Random forest regression for optimizing variable planting rates for corn and soybean using topographical and soil data

Margaret R. Krause1,2 | Savanna Crossman3 | Todd DuMond3 | Rodman Lott3 | Jason Swede3 | Scott Arliss3 | Ron Robbins3 | Daniel Ochs3 | Michael A. Gore1

1 Plant Breeding and Genetics Section, School of Integrative Plant Science, Cornell University, Ithaca, NY, 14853, USA
2 Global Wheat Program, International Maize and Wheat Improvement Center (CIMMYT), Texcoco, Edo. de México 56237, México
3 New York Corn and Soybean Growers Association, PO Box 133, Silver Springs, NY, 14550, USA

Abstract

In recent years, planting machinery that enables precise control of the planting rates has become available for corn (Zea mays L.) and soybean (Glycine max L.). With increasingly available topographical and soil information, there is a growing interest in developing variable rate planting strategies to exploit variation in the agri-landscape in order to maximize production. A random forest regression-based approach was developed to model the interactions between planting rate, hybrid/variety, topography, soil characteristics, weather variables, and their effects on yield by leveraging on-farm variable rate planting trials for corn and soybean conducted at 27 sites in New York between 2014 and 2018 (57 site-years) in collaboration with the New York Corn and Soybean Growers Association. Planting rate ranked highly in terms of random forest regression variable importance while explaining relatively minimal yield variation in the linear context, indicating that yield response to planting rate likely depends on complex interactions with agri-landscape features. Random forest models explained moderate levels of yield within site-years, while the ability to predict yield in untested site-years was low. Relatedly, variable importance measures for the predictors varied considerably across sites. Together, these results suggest that local testing may provide the most accurate optimized planting rate designs due to the unique set of conditions at each site. The proposed method was extended to identify the optimal variable rate planting design for maximizing yield at each site given the underlying conditions, and empirical validation of the resulting designs is currently underway.

1 | INTRODUCTION

In agricultural production systems, site-specific management involves the development of crop management strategies at a finer spatial scale than that of the whole field area (Plant, 2001). Compared to early farming, which...
was performed by hand and thereby facilitated site-specific management, the advent of mechanization in farming ushered in an era of uniform application of inputs on large areas of cropland. Studies have shown that uniform applications can result in suboptimal input use efficiency in parts of the field that may require greater or fewer inputs than the applied fixed rates (Moore & Tyndale-Biscoe, 1999; Mulla & Schepers, 1997), though increases in farm productivity generally outweighed the associated economic losses. However, the growing costs of agricultural inputs (USDA NASS, 2019) and negative environmental impacts of intensive production practices (Tilman, Cassman, Matson, Naylor, & Polasky, 2002) threaten the long-term economic and environmental sustainability of uniform crop management.

Variable rate application systems enable growers to apply inputs at a range of user-defined levels within a field area. Variable rate as it applies to planting is not a new concept. The first variable rate planting systems emerged in the United States during the 1970s but did not become popular due to their initial lack of automation, which necessitated that growers trigger planting rate transitions manually (Lowenberg-DeBoer, 1999). In the late 1990s, the integration of GPS and automated variable rate systems for commercial planting equipment renewed interest in variable rate planting technologies.

However, these improvements did not immediately drive a widespread adoption of variable rate planting, as users remained uncertain of how to develop variable rate planting designs to maximize returns given the underlying field conditions. Some early as well as more recent studies have based the development of variable rate planting designs on yield potential, delineated by previous years’ yield performance and/or the grower’s knowledge of the field (Barnhisel, Bitzer, Grover, & Shearer, 1996; Corassa et al., 2018; Hörbe, Amado, Ferreira, & Alba, 2013). However, Bullock et al. (1998) postulated that knowledge of yield potential would be insufficient to make variable rate technologies economically beneficial to farmers. They identified as a major barrier the community’s limited understanding of how planting rate and various field characteristics such as topography and soil, which can be difficult to measure, may interact to influence yield. They also identified year-to-year variation in weather as a potential challenge, as weather conditions can interact with topography and soil characteristics in a complex manner. For example, Timlin, Pachepsky, Snyder, and Bryant (1998) showed that grain yields for corn were higher in concave areas of fields during dry years, while wet years showed no correlation between yield and curvature. Finally, it is widely recognized that hybrids/varieties may differ from one another in their response to planting rate (McVickar & Shear, 1946; Rutger & Crowder, 1967).

More than 20 years later, use of variable rate planting technology remains limited, despite the increasing availability of high-resolution topographical and soil data and the abundance of commercial variable rate products on the market. In a 2016 survey of corn and soybean growers in New York State, only 10.5% of respondents had adopted variable rate planting technology and 20.0% cited skepticism towards the technology’s potential economic benefit as a key driver of its slow adoption (van Es et al., 2016).

Currently, there is little consensus among the community regarding how to create variable rate planting designs with respect to yield potential, field characteristics, and/or other observations. While many research efforts have attempted to determine the relationships between topographic features, soil characteristics, and yield potential for crops sown under fixed planting rates (Changere & Lal, 1997; Cox, Gerard, Wardlaw, & Abshire, 2003; Frogbrook, Oliver, Salahi, & Ellis, 2002; Katerji, vann Hoorn, Hamdy, Karam, & Mastrorilli, 1995; Kravchenko & Bullock, 2000; Miller, Singer, & Nielsen, 1988), few studies have assessed how planting rate might influence and interact with those relationships. Among those that have, the results have been mixed: some observed potential for the profitability for variable rate planting technology (Shanahan, Doerge, Johnson, & Vigil, 2004), while others found inconsistencies among planting rate optimizations and the interactions between planting rate and topographical/soil features across sites and years (Licht, Lenssen, & Elmore, 2017; Smidt, Conley, Zhu, & Arriaga, 2016).

In a literature review of variable rate technologies, Bullock and Lowenberg-DeBoer (2007) made the observation that the majority of studies pertaining to variable rate planting and chemical applications have reported results for a small number of specific site-years. The authors, among others (Lambert, Lowenberg-DeBoer, &
Malzer, 2006; Liu, Swinton, & Miller, 2006; Ruffo, Pollero, Bullock, & Bullock, 2006), emphasized the need for longer-term, multi-location experimental data in order to expand the inference space and sample the population of possible environmental conditions at a high level in order to evaluate the value and feasibility of variable rate planting.

To conduct experiments on such a scale, coordination between growers and researchers is needed in addition to a long-term commitment to on-farm testing. The New York Corn and Soybean Growers Association (NYCSGA) is a statewide non-profit organization that represents the interests of New York corn and soybean producers and sponsors relevant research on production, utilization, and marketing. In 2013, the NYCSGA developed a research initiative to optimize variable rate planting technology for corn and soybean growers in New York State. NYCSGA members throughout central, western, and northern New York conducted on-farm variable rate field trials between 2014 and 2018 with the objective of developing a strategy for variable rate planting to exploit the extensive native variation of New York’s agricultural landscape. To sample this variation, spatial data related to topographical features, soil type, and soil nutrients were collected at a high resolution.

This collaborative project driven by growers has created a wealth of information relating yield to planting rate and its interactions with a wide range of environmental factors. However, the magnitude and complexity of these data present a number of statistical challenges when conducting analysis. Many oft-recorded variables in environmental or agricultural studies exhibit significant correlations among one another. For example, topographical attributes such as elevation and slope can influence the movement of water through and over the landscape, thereby affecting soil development (Moore, Gessler, Nielsen, & Peterson, 1993). Relationships between topographical features and soil characteristics such as sand, silt, organic matter, pH, cation exchange capacity, and extractable phosphorus have been reported (Brubaker, Jones, Lewis, & Frank, 1993; Moore et al., 1993; Tan, Lal, Smeck, & Calhoun, 2004). Likewise, strong relationships between soil variables themselves are not uncommon. In a linear-regression analysis of the effects of soil characteristics on corn yield response to variable planting rates, Licht et al. (2017) removed variables describing silt and available water-holding capacity due to their collinearity with sand, clay, and soil organic matter. In addition to correlations between variables, non-linear relationships can often occur in agricultural systems (reviewed in Archontoulis & Miguez, 2014). For example, increasing temperatures are known to benefit corn yields but become harmful once they reach 30°C and above (Schlenker & Roberts, 2006). Despite the limited ability of linear regression to accurately model correlated characters and nonlinear relationships, many published studies evaluating the potential of variable rate planting technology have relied on linear approaches to assess the relationships between planting rate, yield, and environmental features (Hörbe et al., 2013; Licht et al., 2017; Shanahan et al., 2004).

Novel statistical approaches to analyze increasingly large and complex agronomic datasets are needed to fully exploit the information they contain. Random forest regression is an ensemble learning, nonparametric method that constructs multiple decision trees using random subsets of observations and predictor variables (Breiman, 2001). Random forest regression models have become increasingly popular in a range of scientific fields due to their ability to accommodate highly correlated variables, complex nonlinear interactions, and predictors of varying types and ranges. Random forest regression methods have been shown to reduce mean squared error and improve prediction accuracy compared to linear modeling approaches across a range of diverse applications involving highly dimensional data (Jeong et al., 2016; Mutanga, Adam, & Cho, 2012; Silva et al., 2017; Zhang et al., 2017).

Random forest regression has been used effectively for various environmental and agricultural applications such as predicting soil texture from terrain variables (Ließ, Glaser, & Huwe, 2012) and estimating vegetative biomass from remotely sensed data (Mutanga et al., 2012). In the context of modeling crop yields, random forest approaches have been applied to predict yields in corn, wheat, potato, sugarcane, sorghum, and groundnut based on climate variables (Everingham, Sexton, Skocaj, & Inman-Bamber, 2016; Hoffman, Kemanian, & Forest, 2017; Jeong et al., 2016). Tulbure, Wimberly, Bie, and Owens (2012) leveraged random forest regression to identify variables related to weather, genetics, and management practices that were underlying spatiotemporal variability in switchgrass yield. Smidt et al. (2016) used a similar approach to identify the key predictors of soybean yield among topographical and soil field attributes under a variable planting rate experimental design. This type of approach may also be useful for the subsequent development of optimized variable rate planting schemes.

The work presented here aims to 1) investigate the effects of planting rate, hybrid/variety, topographical features, soil characteristics, weather conditions, and their interactions on corn and soybean yield and 2) propose a random forest regression-based method for leveraging this information to build variable rate planting designs.
2 | MATERIALS AND METHODS

2.1 | Site-year summary

On-farm corn and soybean variable rate field trials were conducted in 27 unique sites throughout central, western, and northern New York (Figure 1) between 2014 and 2018 for a total of 57 site-years (32 corn, 25 soybean). The field area, row spacing, experimental design, and hybrids/varieties sown varied among site-years (Table 1). Field areas ranged from 10.8 to 59.9 ha with an average of 28.0 ha. For corn, 16 hybrids were used throughout the course of the experiment, though hybrids P0533AM1 and P0216/P0216AM were used most predominantly, appearing in 15 and 8 of the 32 site-years, respectively. For soybean, 21 varieties were used. Statewide weather data were obtained by the National Climatic Data Center (2020) (Figure 2). Monthly average temperature and precipitation from the nearest National Climatic Data Center weather observation station were likewise extracted for each site-year (National Climatic Data Center, 2020).

2.2 | On-farm experimental designs

All on-farm variable rate trials for corn were sown with a randomized design consisting of 0.81-ha (2 ac) blocks sown to one of four target planting rates (Figure 3). For corn, the following four target planting rates were tested: 66,718 (27,000), 79,074 (32,000), 91,429 (37,000), and 103,784 (42,000) seeds ha$^{-1}$ (ac$^{-1}$). Out of the 32 site-years for corn, 17 were sown with two hybrids and one was sown with three hybrids. All on-farm trials containing more than one hybrid were sown with a split planter, creating alternating strips of each hybrid within the field (Figure 3).
| Field | Year | Crop | Area | Row spacing | Hybrid/Variety | Planting rate design | Hybrid/Variety design | Experimental unit grid length/width | Average yield |
|-------|------|------|------|-------------|----------------|----------------------|----------------------|-----------------------------------|---------------|
| An6   | 2014 | Corn | 22.81 | 50.8 | P9690AM; P9675AMXT | Randomized Blocks | Alternating Hybrids/Varieties | 6.09 | 12,004.9 |
| An8   | 2014 | Corn | 29.99 | 50.8 | P0216; P0533AM1 | Randomized Blocks | Alternating Hybrids/Varieties | 6.09 | 12,151.4 |
| An8   | 2015 | Corn | 29.99 | 50.8 | P0216; P0533AM1 | Randomized Blocks | Alternating Hybrids/Varieties | 6.09 | 10,100.1 |
| Ba1   | 2015 | Corn | 21.23 | 76.2 | P0216; P0533AM1 | Randomized Blocks | Alternating Hybrids/Varieties | 9.14 | 9,805.5 |
| Be2   | 2014 | Corn | 51.15 | 76.2 | P0216; P0533AM1 | Randomized Blocks | Alternating Hybrids/Varieties | 9.14 | 15,126.9 |
| Bi    | 2015 | Corn | 14.92 | 101.6 | P0533AM1 | Randomized Blocks | Single Hybrid/Variety | 13.33 | 14,923.8 |
| Bi    | 2016 | Corn | 14.92 | 101.6 | P0157AMX | Randomized Blocks | Single Hybrid/Variety | 13.33 | 14,583.8 |
| Bi    | 2017 | Corn | 14.92 | 101.6 | 197-68STXRIB | Randomized Blocks | Single Hybrid/Variety | 13.33 | 14,138.5 |
| Cr    | 2014 | Corn | 29.55 | 101.6 | DKC46-20RIB; P0094AMX | Randomized Blocks | Alternating Hybrids/Varieties | 13.33 | 11,382.2 |
| Cr    | 2015 | Corn | 29.55 | 101.6 | P0216AM | Randomized Blocks | Single Hybrid/Variety | 13.33 | 10,218.9 |
| Dm2   | 2014 | Corn | 23.47 | 50.8 | P0216AM; P0533AM1 | Randomized Blocks | Alternating Hybrids/Varieties | 6.09 | 12,695.0 |
| Du1   | 2014 | Corn | 24.34 | 76.2 | P0216AM; P0533AM1 | Randomized Blocks | Alternating Hybrids/Varieties | 9.14 | 14,153.5 |
| Du3   | 2016 | Corn | 39.63 | 76.2 | P0216AM; P0533XR | Randomized Blocks | Alternating Hybrids/Varieties | 9.14 | 11,819.9 |
| He15  | 2015 | Corn | 23.75 | 50.8 | P0157AMX; DK52-85 | Randomized Blocks | Alternating Hybrids/Varieties | 6.09 | 11,447.7 |
| Hk    | 2015 | Corn | 35.07 | 50.8 | P0533; P0604 | Randomized Blocks | Alternating Hybrids/Varieties | 6.09 | 12,216.0 |
| Ke    | 2015 | Corn | 31.31 | 76.2 | P0216AM; P0533AM1 | Randomized Blocks | Alternating Hybrids/Varieties | 9.14 | 9,935.2 |
| Mc2   | 2015 | Corn | 15.83 | 50.8 | P0604; P0533AM1 | Randomized Blocks | Alternating Hybrids/Varieties | 6.09 | 12,221.4 |
| Mc3   | 2015 | Corn | 17.78 | 50.8 | P0506AM; P0533AM1 | Randomized Blocks | Alternating Hybrids/Varieties | 6.09 | 14,088.9 |

(Continues)
| Field | Year | Crop | Area  | Row spacing | Hybrid/Variety | Planting rate design | Hybrid/Variety design | Experimental unit grid length/width | Average yield |
|-------|------|------|-------|-------------|----------------|---------------------|----------------------|----------------------------------|--------------|
| Ov29  | 2014 | Corn | 23.97 | 101.6       | P0094AMX       | Randomized Blocks   | Single Hybrid/Variety | 13.33                           | 13,614.8     |
| Ov29  | 2015 | Corn | 23.97 | 101.6       | P0094AMX       | Randomized Blocks   | Single Hybrid/Variety | 13.33                           | 10,143.0     |
| Ov30  | 2014 | Corn | 10.80 | 101.6       | P0094AMX       | Randomized Blocks   | Single Hybrid/Variety | 13.33                           | 13,067.2     |
| Ov30  | 2015 | Corn | 10.80 | 101.6       | P0604AM        | Randomized Blocks   | Single Hybrid/Variety | 13.33                           | 9,153.1      |
| Ri23  | 2015 | Corn | 40.71 | 76.2        | P0216AM;P0533AM1 | Randomized Blocks   | Alternating Hybrids/Varieties | 6.09                           | 9,344.9      |
| Ri27  | 2015 | Corn | 12.74 | 50.8        | P0216AM;P0533AM1 | Randomized Blocks   | Alternating Hybrids/Varieties | 6.09                           | 15,735.3     |
| Sc    | 2015 | Corn | 23.71 | 101.6       | P0157AMX       | Randomized Blocks   | Single Hybrid/Variety | 13.33                           | 12,861.4     |
| Sc    | 2016 | Corn | 23.71 | 101.6       | P0533AM1       | Randomized Blocks   | Single Hybrid/Variety | 13.33                           | 14,639.0     |
| Sc    | 2017 | Corn | 23.71 | 101.6       | P0157AMX       | Randomized Blocks   | Single Hybrid/Variety | 13.33                           | 15,575.6     |
| St    | 2015 | Corn | 47.71 | 76.2        | P0216AM;P0533AM1 | Randomized Blocks   | Alternating Hybrids/Varieties | 9.14                           | 9,012.7      |
| Sw37  | 2015 | Corn | 23.72 | 76.2        | 49-72          | Randomized Blocks   | Single Hybrid/Variety | 9.14                           | 14,590.2     |
| Sw52  | 2016 | Corn | 19.07 | 76.2        | 49-72          | Randomized Blocks   | Single Hybrid/Variety | 9.14                           | 11,608.2     |
| Sw52  | 2017 | Corn | 19.07 | 76.2        | 49-72          | Randomized Blocks   | Single Hybrid/Variety | 9.14                           | 15,067.5     |
| Vc1   | 2016 | Corn | 59.93 | 50.8        | P0506AM;P0533AM1;P0157AMX | Randomized Blocks | Alternating Hybrids/Varieties | 6.09                           | 9,120.3      |
| An6   | 2015 | Soybean | 22.81 | 101.6       | 2108R          | Randomized Blocks   | Single Hybrid/Variety | 13.33                           | 5,545.4      |
| Bb1   | 2016 | Soybean | 21.23 | 101.6       | AG17X6         | Alternating Strips  | Single Hybrid/Variety | 13.33                           | 2,991.7      |
| Do    | 2015 | Soybean | 56.64 | 101.6       | AG3034         | Randomized Blocks   | Single Hybrid/Variety | 13.33                           | 2,612.3      |
| Do    | 2016 | Soybean | 56.64 | 101.6       | AG3334         | Randomized Blocks   | Single Hybrid/Variety | 13.33                           | 2,794.3      |
| Du1   | 2015 | Soybean | 24.34 | 101.6       | P24T19R        | Alternating Strips  | Single Hybrid/Variety | 13.33                           | 3,225.6      |
| Du3   | 2015 | Soybean | 39.63 | 101.6       | P24T19R;92Y91  | Alternating Strips  | Sections              | 13.33                           | 3,215.1      |
| He15  | 2014 | Soybean | 23.75 | 101.6       | P92Y12;P92Y51  | Randomized Blocks   | Alternating Hybrids/Varieties | 13.33                           | 3,627.3      |
| He15  | 2016 | Soybean | 23.75 | 101.6       | P24T05         | Randomized Blocks   | Single Hybrid/Variety | 13.33                           | 2,115.4      |
| Hk    | 2014 | Soybean | 35.07 | 127.0       | P92Y51         | Randomized Blocks   | Single Hybrid/Variety | 16.67                           | 4,219.0      |
| Ke    | 2016 | Soybean | 31.31 | 101.6       | P19T01R        | Alternating Strips  | Single Hybrid/Variety | 13.33                           | 3,410.3      |
| Lb    | 2016 | Soybean | 28.53 | 50.8        | P19T/P01T;P01T04R15 | Randomized Blocks | Sections              | 6.09                           | 1,508.3      |
| Mc1   | 2014 | Soybean | 25.17 | 101.6       | AG3030         | Randomized Blocks   | Single Hybrid/Variety | 13.33                           | 4,251.1      |

(Continues)
| Field | Year | Crop     | Area  | Row spacing | Hybrid/Variety | Planting rate design | Hybrid/Variety design | Experimental unit grid length/width | Average yield |
|-------|------|----------|-------|-------------|----------------|----------------------|------------------------|----------------------------------|---------------|
| Mc1   | 2018 | Soybean  | 25.17 | 101.6       | AG3334; SG1776 | Randomized Blocks    | Sections               | 13.33                           | 5,205.1        |
| Mc2   | 2014 | Soybean  | 15.83 | 101.6       | P93Y22         | Randomized Blocks    | Single Hybrid/Variety  | 13.33                           | 3,953.2        |
| Mc2   | 2016 | Soybean  | 15.83 | 101.6       | AG2035         | Randomized Blocks    | Single Hybrid/Variety  | 13.33                           | 2,549.6        |
| Mc3   | 2014 | Soybean  | 17.78 | 101.6       | AG3030         | Randomized Blocks    | Single Hybrid/Variety  | 13.33                           | 5,156.2        |
| Ov29  | 2018 | Soybean  | 23.97 | 76.2        | SG1863XT; P16A35X | Randomized Blocks    | Sections               | 9.14                            | 6,044.4        |
| Ov30  | 2018 | Soybean  | 10.80 | 76.2        | SG1863XT; P16A35X | Randomized Blocks    | Sections               | 9.14                            | 6,085.9        |
| St    | 2014 | Soybean  | 47.71 | 101.6       | P92Y51         | Alternating Strips   | Single Hybrid/Variety  | 13.33                           | 4,116.6        |
| Sw37  | 2014 | Soybean  | 23.72 | 76.2        | 2031           | Randomized Blocks    | Single Hybrid/Variety  | 9.14                            | 3,547.4        |
| Sw37  | 2016 | Soybean  | 23.72 | 76.2        | AG2031         | Randomized Blocks    | Single Hybrid/Variety  | 9.14                            | 3,522.0        |
| Sw52  | 2015 | Soybean  | 19.07 | 76.2        | AG2431         | Randomized Blocks    | Single Hybrid/Variety  | 9.14                            | 4,027.1        |
| Tp80  | 2016 | Soybean  | 31.27 | 76.2        | Unknown        | Randomized Blocks    | Single Hybrid/Variety  | 9.14                            | 3,018.7        |
| Vc1   | 2015 | Soybean  | 59.93 | 101.6       | P22T41; P92Y51 | Randomized Blocks    | Sections               | 13.33                           | 2,294.0        |
| Vc1   | 2017 | Soybean  | 59.93 | 101.6       | P22T41; AG2035 | Randomized Blocks    | Sections               | 13.33                           | 2,765.7        |
FIGURE 2  Statewide departures from 1981-2010 monthly temperature and precipitation averages during the 2014–2018 growing seasons

FIGURE 3  Field designs for planting rate and hybrid/variety
For soybean, the same 0.81-ha target planting rate block system was used with the following target rates: 197,684 (80,000), 296,526 (120,000), 395,368 (160,000), and 494,210 (200,000) seeds ha\(^{-1}\) (ac\(^{-1}\)). Rates were assigned to blocks at random for 20 of the 25 site-years. Because variable rate soybean planters were not available for the remaining 5 site-years, target planting rates were assigned to alternating rows or columns of the blocks, enabling adjustments to be made to the planting rate between passes of the planter (Figure 3). A total of 7 of the 25 site-years were sown with two soybean varieties. A split soybean planter was only available for site-year He15_2014, which was sown with alternating varieties. The remaining 6 site-years were sown such that each variety occupied a section of the field (Figure 3).

A central aim of the experiment was to integrate with the cooperating growers’ existing operations to the extent possible. Therefore, cooperating growers were requested to continue their standard protocols for fertility and pest management. The row spacing was uniform within each site-year and was likewise in accordance with the cooperating growers’ standard practices. The row spacings used were 50.8, 76.2, 101.6, and 127.0 cm. The target planting rates were chosen to represent the typical range of rates growing in the region were using at the beginning of the experiment.

2.3 Experimental unit grid

The 0.81-ha target planting rate block system was developed to minimize error associated with transitioning between planting rates. There are, however, limitations to treating the 0.81-ha target planting rate blocks as the experimental units. Blocks that fell along the edges of the field were often irregularly shaped (as seen in Figure 3). As a result, the number of full-sized 0.81-ha blocks in smaller fields was low and insufficient for analysis. Furthermore, due to the high variability of the topographical and soil features observed in some fields, some of the 0.81-ha blocks contained a wide range of variability. Lastly, because split planters were used to sow multiple hybrids/varieties in some site-years, individual blocks contained more than one hybrid/variety. Therefore, by using the blocks as the experimental units, it would not be possible to evaluate the interactions of hybrid/variety with planting rate, soil features, and topographical characteristics as well as their effects on yield.

To address these issues, a finer resolution grid-based system was developed to serve as the experimental units. Square grids were created in QGIS (QGIS Development Team, 2020) with the length and width of the grid cells set equal to the width of the planter (Table 1), unless a split planter was used to sow multiple hybrids/varieties (Figure 4). In that case, the length and width of the grid cells were set to half the width of the planter. Each grid cell was assigned a unique identifier (ID). All data types were aggregated to the experimental unit grid to resolve the data into a tabular format for analysis.

2.4 Data types and quality control

For each site-year, five spatial data layers were used to assess the relationship between yield, planting rate, hybrid/variety, topographical features, and soil characteristics. The layers were as follows: 1) harvest, 2) target/as-applied planting rate, 3) topography, 4) grid soil sampling, and 5) Natural Resources Conservation Service (NRCS) Soil Survey (Figure 5).

The harvest point layer contained the levels of dry volume yield harvested throughout the field. In the form of polygons, the target planting rate layer contained the randomized 0.81-ha block design with which the trials were sown, while the as-applied planting rate point layer recorded the precise rates applied by the planter as well as where each hybrid/variety was sown. These three layers were used in conjunction with one another to perform various quality control measures to the as-applied planting rate and harvest data (Supplemental Figure S1A-F). Briefly, the headlands were removed from both the as-applied planting rate and harvest layers in QGIS. The remaining data-points were then assigned ID numbers corresponding to the experimental unit grid cell in which they appeared. The as-applied planting rate and harvest layers were then analyzed in R (R Core Team, 2019) to remove outlying grid cells based on the mean, variance, and number of data points within each cell. For grid cells that spanned more than one of the 0.81-ha randomized blocks, the target planting rate could not be resolved to a single rate. Those grid cells were therefore removed from the analysis. Similarly, grid cells were also removed if they contained data corresponding to more than one hybrid/variety. Finally, grid cells in which the as-applied planting rate deviated from the target planting rate beyond a threshold of ±1,011.7 seeds ha\(^{-1}\) for corn and ±8,093.7 seeds ha\(^{-1}\) for soybean were also removed.

Topographical variables for absolute elevation, slope, aspect, and total curvature were derived from elevation measurements recorded by the GPS systems aboard the planting machinery (Supplemental Figure S1G). Using the as-applied planting rate layer, elevation means were calculated in QGIS using all of the data points within each experimental unit grid cell. Slope, aspect, and total curvature were then calculated in R using 3 × 3 neighborhoods of experimental unit grid cells following Zevenbergen and Thorne (1987).
Although fine-scale spatial data is generally desirable for obtaining model estimates with higher accuracy in precision agriculture applications (Bishop, Minasny, & McBratney, 2006), fine local variability in elevation data can produce considerable noise when computing topographical features using standard 3 × 3 neighborhood algorithms, thereby misrepresenting the terrain structure (Gallant & Wilson, 2000). Initial examination of the derived total curvature profiles, which quantify the direction and degree of convexity/concavity and were developed for the 6.09–15.24 m experimental unit grid cells, showed limited visible spatial trends and high variability. Lashermes, Foufoula-Georgiou, and Dietrich (2007), Roering, Marshall, Booth, Mort, and Jin (2010), and Hurst, Mudd, Walcott, Attal, and Yoo (2012) demonstrated that a spatial scale of 12–15 m or greater is required to stabilize variability in calculated curvature. To address this issue, elevation data were upscaled to 2 × 2 neighborhoods of experimental unit grid cells, thereby increasing the spatial scale to 12.18–30.48 m, and the resulting grid was used for calculation of total curvature according to Zevenbergen and Thorne (1987) (Supplemental Figure S1G).

Integrated Ag Services (Mildford Center, OH) conducted grid soil sampling along a 0.20-ha grid pattern, and Spectrum Analytic (Washington Court House, OH) provided the topsoil analysis. Point data on up to 12 soil features were provided for each site: aluminum (Al), phosphorus (P), potassium (K), potassium saturation (KSt), calcium (Ca), calcium saturation (CaSt), magnesium (Mg), magnesium saturation (MgSt), pH, buffer pH (BpH), cation exchange capacity (CEC), and soil organic matter (OM). K, Mg, and CEC were recorded for all 27 sites (57 site-years). The remaining variables were unavailable for ≤4 sites (≤12 site-years) with the exception of Al, which was available for only 17 of the 27 sites (37 site-years). Using the “gstat” package in R, the soil sample measurements were block kriged to the experimental unit grid (Supplemental Figure SIG). If a semivariogram for a particular variable could not be fit due to low spatial autocorrelation, the values for that variable were set to missing.
The NRCS Soil Survey layer in the form of polygons was obtained for each site using the SSURGO database (Soil Survey Staff, 2015). The soil type attribute was extracted for each experimental unit grid cell using QGIS. If an experimental unit grid cell spanned multiple NRCS Soil Survey polygons with differing reported soil types, the soil type variable for that cell was set to missing (Supplemental Figure S11).

Final datasets containing variables describing the yield, target and as-applied planting rates, hybrids/varieties, topography, grid soil sampling, and NRCS Soil Survey soil type were merged into a combined data frame in R according to the experimental unit grid cell IDs. Excluding yield as the response variable, the dimensions of the final datasets for each site-year used for model fitting were \( n \times p \) where \( n \) is the number of experimental unit grid cells and \( p \) is the number of predictors.

### 2.5 Variable correlation, ANOVA, and linear regression

Pearson’s correlations were calculated between all variables within each site-year. Treatment effects for planting rate, hybrid/variety, and soil type were estimated using ANOVA. Yield was regressed on each topographical and soil sampling variable to estimate the percentage of yield variation explained (\( R^2 \)). Yield was also regressed on planting rate, hybrid/variety, and their interaction to assess their collective influence on yield in a linear context. All correlation, ANOVA, and linear regression analyses were carried out in R.

### 2.6 Random forest regression variable importance

Random forest regression models were fit to 1) calculate variable importance, 2) assess yield prediction accuracy, and 3) develop optimized planting rate designs. Forests were grown using the “cforest” function of the “party” package for R (Hothorn, Hornik, & Zeileis, 2006; Strobl, Boulesteix, Kneib, Augustin, & Zeileis, 2008; Strobl, Boulesteix, Zeileis, & Hothorn, 2007). Through simulation, Strobl et al. (2007) demonstrated that bias could occur in the variable importance measures when predictors vary in their scale of measurement or, for categorical variables, in their number of levels. To avoid bias, Strobl et al. (2007) proposed “cforest” as an alternative implementation of the commonly used regression functions in the “randomForest” package. The algorithm is based on conditional inference trees and applies subsampling without replacement, which was shown to produce reliable variable importance measures when predictors varied in their scale of measurement or number of categories.

Random forest regression models were fit using each site-year as a training set in order to evaluate variable importance within individual site-years. The hybrid/variety term was only included in the model for those site-years in which more than one hybrid/variety was sown. Variable importance measures for each predictor were calculated for each site-year using the “varimp” function in the “party” package for R. Briefly, variable importance is calculated by permuting each predictor variable to determine the difference in prediction accuracy before and after permutation. Since the magnitudes of the variable importance measures are dependent on the mean squared error of the model, which varied among site-years, the variable importance measures were scaled using the “scale” function in R to enable comparisons across site-years.

In addition to analyses of individual site-years, a “composite” random forest model was fit for each crop using data from all available site-years. The analysis of data from multiple site-years permitted the inclusion of variables describing row spacing, monthly average temperature and precipitation, and the latitude and longitude of the site locations. These variables were uniform within each individual site-year but may play an important role in modulating yield either by themselves or as part of interactions with other variables when considering multiple site-years together. Yield data was centered and scaled within each site-year using the “scale” function in R for use in the composite model. Because relative elevation within sites would likely be more informative than absolute altitude for yield prediction across multiple site-years, the elevation data was similarly scaled within each site-year using the “scale” function in R.

A subset of the full dataset was utilized to evaluate the effect of the \( mtry \) hyperparameter on model accuracy (data not shown). In summary, prediction models were trained for each site-year using all possible values for \( mtry \) [1,...,\( p \)]. The fitted models were then applied to the remaining site-years in the subset planted with the same crop. Overall, prediction accuracies showed minimal variation among the evaluated \( mtry \) values. For each fitted model, there was no \( mtry \) value that consistently provided the highest accuracies across the predicted site-years. Therefore, the default \( mtry \) value of \( p/3 \) was used for this study (Breiman, 2001). The \( ntree \) value was set to 1000.

### 2.7 Yield prediction with random forest regression

Since empirical validations of the proposed variable rate planting optimization strategy have not yet been completed, examining model fit and prediction accuracy
presents an opportunity to preliminarily assess the empirical utility of the proposed strategy. Models with poor fit or minimal ability to predict yield may be expected to generate optimized variable rate planting designs with limited value. Model fit and yield prediction accuracy were therefore used as initial metrics for the evaluation of the optimization strategy.

For the random forest regression models fit using data from each individual site-year, the $R^2$ value was retained as a metric to describe model fit. To test the prediction accuracy of the composite random forest modeling approach, a leave-one-out scheme was applied such that the composite model was trained using data from all site-years for a given crop except one. The fitted model then was applied using the remaining site-year as the test set to predict yield. While uncommon, if the soil type for a given experimental unit grid cell in the test set was unique to that site-year and therefore was not observed in the training set, the soil type for that grid cell was set to the most occurring soil type in the test set as random forest regression does not accommodate missing values. Similarly, if a hybrid/variety sown in the test set site-year was not observed in any other site-years, the hybrid/variety term was omitted from the model. As with model fitting for variable importance calculations, the yield and elevation data were scaled within each site-year using the “scale” function in R. Prediction accuracy was assessed as the Pearson’s correlation between the predicted and observed values for yield.

2.8 | Optimized variable rate planting designs

Optimal planting rate designs were developed for each site-year using 1) random forest regression models trained on each individual site-year’s data and 2) the composite random forest regression model fit using the leave-one-out scheme. The former represents a scenario in which the randomized block design was applied locally to characterize the planting rate response within the site, while the latter mimics a situation in which the site has not been previously tested for planting rate response. In the latter case, the optimized design is based on randomized block experimental data from all other site-years. For each experimental unit grid cell, the trained models were used to predict yield at each of the four tested planting rates. The planting rate that provided the highest yield prediction within each grid cell was then identified to create the optimized design.

3 | RESULTS

3.1 | Grain yield summary statistics

Corn yields were lowest and most variable during the 2015 growing season, which was characterized by extremely wet conditions during May and June followed by dry conditions from early July through mid-Sept. (Figure 2). Scouting reports recorded in early July noted multiple sites with standing water in some low-lying sections (data not shown). Corn yields were likewise variable in 2016, due to severe drought conditions in June and July in western and central New York (Figure 2). For soybean, 2015 yields were the most variable with respect to within-site-year standard deviation, while yields were the lowest in 2016.

3.2 | Effect of planting rate and hybrid/variety on yield

Although ANOVA for planting rate showed a significant difference in yield between at least one pair of planting rate levels ($p < .05$) for 31 out of 32 site-years for corn, planting rate on average explained only 4.1% of yield variation and did not exceed 17.4% for any site-year (Figure 6). For soybean, a significant ($p < .05$) treatment effect for planting rate was observed in all 25 site-years, but the average amount of yield variation explained was likewise low at 10.8%. In contrast with the corn site-years, however, planting rate explained a large amount of yield variation in some site-years (up to 40.8%) for soybean.

Significant treatment effects ($p < .05$) were less prevalent for hybrid/variety, appearing in 13 out of the 19 site-years for which multiple hybrids of corn were sown and for 5 out of 9 site-years for which multiple varieties of soybean were sown. As with planting rate, the amount of yield variation explained by the hybrid/variety variable was very low for corn, averaging 1.0% with a range of 0.0 to 4.1%. For soybean, hybrid/variety explained 8.5% of yield variation on average, though this ranged up to 30.1%. Notably, hybrid/variety explained over 25% of variation in soybean yields in 2 site-years, Vc1_2015 and Mc1_2018, in which the differences in performance between the varieties grown were 771 and 395 kg ha$^{-1}$, respectively.

Considering the composite effects of planting rate, hybrid/variety, and their interaction, as before the percent of yield variation explained was much greater for soybean than corn. These variables explained 2.8 and 22.8% of yield variation on average in corn and soybean, respectively. Most notably, these variables explained 61.9% of the variation in yield for site-year Vc1_2015 with both varieties
FIGURE 6 $R^2$ values of yield regressed onto each variable in each site-year with simple linear regression

P22T41 and P92Y51 showing positive linear responses to planting rate.

### 3.3 Topographical and soil summary statistics

The 27 field sites tested varied greatly in their topographical and soil feature profiles. The lowest and highest elevations recorded were 105 and 368 masl, and the difference between the lowest and highest points at a site was 16 m on average. The largest gradient within a single site with respect to elevation was observed at Ke, which ranged from 116 to 164 masl for a difference of 48 m. The United States Environmental Protection Agency classifies crop production on slopes $\geq 9\%$, or $5.14^\circ$, as “agricultural land cover on steep slopes.” Of the 27 field sites, 19 contained areas of sloped terrain at grades of $5.14^\circ$ or above.

Soils in New York State are generally acidic. The soil grid sampling data showed that the pH ranged from 5.8 to 7.2 for the tested field sites. High levels of aluminum were likewise observed, ranging from 657 to 767 mg/kg. CEC levels were moderate, ranging from 6.2 to 16.5 cmolc/kg, while soil OM was moderate to low, averaging 2.3%. According to the NRCS Soil Survey, 48 unique soil types were present across the tested sites, with 13 appearing in only one site. Each site contained an average of 6 different soil types, with type “Cazenovia” as the most common, appearing in 10 of the sites.

In the context of yield prediction and variable rate planting optimization, it would be desirable for the correlation for a pair of topographical or soil characteristics to have a high magnitude mean and low standard deviation across the testing sites, indicating that the two variables are not only strongly related, but their relationship is consistent across sites. A wide variety of relationships were observed between the measured variables (Figure 7). As expected, pH and Al were observed to have a consistent, strong negative correlation, while CEC and OM were positively correlated. The relationship between OM and elevation was observed to be slightly negative, on average, although highly variable across the sites.

When considering each variable’s relationship with yield, soil type had the most consistent relationship for both corn and soybean. It explained an average of 6.8 and 10.4% of the variation in corn and soybean yields, respectively, across the site-years. Overall, the topographical variables explained relatively low levels of yield variation, averaging 2.4 and 4.0% for corn and soybean, respectively, though the elevation, slope, and curvature variables explained more than 10% of the variation in yield in at least one site-year. For the soil sampling variables, each variable explained between 2.0 and 8.0% of yield variation on average, though each variable explained more than 15.0% of the yield variation in at least one site-year. In summary, all variables typically explained moderate amounts of yield variation within some site-years, but
FIGURE 7  Means and standard deviations of the Pearson’s correlations between each pair of variables across the 27 field sites

FIGURE 8  Scaled random forest regression variable importance measures for each variable in each site-year

these relationships were not consistent across site-years overall.

3.4  Random forest regression variable importance

Random forest regression models were fit within each site-year to evaluate variable importance. Planting rate consistently had higher values for random forest regression variable importance for both crops but particularly for soybean (Figure 8). Likewise, the hybrid/variety variable was of high importance in some site-years for soybean, though the number of site-years in which multiple soybean varieties were sown was limited. Of the topographical variables, elevation and curvature were important for both crops in many site-years. The soil sampling variables were
more frequently of greater importance for corn than for soybean. Overall, most of the variables were identified as highly important in one or more site-years, though the level of importance was inconsistent across site-years.

Random forest regression variable importance measures were also assessed for the composite models, which contained data from all site-years available for each crop. For corn, soil type was the most important variable, and its magnitude was twice that of Mg, which was the second most important predictor (Figure 9). The hybrid/variety and planting rate variables showed similar levels of importance to Mg. Of the remaining predictors, elevation and some of the soil sampling variables had moderate levels of importance, followed by row spacing, the remaining topographical variables, latitude/longitude, and the weather variables. For soybean, the most important variables were hybrid/variety and planting rate, which showed similar levels of importance. Soil type was the third most important variable with a moderate level of importance, followed by all remaining variables with similarly low levels of importance.

3.5 Variable rate planting optimization

Optimized planting rate designs were developed from the random forest regression models to maximize predicted yields (Figure 10). In general, the optimized designs were based on each individual site-year’s data favored the higher rates for both corn and soybean. For corn, the four planting rates in increasing order were assigned to, on average, 18.6, 21.8, 33.3, and 28.2% of the field. For soybean, optimized designs overwhelming favored the highest rate, which was assigned to 60.1% of the field, on average, while the remaining three rates in increasing order were assigned to 9.7, 13.8, and 18.9% of the field, on average. For some soybean site-years, the optimized design effectively represented a fixed rate planting strategy at the highest rate.

Corn and soybean were evaluated in more than one year in 7 and 6 of the field sites, respectively. The optimized designs showed considerable differences depending on which year was used as the training set. For corn, the proportion of grid cells assigned to the same rate across years was 31.5% on average, ranging from 12.7 to 44.3%. For soybean, the average was 33.2% with a range of 5.3 to 74.3%.

Optimized designs were also developed for each site-year using the composite model, which was fit with data from all other site-years evaluated under the same crop. For corn, the four planting rates were distributed similarly compared with the optimizations developed with data from individual site-years. The four planting rates, in increasing order, were assigned to on average 25.4, 18.6, 34.3, and 21.7% of the field. For soybean, the optimized designs were again overwhelmingly assigned to the higher rates. On average, the four rates in increasing order were
assigned to 3.1, 8.4, 10.7 and 77.8% of the grid cells. When comparing optimizations developed using each individual site-year’s data versus those developed by the composite model, the proportion of grid cells assigned to the same rate across the two optimized designs averaged 33.2% for corn and 56.4% for soybean.

3.6 Yield prediction with random forest regression

For random forest regression models trained using the data from individual site-years, the $R^2$ values were moderate for most site-years. The $R^2$ ranged from 0.16 to 0.78 with a mean of 0.47 for corn, and for soybean, values ranged from 0.17 to 0.72 with a mean of 0.50 (Table 2). Yield prediction accuracies of the composite model ranged from −0.08 to 0.74 with a mean of 0.24 for corn, while accuracies for soybean ranged from −0.07 to 0.58 with a mean of 0.23 (Table 2). Notably, nearly all site-years with prediction accuracies greater than 0.3 were for sites that had been evaluated under the same crop in multiple years. The highest accuracies were observed for the site “Sc”, which was evaluated under corn in 2015, 2016, and 2017.

4 DISCUSSION

In the context of linear modeling, planting rate explained relatively low levels of the yield variation. In most site-years, there was no consistent yield advantage of one of the tested planting rates over the others. In contrast, planting rate ranked highly in terms of random forest variable importance for both corn and soybean when considering all site-years individually and in combination. While perhaps seemingly contradictory, this contrast in results reflects fundamental differences between the linear modeling and random forest regression approaches and may provide support for the use variable rate planting. In the context of linear modeling, had a consistent site-wide “optimum” planting rate been realized, the most effective management strategy would therefore be to sow at a fixed rate according to the planting rate that provided the greatest yield advantage. The lack of an apparent site-wide optimum planting rate may instead suggest that subsections of the field may have different optimums, according to underlying field conditions. Comparatively, the results of random forest regression, which considers all variables simultaneously and is capable of modeling complex nonlinear interactions, showed that planting rate ranked highly in terms of variable importance for both corn and soybean. Together, these results provide support for the hypotheses that 1) there may be important complex interactions of planting rate with topographical, soil, and other variables that are driving responses in yield and that 2) it may be possible to identify an optimal planting rate for the conditions in each area of the field in order to maximize yields.

These observations also highlight the importance of selecting an appropriate modeling approach for large agro-nomic datasets. A similar study found that artificial neural networks, which—like random forest regression—are capable of modeling nonlinear interactions, were superior to linear modeling approaches when predicting corn yields using soil and landscape features (Miao, Mulla, & Robert, 2006). Random forest regression is among a suite of established and emerging techniques in machine learning that are becoming more widely used in agricultural research to model complex systems (Chlingaryan, Sukkarieh, & Whelan, 2018; Häring, Dietz, Osenstetter, Koschitzki, & Schröder, 2012; Henderson, Bui, Moran, & Simon, 2005; Liakos, Busato, Moshou, Pearson, & Bochtis, 2018; Xiong et al., 2014). As the amount and variety of environmental, climate, management, and economic data available to growers increase through innovations in precision agriculture, it will be important for the community to consider the limitations of linear modeling and to begin to leverage the advances made in machine learning-based approaches.

In addition to the planting rate variable, the Mg soil sampling term was observed to be highly important in the composite random forest regression model for corn, though it showed lower levels of importance and low $R^2$ values when site-years were analyzed individually. Magnesium can be a limiting factor for corn growth and development on acidic soils, though magnesium deficiencies that produce significant impacts yield are not common (Fox & Piekiele, 1984; Foy & Barber, 1958; Key & Kurtz, 1960; Wang et al., 2020). A possible explanation for the contrast in results between the analysis of individual site-years and the composite model may be that the Mg variable modulates corn yields through interactions with the weather, row spacing, and/or latitude/longitude variables, which were not assessed in the individual site-year analyses as they did not vary. Alternatively, while Mg did not have a strong relationship with yield within the individual site-years, the consideration of multiple site-years captured a wider distribution of variability in Mg that may have driven a response in yield alone and/or in combination with other variables. The identification of this trend highlights the potential value of leveraging longer-term multi-location datasets to uncover important relationships between variables that may not be detected when considering site-years individually.

Both linear modeling and random forest regression identified soil type as having a strong and consistent relationship with yield across site-years for both crops. This result is in agreement with Smidt et al. (2016), which leveraged a range of topographical and soil variables to model...
**TABLE 2** $R^2$ of the random forest regression models fit for each individual site-year and prediction accuracies for the leave-one-out composite model scheme

| Field | Year | Crop   | Individual site-year model $R^2$ | Composite model prediction accuracy |
|-------|------|--------|----------------------------------|-------------------------------------|
| An6   | 2014 | Corn   | 0.38                             | 0.14                                |
| An8   | 2014 | Corn   | 0.36                             | 0.25                                |
| An8   | 2015 | Corn   | 0.72                             | 0.21                                |
| Ba1   | 2015 | Corn   | 0.58                             | 0.16                                |
| Be2   | 2014 | Corn   | 0.29                             | 0.15                                |
| Bi    | 2015 | Corn   | 0.16                             | 0.42                                |
| Bi    | 2016 | Corn   | 0.32                             | 0.58                                |
| Bi    | 2017 | Corn   | 0.25                             | 0.20                                |
| Cr    | 2014 | Corn   | 0.29                             | 0.38                                |
| Cr    | 2015 | Corn   | 0.46                             | 0.30                                |
| Dm2   | 2014 | Corn   | 0.51                             | 0.28                                |
| Du1   | 2014 | Corn   | 0.24                             | 0.17                                |
| Du3   | 2016 | Corn   | 0.55                             | 0.01                                |
| Hei5  | 2015 | Corn   | 0.43                             | 0.27                                |
| Hk    | 2015 | Corn   | 0.58                             | 0.16                                |
| Ke    | 2015 | Corn   | 0.51                             | 0.12                                |
| Mc2   | 2015 | Corn   | 0.53                             | -0.07                               |
| Mc3   | 2015 | Corn   | 0.41                             | 0.27                                |
| Ov29  | 2014 | Corn   | 0.32                             | 0.36                                |
| Ov29  | 2015 | Corn   | 0.50                             | 0.38                                |
| Ov30  | 2014 | Corn   | 0.24                             | 0.26                                |
| Ov30  | 2015 | Corn   | 0.52                             | 0.51                                |
| Ri23  | 2015 | Corn   | 0.78                             | 0.08                                |
| Ri27  | 2015 | Corn   | 0.49                             | 0.02                                |
| Sc    | 2015 | Corn   | 0.70                             | 0.73                                |
| Sc    | 2016 | Corn   | 0.63                             | 0.74                                |
| Sc    | 2017 | Corn   | 0.59                             | 0.67                                |
| St    | 2015 | Corn   | 0.56                             | 0.07                                |
| Sw37  | 2015 | Corn   | 0.52                             | -0.08                               |
| Sw52  | 2016 | Corn   | 0.72                             | 0.02                                |
| Sw52  | 2017 | Corn   | 0.46                             | -0.01                               |
| Vc1   | 2016 | Corn   | 0.53                             | -0.06                               |
| An6   | 2015 | Soybean | 0.45                              | 0.27                                |
| Ba1   | 2016 | Soybean | 0.52                              | 0.11                                |
| Do    | 2015 | Soybean | 0.66                              | 0.58                                |
| Do    | 2016 | Soybean | 0.42                              | 0.14                                |
| Du1   | 2015 | Soybean | 0.38                              | 0.31                                |
| Du3   | 2015 | Soybean | 0.50                              | 0.50                                |
| Hei5  | 2014 | Soybean | 0.51                              | 0.26                                |
| Hei5  | 2016 | Soybean | 0.49                              | 0.16                                |
| Hk    | 2014 | Soybean | 0.50                              | 0.33                                |
| Ke    | 2016 | Soybean | 0.51                              | -0.02                               |
| Lb    | 2016 | Soybean | 0.34                              | 0.06                                |

(Continues)
soybean yields in the context of a variable rate planting experimental design. Soil map unit, which describes the soil type, showed the highest level of variable importance in random forest regression. This result was attributed to the wide range of diversity among the 39 unique soil types observed at the experiment sites. The sites evaluated in the present study similarly represented a broad range of soil diversity: a total of 48 soil types were observed across the tested locations. The northeast region of the United States has been shown to exhibit higher variability in a range of physical, biological, and chemical soil indicators compared to the Midwest and Mid-Atlantic regions (Fine, Ristow, Schindelbeck, & van Es, 2016). The extensive variability of New York soils and their observed relationships with yield are encouraging for the development of variable rate planting strategies for New York. Compared with the acquisition of other data types used in this study, the soil type layer is readily accessible from the NRCS, and its procurement would not require a monetary investment from growers.

The proposed random forest regression model was extended to build optimized variable rate planting designs for maximizing yields. Marked differences were observed between corn and soybean with respect to the resulting optimized designs. For corn, although higher rates were generally favored, all four tested planting rates were assigned to considerable parts of the field in most of the optimized designs. For soybean, many of the optimized designs were predominantly assigned to the highest of the four rates. While this may suggest that fixed rate planting may be more suitable for those sites, it is also important to consider the trade-off between marginal yield gains from higher planting rates and the cost of the seed, as it would be no longer profitable to plant at higher rates if revenue from the yield gains sold at the market price does not exceed the cost of the additional seed. The proposed random forest-based approach may also be utilized to optimize planting rate designs for maximized profit. An example of optimized designs based on theoretical seed cost-to-market price ratios is shown in Figure 10, demonstrating that lower planting rates are assigned to greater proportions of the field as the cost-price ratio becomes less favorable to growers. This highlights the importance of leveraging market trends to inform management decisions concerning variable rate technologies, and the proposed approach is fully capable of accommodating user-defined commodity price and seed cost information to optimize planting rate designs for profit.

For sites that were evaluated under the same crop in multiple years, the resulting optimized planting rate designs developed with data from each individual site-year showed considerable differences depending on which year’s data was used to train the model. These inconsistencies are likely due to variations in environmental conditions across years (Reeves & Cox, 2013; Wells, 1993) and/or differential sensitivities to planting rate among the hybrids/varieties sown (Agudamu & Shiraiwa, 2016).

Year-to-year variation in weather conditions and its impact on yield pose a challenge for the development of variable rate planting strategies. The present study sampled a range of wet and dry conditions, and the proposed composite model approach aimed to capture the

**Table 2 (Continued)**

| Field | Year | Crop | Individual site-year model $R^2$ | Composite model prediction accuracy |
|-------|------|------|---------------------------------|-----------------------------------|
| Mc1   | 2014 | Soybean | 0.28 | 0.14 |
| Mc1   | 2018 | Soybean | 0.56 | 0.26 |
| Mc2   | 2014 | Soybean | 0.17 | 0.12 |
| Mc2   | 2016 | Soybean | 0.69 | -0.07 |
| Mc3   | 2014 | Soybean | 0.54 | 0.38 |
| Ov29  | 2018 | Soybean | 0.56 | 0.10 |
| Ov30  | 2018 | Soybean | 0.50 | 0.20 |
| St    | 2014 | Soybean | 0.32 | 0.22 |
| Sw37  | 2014 | Soybean | 0.56 | 0.37 |
| Sw37  | 2016 | Soybean | 0.65 | 0.17 |
| Sw52  | 2015 | Soybean | 0.53 | 0.03 |
| Tp80  | 2016 | Soybean | 0.60 | 0.21 |
| Vc1   | 2015 | Soybean | 0.72 | 0.53 |
| Vc1   | 2017 | Soybean | 0.58 | 0.32 |
effects of weather and their interaction with plant rate and other variables through the incorporation of terms describing monthly average temperature and precipitation. However, in order to apply this model to develop an optimized variable rate planting design for a target site, weather data corresponding to that site must be supplied to the model. Many of the currently available variable rate input management schemes for chemicals and irrigation actively leverage weather information (Hedley & Yule, 2009; Melkonian, Es, DeGaetano, & Joseph, 2008), but the advantage of these approaches is that they can be designed to respond to in-season weather status, whereas planting rate decisions must be made at the beginning of the season typically with limited prior information about eventual weather outcomes. Through a simulation study of variable rate scenarios in soybean, Paz, Batchelor, and Tylka (2001) found that the net yield gain associated with variable rate planting was dependent on a priori knowledge of the weather status of the upcoming season. While highly accurate a priori weather data are difficult to acquire, long-range weather forecasts for the upcoming season or historical observed data could be supplied to the model instead.

Short product life cycles also complicate the design and implementation of variable rate planting technologies. Results from this study and other works show that corn hybrids and soybean varieties can differ in their response to planting rate as well as in their interactions with other variables (Agudamu & Shiraiwa, 2016; Modarres et al., 1998; Rutger & Crowder, 1967). To address this, some seed companies conduct planting rate trials across a range of environments and publish optimal rate recommendations for the hybrids and varieties they release. This information cannot, however, be customized to the conditions of an individual site, nor does it provide insight into effective variable rate strategies to impose within a site. The proposed random forest regression modeling approach cannot develop variable rate planting strategies for specific hybrids/varieties if they have not been previously tested. However, empirical testing of the variable rate response of each hybrid/variety, as in this study, is time-consuming, and advances in agricultural biotechnologies and molecular breeding have accelerated the rate at which improved hybrids and varieties enter the market. A 2010 study estimated that US corn hybrids reach maximum sales within two-to-three years before declining in the following years (Magnier, Kalaitzandonakes, & Miller, 2010). New hybrids or varieties may therefore become obsolete during the time required to conduct thorough on-farm trials of variable rate response if carried out by the grower. Greater integration of variable rate testing in the germplasm development pipeline may serve to address this issue. Alternatively, variable rate optimization models could be trained using data collected on a range of hybrids/varieties so as to become “agnostic” the hybrid or variety sown and base planting rate optimizations on main effects, though the relative importance of the hybrid/variety term in the random forest regression models developed in this study would suggest that the resulting optimized variable rate designs may be of limited value.

In addition, the expense of conducting on-farm trials of variable rate response is not trivial. Grid soil sampling was carried out at each field site, and research associates trained in data science were needed to process and analyze the data. In addition, the current approach requires that at least some growers plant and evaluate variable rate randomized block designs to develop a training set. Once this training set has been created, however, the ability to build optimized designs for sites that have not undergone variable rate randomized block testing would be of great value. However, results showed that the optimized variable rate planting designs derived using a site-year’s own data were considerably different from those developed using the data from all other site-years. In addition, leave-one-out yield prediction accuracies calculated for the composite model were relatively low for most site-years. The highest levels of prediction accuracy were achieved when the training set contained different years of data from the same site as that of the test set. Considering also that the $R^2$ values of the random forest regression models based on individual site-years were moderate, local randomized block testing is perhaps likely to provide the most accurate optimizations; however, there may be alternative approaches for building the most useful training sets with which to develop a variable rate planting optimization for an untested site. For example, provided yield, topographical, and soil data are available for an untested site, random forest regression models predicting yield could be fit, and the resulting variable importance measures could be utilized to identify the site-year(s) evaluated under the randomized block design that have similar profiles of random forest variable importance measures to use as a training set. Alternatively, data from multiple site-years could be combined in an optimal fashion to maximize prediction accuracy. Heslot, Jannink, and Sorrells (2013) provided an innovative approach for this type of training set optimization in the context of plant breeding that identifies and removes less predictive site-years from the complete set of site-years in order to train a more accurate composite model.

Prior to testing planting rate optimization for untested sites, a critical next step is to empirically validate optimized designs based on local testing. Moderate model $R^2$ values for individual site-years suggest that the data collected are capturing the relationships between yield, planting rate, topography, and soil at a reasonable level, but in order to determine whether the proposed approach offers an advantage over fixed rate sowing, empirical testing of
the optimized designs is needed. The NYCSGA is currently conducting on-farm validation trials in which the optimized designs developed from previous years’ randomized block variable rate trials are sown.

ACKNOWLEDGMENTS
This project was designed and administered by the New York Corn and Soybean Growers Association. Research funding was provided by the New York Corn and Soybean Growers Association and the New York Farm Viability Institute. Additionally, this work was supported by Hatch funding from the USDA National Institute of Food and Agriculture. We are also thankful to the New York Soybean Checkoff, DuPont Pioneer, and the National Science Foundation Graduate Research Fellowship (Grant No. DGE-1650441) for supporting the graduate studies of Margaret R. Krause. The authors express appreciation to all New York State corn and soybean growers who contributed their time, land, and resources for the on-farm field experiments. This manuscript was previously posted to a preprint server (Krause et al., 2020).

ORCID
Margaret R. Krause https://orcid.org/0000-0003-1154-337X
Michael A. Gore https://orcid.org/0000-0001-6896-8024

REFERENCES
Agudamu, T. Y., & Shiraiwa, T. (2016). Branch development responses to planting density and yield stability in soybean cultivars. Plant Production Science, 19, 331–339. https://doi.org/10.1080/1343943X.2016.1157443
Archontoulis, S. V., & Miguez, F. E. (2014). Nonlinear regression models and applications in agricultural research. Agronomy Journal, 107, 786–798. https://doi.org/10.2134/agronj2012.0506
Barnhisel, R. I., Bitzer, M. J., Grover, J. H., & Shearer, S. A. (1996). Agronomic benefits of varying corn seed populations: A Central Kentucky study. In P. C. Robert, R. H. Rust, & W. E. Larson (Eds.), Proceedings of the Third International Conference on Precision Agriculture, 23-26 June, Minneapolis, Minnesota (pp. 957–965). Madison, WI: ASA, CSSA, and SSSA.
Bishop, T. F. A., Minasny, B., & McBratney, A. B. (2006). Uncertainty analysis for soil-terrain models. International Journal of Geographical Information Science, 20, 117–134. https://doi.org/10.1080/13658810500287073
Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32. https://doi.org/10.1023/A:1010935404324
Brubaker, S. C., Jones, A. J., Lewis, D. T., & Frank, K. (1993). Soil properties associated with landscape position. Soil Science Society of America Journal, 57, 235–239. https://doi.org/10.2136/sssaj1993.03615995005000280703
Bullock, D. G., Bullock, D. S., Nafziger, E. D., Doerge, T. A., Paskiewicz, S. T., Carter, P. R., & Peterson, T. A. (1998). Does variable rate seeding of corn pay? Agronomy Journal, 90, 830–836. https://doi.org/10.2134/agronj1998.000219620090000 60019x
Bullock, D. S., & Lowenberg-DeBoer, J. (2007). Using spatial analysis to study the values of variable rate technology and information. Journal of Agricultural Economics, 58, 517–535. https://doi.org/10.1111/j.1477-9552.2007.00116.x
Changere, A., & Lal, R. (1997). Slope position and erosional effects on soil properties and corn production on a Micaceous soil in central Ohio. Journal of Sustainable Agriculture, 11, 5–21. https://doi.org/10.1300/J064v11n01_03
Chingaryan, A., Sukkarieh, S., & Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture. Computers and Electronics in Agriculture, 151, 61–69. https://doi.org/10.1016/j.compag.2018.05.012
Corassa, G. M., Amado, T. J. C., Strieder, M. L., Schwalbert, R., Pires, J. L. F., Carter, P. R., & Ciampitti, I. A. (2018). Optimum soybean seeding rates by yield environment in Southern Brazil. Agronomy Journal, 110, 2430–2438. https://doi.org/10.2134/agronj2018.04.0239
Cox, M. S., Gerard, P. D., Wardlaw, M. C., & Abshire, M. J. (2003). Variability of selected soil properties and their relationships with soybean yield. Soil Science Society of America Journal, 67, 1296–1302. https://doi.org/10.2136/sssaj2003.1296
Everingham, Y., Sexton, J., Skocaj, D., & Inman-Bamber, G. (2016). Accurate prediction of sugarcane yield using a random forest algorithm. Agronomy for Sustainable Development, 36. https://doi.org/10.1007/s13593-016-0364-z
Fine, A. K., Ristow, A., Schindelbeck, R., & van Es, H. (2016). Comparing soil health test results from Northeast, Midwest, and Mid-Atlantic regions. What's Cropping Up?, 26, 109–110. Retrieved from https://blogs.cornell.edu/whatscroppingup/2016/12/05/ comparing-soil-health-test-results-from-northeast-midwest-and-mid-atlantic-regions/
Fox, R. H., & Piekielek, W. P. (1984). Soil magnesium level, corn (Zea mays L.) yield, and magnesium uptake. Communications in Soil Science and Plant Analysis, 15, 109–123. https://doi.org/10.1080/00103628409367459
Foy, C. D., & Barber, S. A. (1958). Magnesium deficiency and corn yield on two acid Indiana soils. Soil Science Society of America Journal, 22, 145–148. https://doi.org/10.2136/sssaj1958.03615995002200020014x
Frogbrook, Z. L., Oliver, M. A., Salahi, M., & Ellis, R. H. (2002). Exploring the spatial relations between cereal yields and soil chemical properties and the implications for sampling. Soil Use Management, 18, 1–9. https://doi.org/10.1079/SUM2002086
Gallant, J. C., & Wilson, J. P. (2000). Primary topographic attributes. In J. P. Wilson & J. C. Gallant (Eds.), Terrain Analysis: Principles and Applications (pp. 51–85). New York: Wiley.
Häring, T., Dietz, E., Osenstetter, S., Koschitzki, T., & Schröder, B. (2012). Spatial disaggregation of complex soil map units: A decision-tree based approach in Bavarian forest soils. Geoderma, 185–186, 37–47. https://doi.org/10.1016/j.geoderma.2012.04.001
Hedley, C. B., & Yule, I. J. (2009). Soil water status mapping and two variable-rate irrigation scenarios. Precision Agriculture, 10, 342–355. https://doi.org/10.1007/s11119-009-9119-z
Henderson, B. L., Bui, E. N., Moran, C. J., & Simon, D. A. P. (2005). Australia-wide predictions of soil properties using decision trees. Geoderma, 124, 383–398. https://doi.org/10.1016/j.geoderma.2004.06.007
Heslot, N., Jannink, J. L., & Sorrells, M. E. (2013). Using genomic prediction to characterize environments and optimize predic-
tion accuracy in applied breeding data. Crop Science, 53, 921–933. https://doi.org/10.2135/cropsci2012.07.0420
Hoffman, A. L., Kemanian, A. R., & Forest, C. E. (2017). Analysis of climate signals in the crop yield record of sub-Saharan Africa. Global Change Biology, 24, 143–157. https://doi.org/10.1111/gcb.13901
Hörbe, T. A. N., Amado, T. J. C., Ferreira, A. O., & Alba, P. J. (2013). Optimization of corn plant population according to management zones in Southern Brazil. Precision Agriculture, 14, 450–465. https://doi.org/10.1007/s11119-013-9308-7
Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. Journal of Computational & Graphical Statistics, 15, 651–674. https://doi.org/10.1198/106186006X133933
Hurst, M. D., Mudd, S. M., Walcott, R., Attal, M., & Yoo, K. (2012). Using hilltop curvature to derive the spatial distribution of erosion rates. Journal of Geophysical Research, Earth Surface, 117.
Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., … Kim, S. H. (2016). Random forests for global and regional crop yield predictions. Plos One, II, e0156571. https://doi.org/10.1371/journal.pone.0156571
Katerji, N., vann Hoorn, J. W., Hamdy, A., Karam, F., & Mastrorilli, M. (1995). Effect of salinity on water stress, growth, and yield of maize and sunflower. Agriculture Water Management, 30, 237–249. https://doi.org/10.1016/0378-3774(95)01228-1
Key, J. L., & Kurtz, L. T. (1960). Response of corn and soybeans to magnesium fertilizers. Agronomy Journal, 52, 300. https://doi.org/10.2134/agronj1960.00021962005200050025x
Krause, M. R., Crossman, S., DuMond, T., Lott, R., Swede, J., Arliss, S., … Gore, M. A. (2020). Random forest regression for optimizing variable planting rates for corn and soybean using high-resolution topographic and soil data. bioRxiv, 952556. https://doi.org/10.1101.2020.02.17.952556
Kravchenko, A. N., & Bullock, D. G. (2000). Correlation of corn and soybean grain yield with topography and soil properties. Agronomy Journal, 92, 75–83. https://doi.org/10.2134/agronj2000.92175x
Lambert, D., Lowenberg-DeBoer, J., & Malzer, G. L. (2006). Economic analysis of spatial-temporal patterns in corn and soybean response to nitrogen and phosphorus. Agronomy Journal, 98, 43–54. https://doi.org/10.2134/agronj2005.0005
Lashermes, B., Foufoula-Georgiou, E., & Dietrich, W. E. (2007). Channel network extraction from high resolution topography using wavelets. Geophysical Research Letters, 34, L23S04. https://doi.org/10.1029/2007GL031140
Licht, M. A., Lenssen, A. W., & Elmore, R. W. (2017). Corn (Zea mays L.) seeding rate optimization in Iowa, USA. Precision Agriculture, 18, 452–469. https://doi.org/10.1007/s11119-016-9464-7
Ließ, M., Glaser, B., & Huwe, B. (2012). Uncertainty in the spatial prediction of soil texture: Comparison of regression tree and Random Forest models. Geoderma, 170, 70–79. https://doi.org/10.1016/j.geoderma.2011.10.010
Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. Sensors, 18, 2674. https://doi.org/10.3390/s18082674
Liu, Y., Swinton, S. M., & Miller, N. R. (2006). Is site-specific yield response consistent over time? Does it pay? American Journal of Agricultural Economics, 88, 471–483. https://doi.org/10.1111/j.1467-8276.2006.00872.x
Lowenberg-DeBoer, J. (1999). Economics of variable rate planting for corn. In P. C. Robert, R. H. Rust, & W. E. Larson (Eds.), Proceedings of the Fourth International Conference on Precision Agriculture, 19-22 July 1998, St. Paul, Minnesota (pp. 1643–1651). Madison, WI: ASA, CSSA, and SSSA.
Magnier, A., Kalaitzandonakes, N. G., & Miller, D. J. (2010). Product life cycles and innovation in the US seed corn industry. In Agricultural and Applied Economics Association Annual Meeting, July 25–27, 2010, Denver, Colorado (pp. 1–30).
McVickar, M. H., & Shear, G. M. (1946). Variations in response of different varieties and hybrids of field corn to planting rate. Agronomy Journal, 38, 933. https://doi.org/10.2134/agronj1946.00021962003800100009x
Melkonian, J., van Es, H. M., DeGaetano, A. T., & Joseph, L. (2008) ADAPT-N: Adaptive nitrogen management for maize using high resolution climate data and model simulations. In R. Khosla (Ed.), Proceedings Ninth International Conference on Precision Agriculture, 18-21 July, Denver, Colorado. Monticello, IL: International Society Precision Agriculture.
Miao, Y., Mulla, D. J., & Robert, P. C. (2006). Identifying important factors influencing corn yield and grain quality variability using artificial neural networks. Precision Agriculture, 7, 117–135. https://doi.org/10.1007/s11119-006-9004-y
Miller, M. P., Singer, M. J., & Nielsen, D. R. (1988). Spatial variability of wheat yield and soil properties on complex hills. Soil Science Society of America Journal, 52, 1133–1141. https://doi.org/10.2136/ssaaj1988.03615995005200040404x
Modarres, A. M., Hamilton, R. I., Dijak, M., Dwyer, L. M., Stewart, D. W., Mather, D. E., & Smith, D. L. (1998). Plant population density effects on maize inbred lines grown in short-season environments. Crop Science, 38, 104–108. https://doi.org/10.2135/cropsci1998.0011183X003800010018x
Moore, G. A., & Tyndale-Biscoe, J. P. (1999). Estimation of the importance of spatially variable nitrogen application and soil moisture holding capacity to wheat production. Precision Agriculture, 1, 27–38. https://doi.org/10.1023/A:100973802295
Moore, I. D., Gessler, P. E., Nielsen, G. A., & Peterson, G. A. (1993). Soil attribute prediction using terrain analysis. Soil Science Society of America Journal, 57, 443–452. https://doi.org/10.2136/sssaj1993.036159950057000200026x
Mulla, D. J., & Schepers, J. S. (2006). Identifying important factors influencing corn yield and grain quality variability using artificial neural networks. Precision Agriculture, 7, 117–135. https://doi.org/10.1007/s11119-006-9004-y
National Climatic Data Center. (2020). Climate data online. Washington, DC: National Climatic Data Center. http://www.ncdc.noaa.gov/cdo-web/ (accessed 9 June 2020).
Paz, J. O., Batchelor, W. D., & Tylka, G. L. (2001). Method to use crop growth models to estimate potential return for variable-rate management in soybeans. Transations of the ASABE, 44, 1335–1341.
Plant, R. E. (2001). Site-specific management: The application of information technology to crop production. Computers and Electronics in Agriculture, 30, 9–29. https://doi.org/10.1016/S0168-1699(00)00152-6
