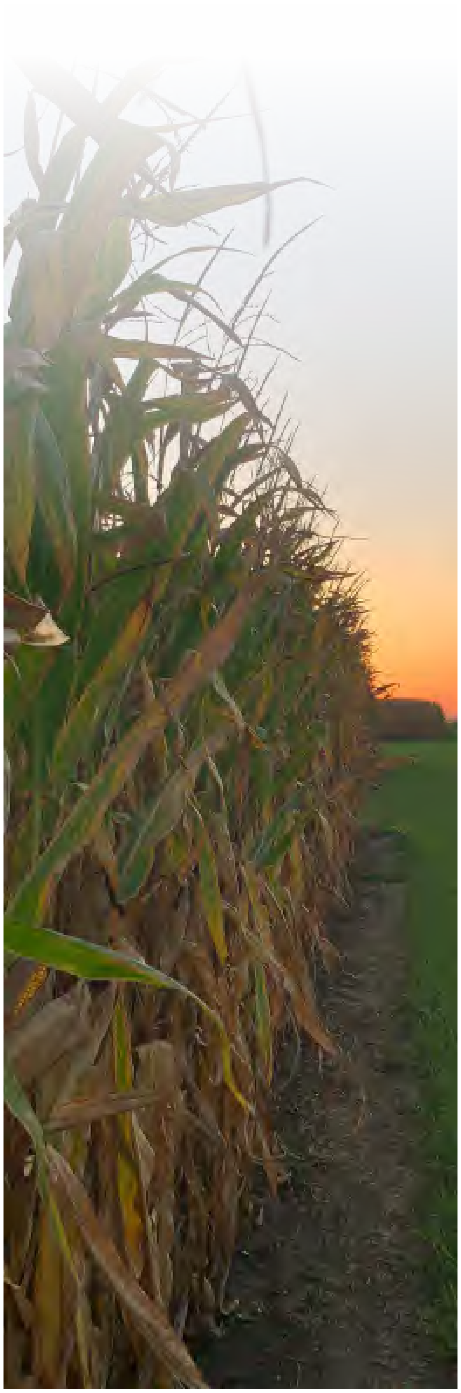
COOL BEAN
University of Wisconsin-Madison | UW Extension
WWW.COOLBEAN.INFO

Use of data science to optimize farm-specific cropping systems

Spyros Mourtzinis, Paul D. Esker, James E. Specht, and Shawn P. Conley



IN A BEAN POD:

- The effect of background management on crop yield is often ignored due to complexity
- Data science was used to develop and analyze diverse datasets
- Machine learning algorithms can identify farm-specific hidden yield potential
- The effect of multiple management interactions on crop yield can greatly influence corn grain and soybean seed yield and should not be ignored
- A single optimal solution does not necessarily exist and different combinations of management practices, when they interact with environment, can still result in similar high yields

INTRODUCTION

Increasing food demand will challenge the agricultural sector globally over the next decades (Godfray et al, 2010). A sustainable solution to this challenge is to increase crop yield without massive cropland area expansion. This can be achieved by identifying and adopting best management practices. To do so requires a more detailed understanding of how crop yield is impacted by climate change (Schlenker and Lobell, 2010; Mourtzinis et al., 2015) and growing-season weather variability (Hoffman et al., 2020). Even with that knowledge, prediction is challenging because various factors interact with each other. For example, variability in soil type can interact with weather conditions and mitigate or aggravate climate-related impacts on crop yield. Additionally, seed genetics (G) and crop management decisions (M), interact with the effect of environment (E: soil and in-season weather conditions), thereby resulting in a near infinite number of combinations of G × E × M that can impact crop yield.

Replicated field experiments have been used to identify best management practices for several decades. Most commonly, the effectiveness of up to three management factors and their interactions are evaluated in a single location due to practical constraints (e.g., cost, logistics). By holding the background management constant, causal relationships are identified, and the effectiveness of the examined management practice/s is assessed. It is assumed that background management practices are optimal or at least relevant to what most farmers use in the region, which in fact may not be realistic for many farmers.

Multi-year-site performance trials that account for large environmental and background management variability is another common practice in agricultural research. Such trials usually estimate an average effect across environments and background cropping systems. Inevitably, the measured yield response magnitude and sign may not apply to all farms in the examined region. Other research approaches involve analysis of producer self-reported data (Rattalino Edreira et al., 2017; Mourtzinis et al., 2018), which can capture yield trends attributable to producer management choice across large regions, but such studies lack sufficient power relative to establishing causality and evaluating complex high-order G × E × M interactions.

Process-based models have been extensively used to evaluate the effect of weather (Frieler et al, 2017) and management (Rong et al., 2019) on crop yield. However, to obtain accurate estimates, the models require extensive calibration, which is not a trivial task due to the large number of parameters. Specifically, it has been shown that management is an important source of uncertainty in process-based models, which can lead to substantial and varying degree of bias in yield estimates across the US, even when using harmonized parameters (Leng and Hall, 2020).

Given all the well-known deficiencies of current agricultural research methods, we argue that a method that allows environment-specific identification of unique cropping systems with the greatest yield potential is essential to meet future food demand. Here, by utilizing corn and soybean yield and management data from publicly available performance tests, plus associated weather data, and by leveraging the power of machine learning (ML) algorithms, we developed a method that can evaluate myriads of potential crop management systems and thereby identify those with the greatest yield potential in specific environments across the US.

METHODS

Two databases including yield, management, and weather data for corn (n=17,013) and soybean (n=24,848) involving US crop performance trials conducted in 28 states between 2016 to 2018 for corn and between 2014 to 2018 for soybean, were developed (Fig. 1). For each crop, an ML algorithm to estimate yield based on soil type and weather conditions (E), seed traits (G) and management practices (M) was developed.

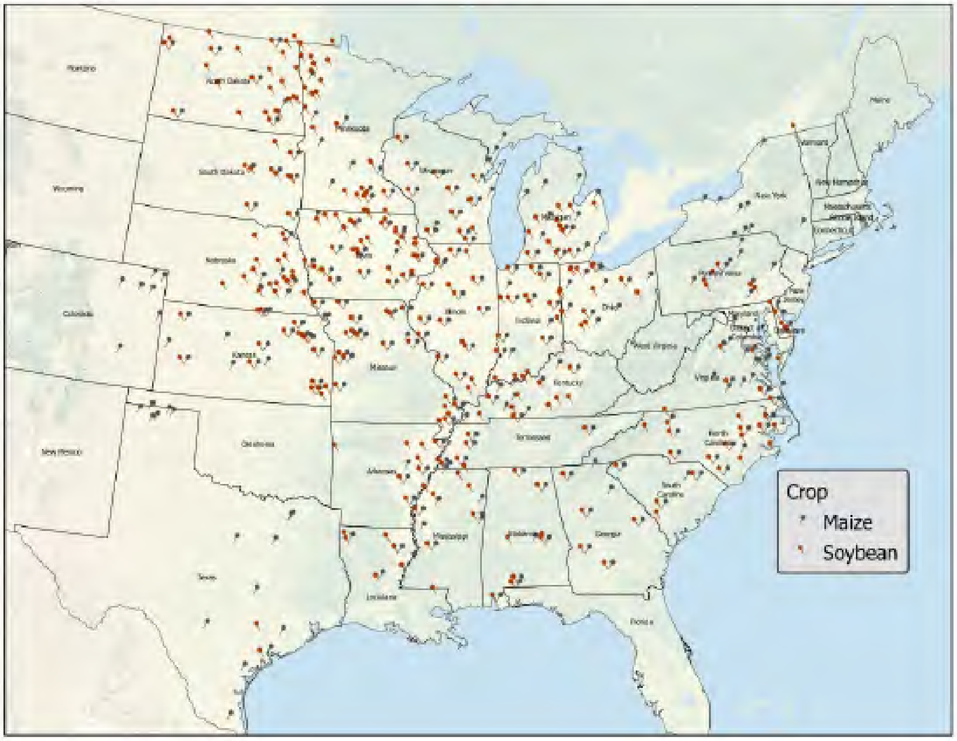
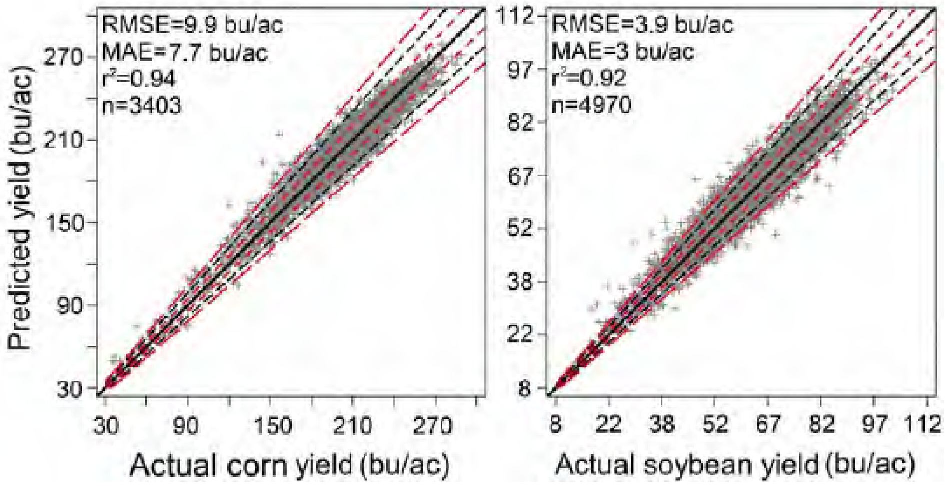


Figure 1. Locations where corn and soybean trials were performed during the examined period. The map was developed in ArcGIS Pro 2.8.0 (https://www.esri.com).

RESULTS AND DISCUSSION

The developed algorithms exhibited a high degree of accuracy when estimating yield in independent datasets (test dataset not used for model calibration) (Fig. 2). For corn, the root mean square error (RMSE) and mean absolute error (MAE) was a respective 4.7 and 3.6% of the dataset average yield (213 bu/ac). For soybean, the respective RMSE and MAE was 6.3 and 4.8% of the dataset average yield (62 bu/ac). As is evident in the graphs (Fig. 2), estimated yields exhibited a high degree of correlation with actual yields for both algorithms in the independent datasets. For corn and soybean, 72.3 and 60% of cases in the test dataset deviated less than 5% from actual yields, respectively. Data points with deviations greater than 15% from actual

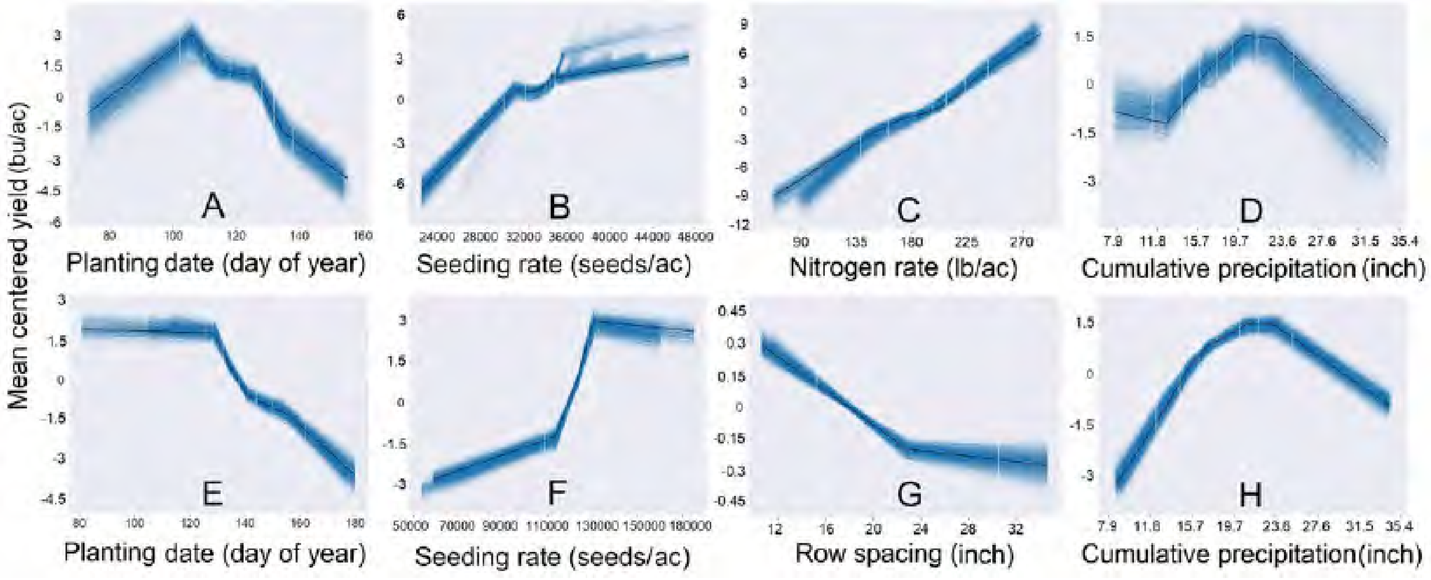
Figure 2. Actual *vs* algorithm-derived corn (left) and soybean (right) yield in test datasets. Black solid line indicates y=x, red short-dashed lines, black dashed lines, and red long-dashed lines indicate ± 5, 10, and 15% deviation from the y=x line. RMSE, root mean square error; MAE, mean absolute error; r^2, coefficient of determination; n=number of observations. Each observation corresponds to a yield of an individual cropping system in a specific environment (location-year).



yield were 1.5% in corn and 3.6% in soybean databases. These results suggest that the developed algorithms can accurately estimate corn and soybean yields utilizing database-generated information involving reported environmental, seed genetic, and crop management variables.

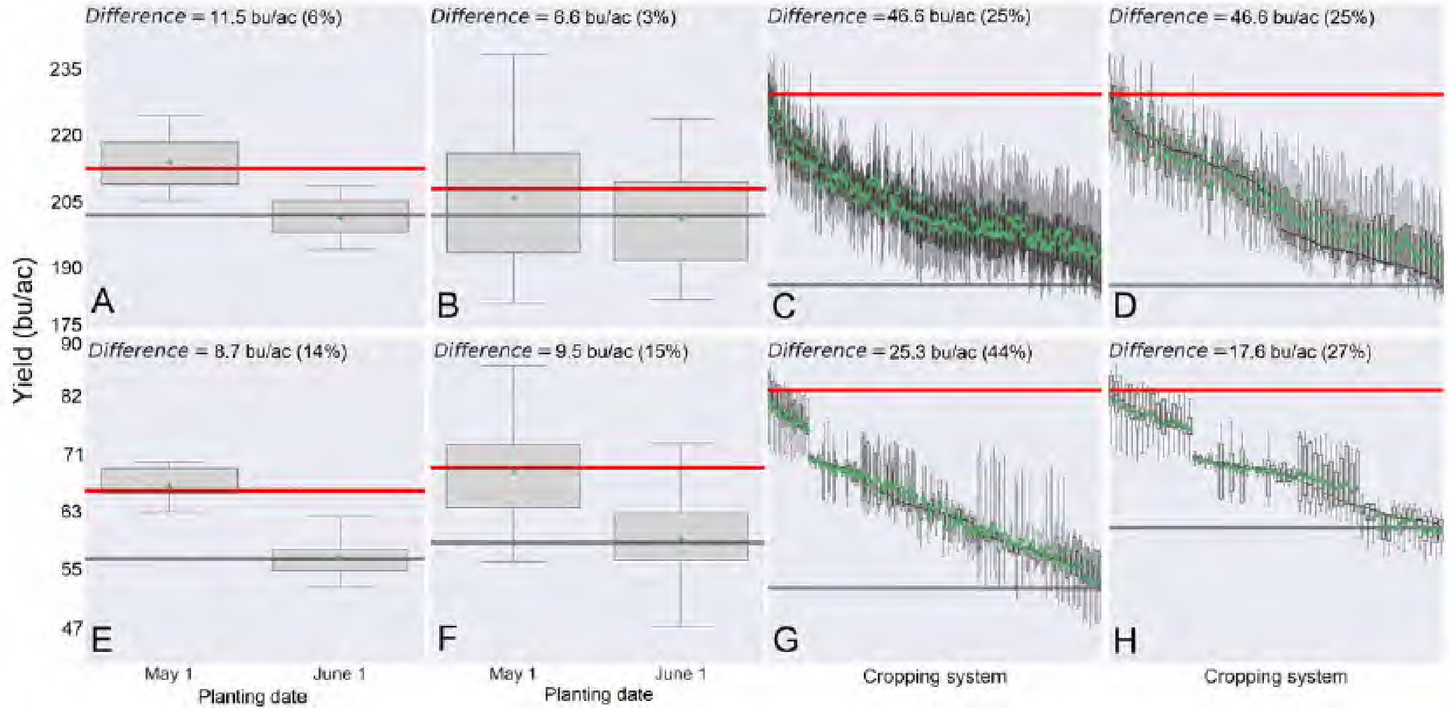
In contrast to statistical models, ML algorithms can be complex, and the effect of single independent variables may not obvious. However, accumulated local effects (ALE) plots (Apley and Zhu, 2016) can aid the understanding and visualization of important and possibly correlated features in ML algorithms. For both crops, indicatively important variables included planting date, seeding rate, nitrogen fertilizer (for corn), row spacing (for soybean) and June to September cumulative precipitation (Fig. 3). Across the entire region and for both crops, the algorithm-derived trends suggest that above average yields occur in late April to early May planting dates, but sharply decrease thereafter. Similar responses have been observed in many regional studies across the US for both, corn (Long et al, 2017) and soybean (Mourtzinis et al, 2019). Similarly, simulated yield curves due to increasing seeding rate are in close agreement with previous corn (Light et al., 2016) and soybean (Gaspar et al., 2020) studies. The corn algorithm has captured the increasing yield due to increasing N fertilizer rate. The soybean algorithm suggests that narrower row spacing resulted in above average yield compared to wider spacing. Such response has been observed in many regions across the US (Andrade et al., 2019). Season cumulative precipitation between 16 and 28 inches resulted in above average yields for both crops.

Figure 3. Accumulated local effect plots for corn planting date (A), seeding rate (B), Nitrogen fertilizer rate (C), and cumulative precipitation between June and September (D), and soybean planting date (E), seeding rate (F), row spacing (G), and cumulative precipitation between June and September (H).



The responses in the ALE plots (Fig. 3) suggest that these algorithms have captured the general expected average responses for important single features. Nevertheless, our databases include hundreds of locations with diverse environments across the US and site-specific crop responses which may vary due to components of the G × E × M interaction. We argue that, instead of examining a single or low-order management interactions, site-specific evaluation of complex high order interactions (a.k.a. cropping systems) can reveal yield differences that current research approaches cannot fully explore and quantify. For example, planting date exerts a well-known impact on corn and soybean yield. For each crop separately, by creating a hypothetical cropping system (a single combination of all variables) in a randomly chosen field in south central Wisconsin (latitude=43.34, longitude=-89.38), and by applying the developed algorithms, we can generate estimates of corn and soybean yield. For that specific field and cropping system (out of the vast number of management combinations a farmer can choose from), corn yield with May 1st planting was 11.5 bu/ac greater (6% increase) than June planting (Fig. 4 A). By creating scenarios with 256 background cropping system choices (Table 1), the resultant algorithm-derived yield estimate difference for the same planting date contrast (averaged across varying cropping systems) was smaller but still positive (3% increase), although the range of possible yield differences was wider (Fig. 4 B). However, when comparing, instead of averaging, the estimated yield potential among the simulated cropping systems, a 46.6 bu/ac yield difference (25% difference) was observed (Fig. 4 C). Interestingly, when focusing on the early sown fields that were expected to exhibit the greatest yield, the same yield difference was observed (Fig. 4 D). This result shows that sub-optimal background management can mitigate the beneficial effect of early planting (Table 2).

Figure 4. Corn yield difference (in bu/ac and percentage) due to planting date (May 1st vs June 1st) for a single identical background cropping system (A), corn yield difference due to planting date when averaged across 256 (3 years × 256 cropping systems=768 year-specific yields) (B), corn yield variability in each of the 256 cropping systems (C), and corn yield variability in each of the 128 cropping systems with early planting (D). Soybean yield difference due to planting date (May 1st vs June 1st) for a single identical background cropping system (E), soybean yield difference due to planting date when averaged across 128 (5 years × 128 cropping systems=640 year-specific yields) (F), soybean yield in each of the 128 cropping systems (G) and soybean yield variability in each of the 64 cropping systems with early planting (H). Within each panel, the horizontal red and grey lines indicate the boxplot with maximum and minimum yield, respectively. In the left four panels, boxes delimit first and third quartiles; solid lines inside boxes indicate median and green triangles indicate means. Upper and lower whiskers extend to maximum and minimum yields. Each corn and soybean cropping system is a respective interaction of management practices in a randomly chosen field in Wisconsin, USA.



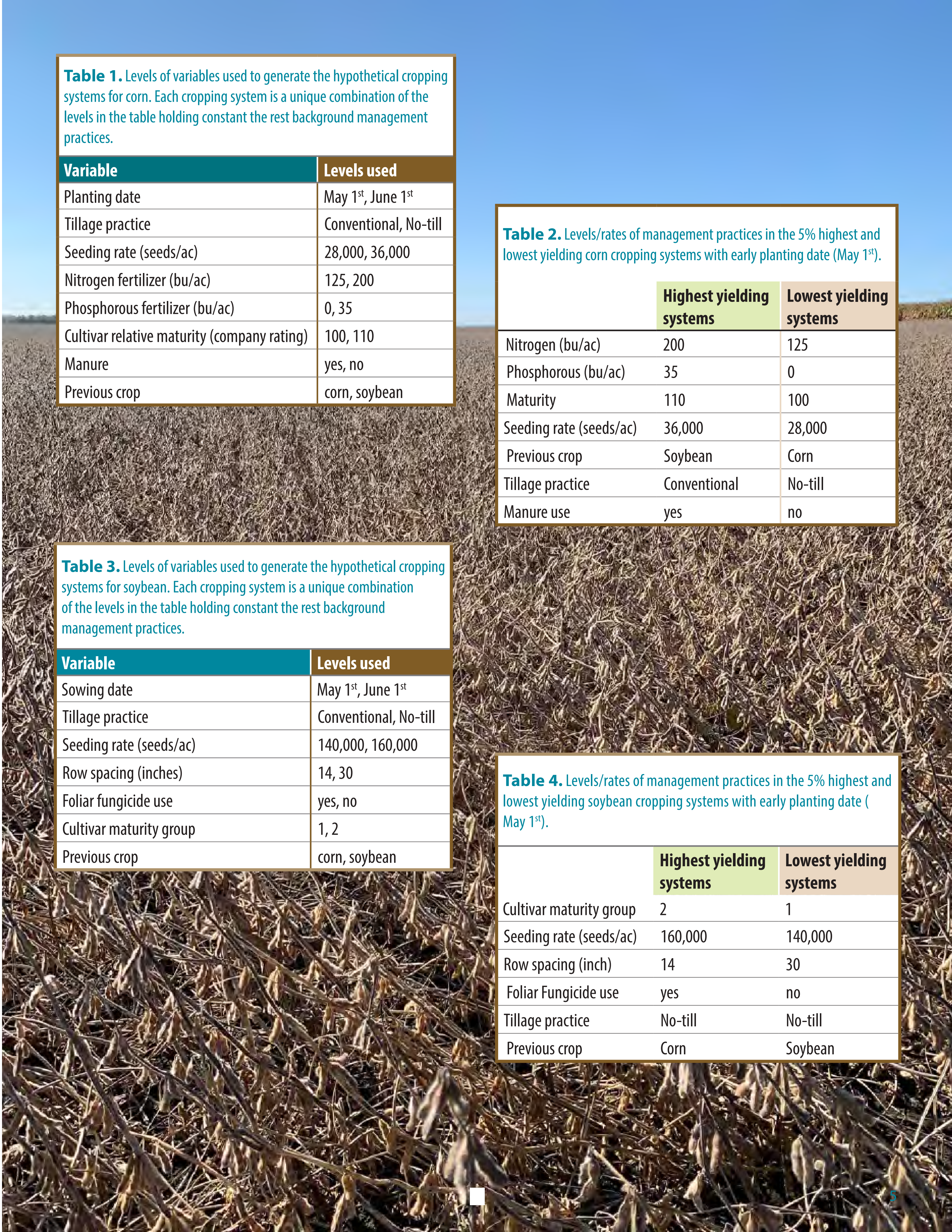


Table 1. Levels of variables used to generate the hypothetical cropping systems for corn. Each cropping system is a unique combination of the levels in the table holding constant the rest background management practices.

Variable	Levels used
Planting date	May 1st, June 1st
Tillage practice	Conventional, No-till
Seeding rate (seeds/ac)	28,000, 36,000
Nitrogen fertilizer (bu/ac)	125, 200
Phosphorous fertilizer (bu/ac)	0, 35
Cultivar relative maturity (company rating)	100, 110
Manure	yes, no
Previous crop	corn, soybean

Table 2. Levels/rates of management practices in the 5% highest and lowest yielding corn cropping systems with early planting date (May 1st).

	Highest yielding systems	Lowest yielding systems
Nitrogen (bu/ac)	200	125
Phosphorous (bu/ac)	35	0
Maturity	110	100
Seeding rate (seeds/ac)	36,000	28,000
Previous crop	Soybean	Corn
Tillage practice	Conventional	No-till
Manure use	yes	no

Table 3. Levels of variables used to generate the hypothetical cropping systems for soybean. Each cropping system is a unique combination of the levels in the table holding constant the rest background management practices.

Variable	Levels used
Sowing date	May 1st, June 1st
Tillage practice	Conventional, No-till
Seeding rate (seeds/ac)	140,000, 160,000
Row spacing (inches)	14, 30
Foliar fungicide use	yes, no
Cultivar maturity group	1, 2
Previous crop	corn, soybean

Table 4. Levels/rates of management practices in the 5% highest and lowest yielding soybean cropping systems with early planting date (May 1st).

	Highest yielding systems	Lowest yielding systems
Cultivar maturity group	2	1
Seeding rate (seeds/ac)	160,000	140,000
Row spacing (inch)	14	30
Foliar Fungicide use	yes	no
Tillage practice	No-till	No-till
Previous crop	Corn	Soybean

In the case of soybean, a May 1st planting resulted in greater yield (8.7 bu/ac; a 14% increase) than a June 1st in the single background cropping system (Fig. 4 E). The result was consistent when yield differences due to planting date were averaged across 128 background cropping system choices (Table 3) (Fig. 4 F). Similar to what was observed in corn, among all cropping systems, yield varied by 25.3 bu/ac (44% difference) (Fig. 4 G). When focusing only on the early sown fields, a 17.6 bu/ac yield difference (27% yield increase) was observed (Fig. 4 H). In agreement with corn, this result highlights the importance of accounting for sub-optimal background management which can mitigate the beneficial effect of early planting (Table 4).

We note here the ability of farmers to change management practices can be limited due to an equipment constraint (e.g., change planter unit row width) or simply impossible (e.g., change the previous year's crop). Thus, recommended management practices that were evaluated in studies that used specific background management may not be applicable in some instances. The benefits of the foregoing approach, which involves extensive up-to-date agronomic datasets and high-level computational programing, can have important and immediate implications in future agricultural trials. Our approach allows for more precise examination of complex management interactions in specific environments (soil type and growing season weather) across the US (region covered in Fig. 1).

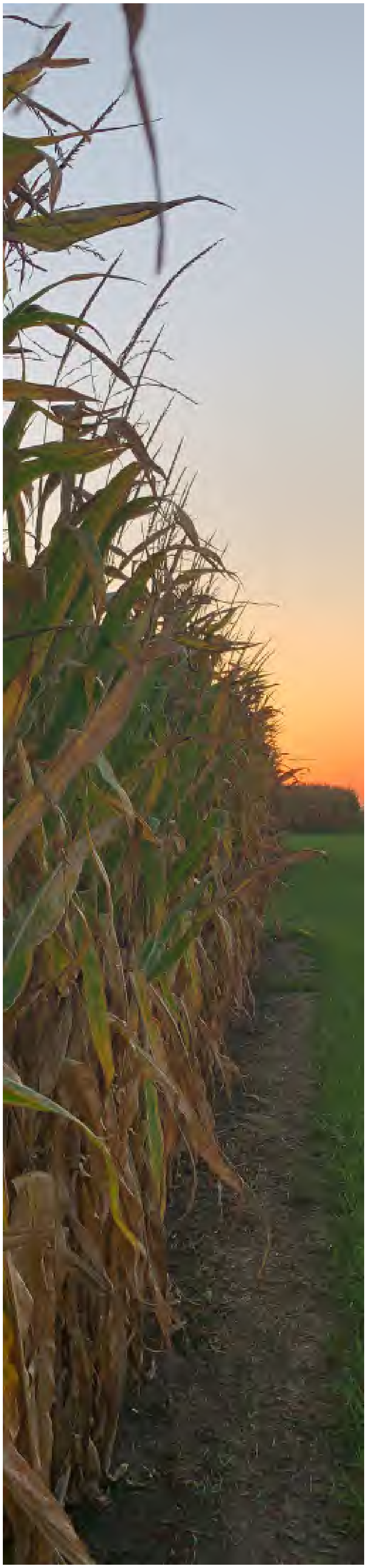
It appears that several different cropping systems can result in similar high yield for both crops (Fig. 4 C, D and G, H). Moreover, it is common for neighboring farms to attain similar crop yield despite the use of a different cropping system, suggesting that a single optimal solution does not necessarily exist and that different combinations of management practices, when they interact with environment, can still result in similar high yields. Since the effect of environment is ever-changing, the high level of complexity of synergies between G × E × M suggests that long-term optimization of single management factor may not be possible, which further highlights the importance of accounting for the effect of the entire cropping system at the field level.

The algorithms we present here can generate hypothetical experimental data that can be used to rapidly examine G × E × M interaction for both corn and soybean across the US. Of the millions of possible G × E × M combinations, our ML algorithms can identify hidden complex patterns between G × E × M combinations for yield optimization that may be non-obvious, but once identified, worthy of field test confirmation. Farmers can use the algorithms to gain insights about optimum management interactions in their location-specific environment (known soil type × expected weather conditions), and to identify farm factors that may be too costly to alter without *a priori* reason (generated by the model) for doing so. Researchers can compare expected yield across thousands of hypothetical cropping systems and use the results as a guide to design more efficient future field-based crop management practice evaluation experiments.

We note that this approach should not be considered as a substitute of replicated trials. To the contrary, replicated field trials performed by Universities are continually needed to serve as an excellent source of high-quality unbiased data which can be used to train even more comprehensive algorithms. The major issue with current performance trial data is that a great amount of management information is not reported. Usually, only information relevant to the examined management factors in each trial are reported, which inevitably results in missing values, or even in absence of important variables (e.g., number and dates of split fertilizer application). As we have highlighted here, the high order and complex background management interactions should not be considered as irrelevant.

CONCLUSIONS

Agricultural experiments repeated every year in hundreds of locations across the US generate a vast amount of crop yield and management datasets which are useful to identify average effects of management practices across a range of environments. Such datasets have, to date, remained disconnected from each other, and are difficult



to combine, standardize, and properly analyze. In the presented work, we overcame these issues by developing large databases and by leveraging the power of ML algorithms. We argue that our algorithms can advance agricultural research and aid in revealing a currently hidden yield potential in each individual farm across the US.

ACKNOWLEDGMENTS

This research was funded in part by the Wisconsin Soybean Marketing Board, The North Central Soybean Research Program (S.P. Conley), and the USDA National Institute of Food and Federal Appropriations under Project PEN04660 and Accession number 1016474 (P.D. Esker).

Adapted from: Mourtzinis, S., Esker, P., Specht, J., & Conley, S. 2021. Advancing agricultural research using machine learning algorithms. DOI:10.1038/s41598-021-97380-7

REFERENCES

H. C. J. Godfray, et al. Food security: The challenge of feeding 9 billion people. Science 812-818 (2010).

W. Schlenker, D. B. Lobell, Robust negative impacts of climate change on African agriculture Environ. Res. Lett. 5 014010 (2010).

S. Mourtzinis, et al. Climate-induced reduction in US-wide soybean yields underpinned by region- and in-season specific responses. Nat. Plants 1, 14026 (2015).

L. A. Hoffman, A.R. Kemanian, C.E. Forest, The response of maize, sorghum, and soybean yield to growing-phase climate revealed with machine learning. Environ. Res. Lett. 15, 094013 (2020).

J. I. Rattalino Edreira, et al. Assessing causes of yield gaps in agricultural areas with diversity in climate and soils. Agric. For. Meteorol. 247, 170-180 (2017).

S. Mourtzinis, et al. Sifting and winnowing: analysis of farmer field data for soybean in the US North-Central region. Field Crops Res. 221, 130-141 (2018).

K. Frieler, et al. Understanding the weather signal in national crop-yield variability. Earths Future. 5, 605–616 (2017).

J. Rong, et al. Exploring management strategies to improve maize yield and nitrogen use efficiency in northeast China using the DNDC and DSSAT models. Comput. Electron. Agric. 104988 (2019).

G. Leng, J. W. Hall, Predicting spatial and temporal variability in crop yields: an inter-comparison of machine learning, regression and process-based models. Environ. Res. Lett. 15 044027 (2020).

D. W. Apley, J. Zhu, Visualizing the effects of predictor variables in black box supervised learning models. arXiv:1612.08468v2 (2016).

N. V. Long, Y. Assefa, R. Schwalbert, I. A. Ciampitti, Maize Yield and Planting Date Relationship: A Synthesis-Analysis for US High-Yielding Contest-Winner and Field Research Data. Front. In Plant Sci. 8, 2106 (2017).

S. Mourtzinis, J. E. Specht, S. P. Conley, Defining optimal soybean sowing dates across the US. Sci. Rep. 9:2800 (2019).

Y. Assefa, et al. Yield responses to planting density for US modern corn hybrids: A synthesis-analysis. Crop Sci. 56, 2802-2817 (2016).

M. A. Light, A. W. Lenssen, R. W. Elmore, Corn (Zea mays L.) seeding rate optimization in Iowa, USA. Precision Agric. 18, 452-469 (2016).

A. Gaspar, et al. Defining optimal soybean seeding rates and associated risk across North America. Agron. J. 1–12 (2020).

J. Andrade, et al. Assessing the influence of row spacing on soybean yield using experimental and producer survey data. Field Crops Res. 230, 98-106 (2019).