it is difficult to reveal and quantify interactive effects of 2- and 3-way interactions of continuous variables and, perhaps more importantly, it is difficult to interpret these ranks. Additionally, because in unstructured datasets not all levels of a variable always exist for all levels of the interacting variables, the risk of extrapolation beyond the actual range increases and, thus, interpretation of the interactions can be misleading. Amongst all statistical methods evaluated here, conditional inference tree analysis appeared as the most robust approach to identify and rank factors (and their interactions) with greatest impact on farmer yields while facilitating the interpretation of the results. Another advantage of this method, in relation with other statistical techniques, is that fields were stratified so that interactions between management and/or soil factors were restricted within the actual range of management and/or soil properties. For example, while we are aware about the power of random forest regression to develop yield prediction models, this method resembled a "black box" approach because interpretation of model results was extremely difficult.

The conditional inference tree analysis performed for rainfed fields located within one of the eastern TEDs (TED 1) is shown in Fig. 5. Sowing date was the most important variable influencing farmer soybean fields. On average, fields that were sown between day of year (DOY) 119 and 123 (late April and early May) yielded $4\,Mg\,ha^{-1}$ (left terminal node), which is 9% higher than average yield in late-sown fields. In late-sown fields (DOY from 124 to 167, which corresponded to late May-early June), highest yields were achieved in fields with relatively higher TWI (> 9.2) and lower subsoil pH (< 7.2), but these yields were still lower than those reported for early-sown fields. The three variables of the explanatory model (sowing date, TWI, and subsoil pH) captured approximately one third of total yield variability within the TED ($R^2 = 0.29$).

Sowing date was also the most important factor influencing soybean yields in TEDs 4R, 5R, 6R, and 8I (Table 5, Fig. 6). Remarkably, late-sown fields could not achieve yields comparable to early-sown fields

Table 4
List of top 15 management and soil variables found to be strongest candidates at influencing soybean yields in each technology extrapolation domain (TED). Values next to each variable represent the number of statistical models (out of the 26 listed in Table 2) that detected a given variable as statistically significant on its influence on soybean field for a given TED.

TED 1R		TED 2R		TED 3R		TED 4R		TED 5R		TED 6R		TED 7RI		TED 8I		TED 9I	
Row spacing	25	Tillage	22	TWI	26	Sowing date	26	Sowing date	26	Sowing date	26	Irrigation	26	Sowing date	26	MG	25
ST	24	Row spacing	22	MG	25	Row spacing	26	ST	24	ST	25	Foliar fungicide	25	P fertilizer	24	TWI	25
Sowing date	24	Sowing date	21	Foliar fungicide	21	Subsoil pH	24	Foliar fungicide	24	Manure	24	Foliar insecticide	24	ST	24	Drainage	21
Topsoil pH	18	Foliar insecticide	21	Seeding rate	20	ST	24	Seeding rate	22	IDC	24	Lime	20	Foliar insecticide	22	Foliar fungicide	21
TWI	18	Subsoil pH	20	Drainage	19	Foliar fungicide	19	Topsoil pH	23	Starter fertilizer	23	TWI	19	Starter fertilizer	21	P fertilizer	21
Seeding rate	17	Seeding rate	20	Residue	16	K fertilizer	21	MG	16	Foliar insecticide	22	Sowing date	17	Subsoil pH	19	Residue	21
Subsoil pH	15	Seed treatment	20	Foliar insecticide	10	ST-fung	21	Subsoil pH	15	K fertilizer	21	Seeding rate	17	Seeding rate	15	Row spacing	21
ST-insect	14	K fertilizer	19	ST- fung	9	Tillage	20	K fertilizer	14	ST-insect	20	ST	14	Topsoil pH	14	Seeding rate	20
P fertilizer	13	ST- insect	19	P fertilizer	7	TWI	16	Foliar insecticide	13	Tillage	20	Topsoil pH	11	Row spacing	14	Sowing date	18
Drainage	12	P fertilizer	19	Subsoil pH	6	Residue	14	Row spacing	13	Lime	19	Subsoil pH	10	MG	13	ST	18
Manure	12	TWI	18	ST-insect	6	Drainage	13	TWI	13	Row spacing	19	ST- fung	10	ST- fung	13	Foliar insecticide	17
Starter fertilizer	12	Manure	17	Tillage	6	Foliar insecticide	13	Drainage	12	Foliar fungicide	19	MG	9	Drainage	12	Subsoil pH	17
Foliar insecticide	10	Drainage	15	Topsoil pH	5	Nematodes	13	ST-fung	12	P fertilizer	17	P fertilizer	8	K fertilizer	12	ST-fung	16
Lime	10	Topsoil pH	15	ST	5	Seeding rate	13	Tillage	12	TWI	15	Row spacing	7	ST-insect	11	Tillage	16
MG	10	Foliar fungicide	13	Sowing date	4	MG	12	Lime	11	Residue	14	ST- insect	7	Tillage	11	Topsoil pH	15

ST: treated seed; ST-fung: fungicide-treated seed; ST-insect: insecticide-treated seed; TWI: topographic wetness index; P: phosphorous; MG: maturity group; K: potassium; Residue: residue management before sowing; Drainage: artificial drainage; IDC: iron deficiency chlorosis.

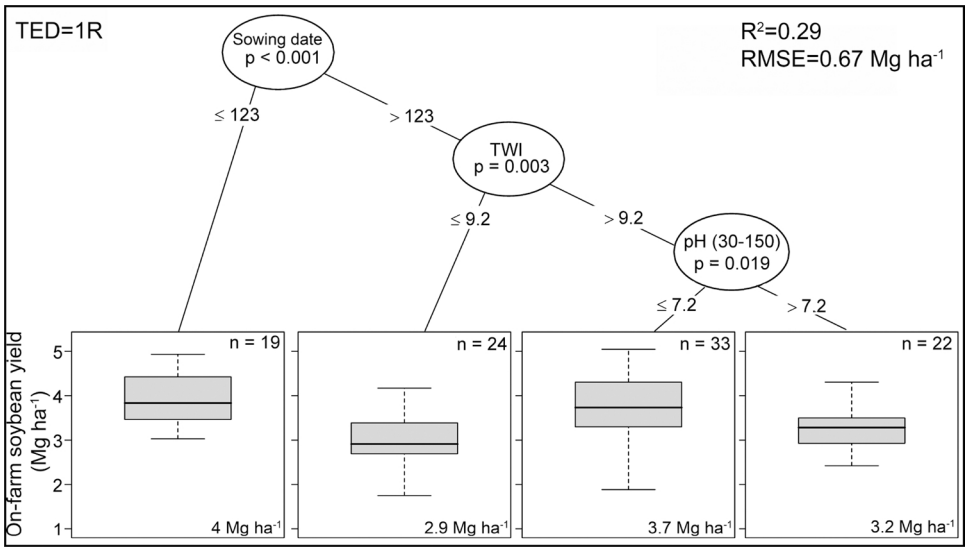

Fig. 5. Conditional inference tree for technology extrapolation domain (TED) 1R, located in the eastern region of the US North-Central region. In each boxplot, the central rectangle spans the first to the third yield quartiles. The solid line inside the rectangle shows the mean, which is also reported in the bottom right corner. The upper and lower whiskers represent the maximum and minimum values, respectively. TWI = topography wetness index.

Table 5
Summary of conditional inference trees in technology extrapolation domains (TEDs) 2R, 3R, 4R, 5R, 8I, and 9I. Values in brackets indicate number of fields (n) and average yield (Y, Mg ha^{-1}).

TED#	N1	N2	N3	N4	[n, Y]	R^2	RMSE (Mg ha^{-1})
2R	Row spacing (narrow)				[36,3.7]		
	Row spacing					0.10	0.6
	(intermediate, wide)				[82,4.3]		
3R	Foliar insecticide	TWI (9.7–11.7)			[58,2.8]	0.19	0.48
	(yes)	TWI (8.2–9.7)			[23,2.5]		
	Foliar	MG (0.9–1.5)			[23,2.9]		
	insecticide (no)	MG (0.08–0.9)	MG (0.08–0.6)		[39,2.3]		
			MG (0.6–0.9)		[58,2.4]		
4R	Sowing date	Foliar fungicide (no)			[39,4.1]	0.31	0.57
	(DOY 108–136)	Foliar fungicide (yes)			[39,4.4]		
	Sowing date	Row spacing			[52,3.4]		
		(narrow, medium)					
	(137–164 DOY)	Row spacing (wide)			[49,3.7]		
5R	Sowing date	Subsoil pH (5.5–6.5)			[41,4.3]	0.24	0.54
	(DOY 107–132)	Subsoil pH (6.6–8.1)			[23,3.9]		
	Sowing date	Foliar fungicide (no)	Sowing date (DOY 133–140)		[27,3.7]		
	(DOY 137–164)		Sowing date (DOY 141–161)		[39,3.4]		
		Foliar fungicide (yes)			[23,3.9]		
8I	Sowing date	Foliar insecticide (yes)			[22,5.2]	0.26	0.44
	(DOY 113–142)	Foliar insecticide (no)	Sowing date (DOY 113–124)		[50,5.0]		
			Sowing date (DOY 125–142)	TWI (8.3–10)	[18,4.5]		
				TWI (10.1–11.7)	[38,4.8]		
	Sowing date (143–175 DOY)				[50,5.0]		
9I	Seeding rate (30–36 m^{-2})				[15,4.3]	0.34	0.52
	Seeding rate	TWI (8.4–9.1)			[18,4.5]		
	(36–53 m^{-2})	TWI (9.1–11)	MG (2.4–2.7)		[45,5.2]		
			MG (2.7–4.2)		[25,4.9]		

Nth: node number; TWI: topography wetness index; MG: maturity group; DOY: day of year.

under any suite of management practices and soil and terrain parameters. Foliar fungicide or insecticide was also identified as management factors increasing soybean yield in 5 of 9 TEDs (Fig. 7, Table 5). Higher yields were also generally related to high TWI, which may reflect a more favorable position in the landscape in relation to crop water supply and likely better soil quality (see Section 3.2). Other management factors influencing yield in at least one TED were row spacing, maturity group, tillage method, and seeding rate (Figs. 6 and 7, and Table 5).

Conditional inference trees also allowed us to capture M × E interactions. For example, MG was a significant secondary (TEDs 3 and 7) and tertiary key management practice (TEDs 3 and 9). In the short-season environment of TED 3, higher yields were associated with late MGs (Table 5). This finding was not biased by the latitudinal distribution of MG varieties within TED 3 as the influence of MG persisted even when the analysis was conducted separately for the southern and northern portions of this TED. In contrast, in favorable irrigated environments (TEDs 7 and 9), higher yields were achieved with early MGs (Fig. 7, Table 5). These findings are consistent with Specht et al. (1986, 2001), who noted that Midwestern U.S. full-season maturity cultivars in rainfed environments usually yield better than earlier-maturing ones, but generally yield less under irrigation. Drought can shorten reproductive development in the early-maturing cultivars aligning those stages with the hotter part of the growing season, which tends to exacerbate the impact of water deficit. Our analysis also revealed an interesting interaction between presence of soybean cyst nematode (SCN) and tillage: SCN led to lower yields in TED 6 (Fig. 6), but this yield reduction was 6% higher in no-till *versus* tilled fields.

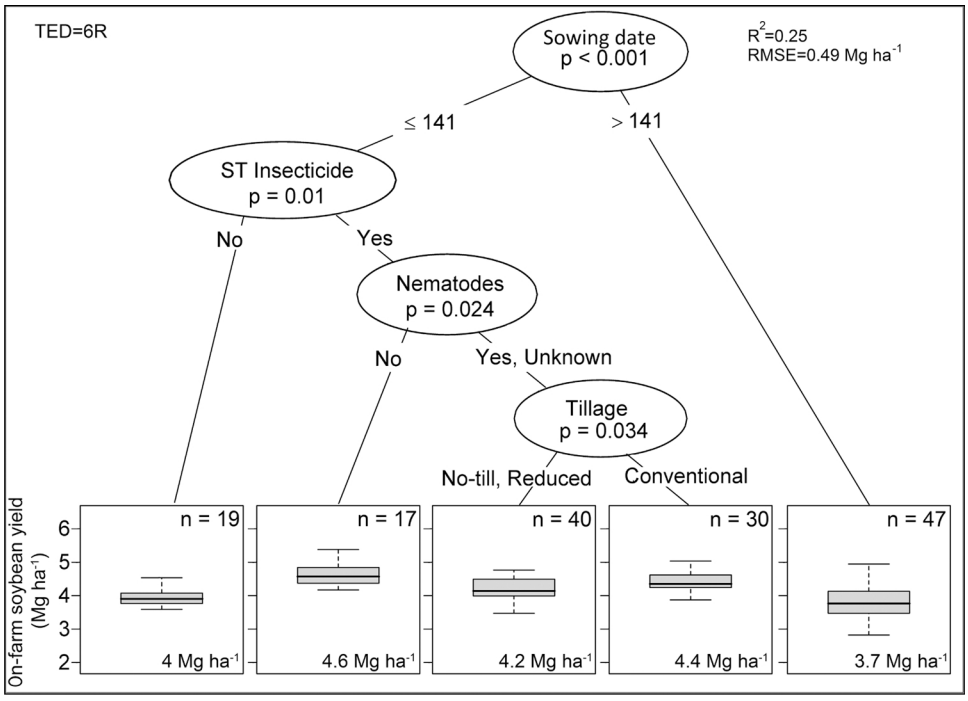

Fig. 6. Conditional inference tree for technology extrapolation domain (TED) 6R located in the southern fringe of the US North-Central region.. In each boxplot, the central rectangle spans the first to the third yield quartiles. The solid line inside the rectangle shows the mean, which is also reported in the bottom right corner. The upper and lower whiskers represent the maximum and minimum values, respectively.

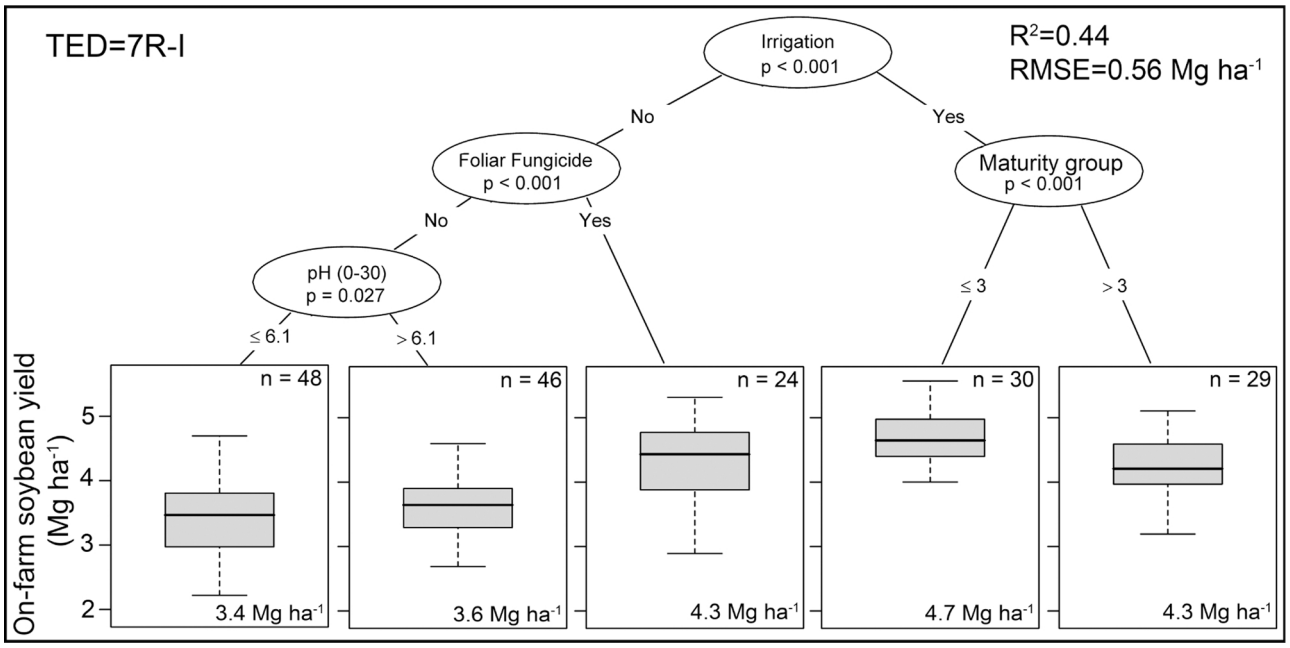

Fig. 7. Conditional inference tree for technology extrapolation domain (TED) 7RI, located in the western fringe of the US North-Central region, and which includes both rainfed and irrigated soybean fields. In each boxplot, the central rectangle spans the first to the third yield quartiles. The solid line inside the rectangle shows the mean, which is also reported in the bottom right corner. The upper and lower whiskers represent the maximum and minimum values, respectively.

4. Discussion

Analysis of farmer survey data using multiple statistical methods and a spatial framework allowed us to identify the most critical management factors explaining field-to-field yield variation in major US soybean producing areas. It was remarkable that a reasonable number of TEDs (9) and number of fields per TED (*ca.* 150) was sufficient to identify the most important yield-limiting factors for an agricultural area that includes 7 million ha sown with soybean, which, in turn, represent 21% of US soybean area. A larger number of fields would have been needed for harsher rainfed environments with very high yield variation. We note, however, that these environments typically account for a smaller share of regional and national crop production. Overall, we believe that future agronomic studies based on farmer survey data can greatly benefit from *ex-ante* identification of most important TEDs in relation to crop area and production as well as determination of the minimum number of farmer survey data that needs to be collected from each of them based on expected yield differences and variability. Such an *ex-ante* analysis based on a spatial framework that accounts for key biophysical variables explaining yield variation and response to management practices, together with a power analysis to determine the minimum number of fields, can help prioritize resources by targeting TEDs with largest area and/or production (or other criteria) and ensuring that a sufficient (but not excessive number) of surveys are

collected for each TED.

We used 26 different statistical models to analyze the data and we identified management and soil variables that were consistently correlated with soybean yield. For a given TED, selected variables varied across methods, which we attributed to the specific properties of each technique. For example, regression procedures such as LASSO, LAR, and elastic net have properties that can mitigate data multi-collinearity issues (Zou and Hastie, 2005, Dormann et al., 2013). Hence, as also noted in Krupnik et al. (2015), we found conditional inference tree analysis to outperform other statistical approaches when the goal is to both analyze and interpret unstructured farmer survey data. In our analysis for soybean in the US North-Central region, across TEDs, conditional inference models explained 10–44% of field-to-field yield variation using only one to four explanatory variables. Allowing the development of larger tress using increasingly smaller groups of fields would have inflated R^2 values at expense of higher uncertainty and increasing difficult to interpret the results. Likewise, our analysis was not intended to capture all possible sources of variability (*e.g.,* climate variables). Instead, our objective was to identify key management, soil, and terrain factors influencing yield within each TED that could eventually be manipulated by farmers.

Sowing date and foliar fungicide and/or insecticide were the most persistent factors associated with yield variation. These results are consistent with findings from previous research based on farmer data collected from small geographic areas (*e.g.*, Grassini et al., 2015), multi-year, multi-location replicated field experiments (*e.g.*, Bastidas et al., 2008, Rowntree et al., 2013; Mourtzinis et al., 2016) and simulation modeling (*e.g.*, Specht et al., 2014). But, in contrast to these previous studies, our analysis also exposed interesting interactions between management practices, for example, MG x water regime and nematodes x tillage, which are also consistent with experimental data (Specht et al., 1986, 2001; Conley et al., 2011). Interestingly, we could not detect a positive influence of narrow or intermediate row spacing on soybean yield despite the yield benefits of narrow row spacing reported in previous studies (Anaele and Bishnoi, 1992; Oplinger and Philbrook, 1992; Hanna et al., 2008; Chauhan and Opena, 2013). These contrasting results derived from on-farm data *versus* controlled experiments deserve further investigation. Sowing date exhibited a consistent association with yields, with diminishing yield as sowing date was delayed. It was remarkable that the yield loss due to late sowing could not be fully compensated by any combination of other management practices, such as seeding rate or row spacing. In other words, sowing date appears to play a major role in setting the yield potential for a given field, as other factors cannot compensate for late sowing. Hence, timely sowing appears as a key factor to increase the current soybean yields in the US NC region.

Identification of the causes for yield variation is needed but not sufficient for increasing farmer yields. For example, we identified sowing date as a key management factor explaining yield variation within the same TED. Hence, one would tend to think that it is relatively easy for a large number of farmers in the US North Central region to increase current soybean yield by sowing earlier, especially considering that early sowing date *per se* does not involve higher costs and labor. However, there are many reasons why farmers may still be reluctant to sow soybean earlier. The first constraint is a combination of farm logistics and cultural preference as many farmers only have one planter and they prefer to use it for sowing maize first. The second limitation is associated with biophysical factors (*i.e.*, water excess, cold weather) that could delay sowing time in many years. Finally, farmers tend to overestimate the risk associated with seed chilling injury, early frost, and seed and/or plant stand loss associated with early sowing despite the well-documented benefits of early sowing and associated measures to reduce risk, for example, by using seed treatments or monitoring of soil temperature (*e.g.*, Bastidas et al., 2008; Rowntree et al., 2013; Tenorio et al., 2016). Additionally, the current crop insurance program sets a limit to very early sowing for a given area. We note, however, that our analysis showed that a large number of the farmers are sowing soybean much earlier than other farmers within the same TED suggesting that closing the portion of the yield gap due to sowing date is possible through fine tune adjustment of farm logistics and a correct assessment (and mitigation) of risk level. Indeed, over the past three decades, farmers have persistently shifted average soybean sowing times in the US North Central region to earlier calendar dates at a rate of *ca.* 0.5 d $year^{-1}$ (Specht et al., 2014). The present study indicates that there is still large room for improving soybean yields by increasing the rate at which farmers shift toward early sowing.

In a broader context, given the growing pressure for increasing food production on existing cropland area, the approach used here represents a tremendous opportunity to help accelerate rates of yield gain and better prioritize research and extension programs in major crop producing regions of the world. Another strength of the approach is that it screens for suites of 'best' management practices within the context of the current cropping system; hence, it is able to capture the continuous changes in management practices as a result of farmer innovation and adoption of new technologies and identify emerging problems (Loomis, 1984; Passioura, 2010; Grassini et al., 2014b) While replicated field trials will still be needed to establish cause-effect relationships, the information derived from analysis on farmer data as presented here can provide a focus to these trials in regard to which factors (and interactions) to investigate. In other words, our approach can be considered as a complement to research based on randomized replicated field experiments. The approach proposed here is cost-effective and generic enough to be applied in any cropping system in the world as long as underpinning soil and climate data needed to contextualize farmer fields are available.

Acknowledgements

The authors acknowledged the North-Central Soybean Research Program (NCSRP), Nebraska Soybean Board, and Wisconsin Soybean Marketing Board for their support to this project. We also thank NE and ND Extension Educators, Nebraska Natural Resources Districts, and Iowa Soybean Association for helping collect the farmer data. Finally, we thank Lim Davy, Agustina Diale, Laurie Gerber, Clare Gietzel, Mariano Hernandez, Ngu Kah Hui, Caleb Novak, Juliana de Oliveira Hello, Matt Richmond, and Paige Wacker for helping in data inputting.

References

Alexandratos, N., Bruinsma, J., 2012. World Agriculture Towards 2030/2050: the 2012 Revision. FAO, Rome.

Anaele, A.O., Bishnoi, U.R., 1992. Effects of tillage: weed control method and row spacing on soybean yield and certain soil properties. Soil Till. Res. 23, 333–340.

Bastidas, A.M., Setiyono, T.D., Dobermann, A., Cassman, K.G., Elmore, R.W., Graef, G.L., Specht, J.E., 2008. Soybean sowing date: the vegetative, reproductive, and agronomic impacts. Crop Sci. 48, 727–740.

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

Chauhan, B.S., Opena, L.J., 2013. Effect of plant spacing on growth and grain yield of soybean. Am. J. Plant Sci. 4, 2011–2014.

Conley, S.P., Pedersen, P., Esker, P., Gaska, J., 2011. Soybean yield and Heterodera Glycines response to rotation tillage, and source of genetic resistance. Agron. J. 103, 1604–1609.

De Bruin, J.L., Pedersen, P., 2008. Soybean seed yield response to planting date and seeding rate in the Upper Midwest. Agron. J. 100, 696–703.

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., García Marquéz, J.R., Gruber, B., Lafourcade, B., Leitão, P.J., Münkemüller, T., McClean, C., Osborne, P.E., Reineking, B., Schröder, B., Skidmore, A.K., Zurell, D., Lautenbach, S., 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36, 27–46.

FAOSTAT, 2016. Crops and Livestock Trade Database. www.faostat.fao.org.

Ferraro, O.D., Rivero, E.D., Claudio, M.G., 2009. An analysis of the factors that influence sugarcane yield in Northern Argentina using classification and regression trees. Field Crops Res. 112, 149–157.

Gaspar, A.P., Mueller, D.S., Wise, K.A., Chilvers, M.I., Tenuta, A.U., Conley, S.P., 2017. Response of broad spectrum and target specific seed treatments and seeding rate on soybean seed yield profitability, and economic risk across diverse environments. Crop Sci. 56, 2251–2262. http://dx.doi.org/10.2135/cropsci2016.11.0967.

Grassini, P., Thorburn, J., Burr, C., Cassman, K.G., 2011. High-yield irrigated maize in the Western U.S. Corn Belt: i on-farm yield, yield potential, and impact of agronomic practices. Field Crops Res. 120 (1), 142–150.
Grassini, P., Eskridge, K., Cassman, K.G., 2013. Distinguishing between yield advances and yield plateaus in historical crop production trends. Nat. Commun. 4, 2918.
Grassini, P., Torrion, J.A., Cassman, K.G., Yang, H.S., Specht, J.E., 2014a. Drivers of spatial and temporal variation in soybean yield and irrigation requirements in the western US Corn Belt. Field Crops Res. 163, 32–46.
Grassini, P., Specht, J., Tollenaar, T., Ciampitti, I., Cassman, K.G., 2014b. High-yield maize-soybean cropping systems in the U.S. Corn Belt. In: Sadras, V.O., Calderini, D.F. (Eds.), Crop Physiology- Applications for genetic improvement and agronomy, 2nd edition. Elsevier, The Netherlands.
Grassini, P., Torrion, J.A., Yang, H.S., Rees, J., Andersen, D., Cassman, K.G., Specht, J.E., 2015. Soybean yield gaps and water productivity in the western U.S. Corn Belt. Field Crops Res. 179, 150–163.
Hanna, O.S., Conley, S.P., Shaner, E.G., Santini, B.J., 2008. Fungicide application timing and row space effect on soybean canopy penetration and grain yield. Agron. J. 100, 1488–1492.
Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York.
Hothorn, T., Hornick, K., Zeileis, A., 2006. Unbiased recursive partitioning: a conditional inference framework. J. Comput. Graph. Stat. 15 (3), 651–674.
Krupnik, J.T., Ahmed, U.Z., Timsina, J., Yasmina, S., Hossain, F., Mamun, A.A., Mridha, I.M., McDonald, J.A., 2015. Untangling crop management and environmental influences on wheat yield variability in Bangladesh: an application of non-parametric approaches. Agric. Syst. 139, 166–179.
Lobell, B.D., Ortiz-Monasterio, J.I., Asner, G.P., Naylor, R.L., Falcon, W.P., 2005. Combining field surveys, remote sensing, and regression trees to understand yield variations in an irrigated wheat landscape. Agron. J. 97, 241–249.
Loomis, S.R., 1984. Traditional agriculture in America. Annu. Rev. Ecol. Evol. Syst. 15, 449–478.
Mercau, J.L., Sadras, V.O., Satorre, E.H., Messina, C., Balbi, C., Uribelarrea, M., Hall, A.J., 2001. On-farm assessment of regional and seasonal variation in sunflower yield in Argentina. Agric. Syst. 67, 83–103.
Mingers, J., 1987. Expert systems—rule induction with statistical data. J. Oper. Res. Soc. 38, 39–47.
Moore, D.I., Gessler, P.E., Nielsen, G.A., Peterson, G.A., 1993. Soil attribute prediction using terrain analysis. Soil Sci. Soc. Am. J. 57, 443–452.
Mourtzinis, S., Marburger, D.A., Gaska, J.M., Conley, S.P., 2016. Characterizing soybean yield and quality response to multiple prophylactic inputs and synergies. Agron. J. 108, 1–9.
Mourtzinis, S., Kaur, G., Orlowski, J.M., Shapiro, C.A., Lee, C.D., Wortmann, C., Holshouser, D., Nafziger, E.D., Kandel, H., Niekamp, J., Ross, J., Lofton, J., Vonk, J., Roozeboom, K.L., Thelen, K.D., Lindsey, L.E., Staton, M., Naeve, S.L., Casteel, S.N., Wiebold, W.J., Conley, S.P., 2018. Soybean response to nitrogen application across the United States: a synthesis-analysis. Field Crops Res. 215, 74–82.
Oplinger, E.S., Philbrook, B.D., 1992. Soybean planting date, row width, and seeding rate response in three tillage systems. J. Prod. Agric. 5, 94–99.
Passioura, J.B., 2010. Scaling up: the essence of effective agricultural research. Funct. Plant Biol. 37, 585–591.
Rattalino Edreira, R.J.I., Mourtzinis, S., Conley, S.P., Roth, A.C., Ciampitti, I.A., Licht, M.A., Kandel, H., Kyveryga, P.M., Lindsey, L.E., Mueller, D.S., Naeve, S.L., Nafziger, E., Stanley, J., Staton, M.J., Grassini, P., 2017. Assessing causes of yield gaps in agricultural areas with diversity in climate and soils. Agric. Forest Meteorol. 247, 170–180.
Rowntree, S., Suhre, J.J., Weidenbenner, N., Wilson, E., Davis, V., Naeve, S., Casteel, S., Diers, B., Esker, P., Specht, J., Conley, S.P., 2013. Genetic gain × management interactions in soybean: i. Planting date. Crop Sci. 53, 1128–1138.
Sadras, V., Roget, D., O'Leary, G., 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.
Silva, J.V., Reidsma, P., Laborte, A.G., van Ittersum, M.K., 2016. Explaining rice yields and yield gaps in Central Luzon, Philippines: an application of stochastic frontier analysis and crop modelling. Eur. J. Agron. 82, 223–241.
Specht, J.E., Williams, J.H., Weidenbenner, C.J., 1986. Differential responses of soybean genotypes subjected to a seasonal soil water gradient. Crop Sci. 26, 922–934.
Specht, J.E., Chase, K., Macrander, M., Graef, G.L., Chung, J., Markwell, J.P., Germann, M., Orf, J.H., Lark, K.G., 2001. Soybean response to water: a QTL analysis of drought tolerance. Crop Sci. 41, 493–509.
Specht, J.E., Diers, B.W., Nelson, R.L., Toledo, J.F., Torrion, J.A., Grassini, P., 2014. Soybean (Glycine max (L.) Merr.). In: Smith, J.S.C., Carver, B., Diers, B.W., Specht, J.E. (Eds.), Yield Gains in Major US Field Crops: Contributing Factors and Future Prospects. CSSA Special Publication #33, ASA-CSSA-SSSA, Madison, WI.
Tenorio, F.A., Grassini, P., Rees, J., Glewen, K., Mueller, N., Thompson, L., Specht, J., 2016. Early Bird Gets the Worm: Benefits of Early Soybean Planting. UNL CropWatch URL. http://cropwatch.unl.edu/2016/early-bird-gets-worm-benefits-early-soybean-planting.
Tilman, D., Balzer, C., Hill, J., Befort, B.L., 2011. Global food demand and the sustainable intensification of agriculture. Proc. Natl. Acad. Sci. U. S. A. 108, 20260–20264.
Tittonell, P., Shepherd, K.D., Vanlauwe, B., Giller, K.E., 2008. Unravelling the effects of soil and crop management on maize productivity in smallholder agricultural systems of western Kenya–an application of classification and regression tree analysis. Agric. Ecosyst. Environ. 123, 137–150.
USDA-NASS, 2016. USDA-National Agricultural Statistics Service (NASS). Crop U.S. State and County Databases.
Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Philos. Trans. R. Soc. Lond. Ser. B (Stat. Methodol.) 67, 301–320.