Core ideas

- Twenty-six maize fields in two contrasting physiographic areas were studied
- Apparent electrical conductivity sensor estimated subsoil texture within a region
- Optical sensor data showed poor correlation to measured soil properties
- Random forest with sensor data can estimate maize yield in the Coastal Plain
- On-the-go sensors and machine learning can reveal maize yield limiting factors

Soil Sensing and Machine Learning Reveal Factors Affecting Maize Yield in the Mid-Atlantic USA

Rintaro Kinoshita[a,b,*], Masayuki Tani[b], Sonam Sherpa[a,c], Afshin Ghahramani[d], Harold van Es[a]

[a] School of Integrative Plant Science, Soil and Crop Sciences Section, Cornell University, Ithaca, NY 14853-1901, USA

[b] Research Center for Global Agromedicine, Obihiro University of Agriculture and Veterinary Medicine, Obihiro, Hokkaido 080-8555, Japan

[c] International Maize and Wheat Improvement Center (CIMMYT), Patna, Bihar 800025, India

[d] Centre for Sustainable Agricultural Systems, University of Southern Queensland, Toowoomba, QLD 4300, Australia

* Corresponding author (rintaro@obihiro.ac.jp; +81-155-49-5497)

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/agj2.21223.

ABSTRACT

In large-scale arable cropping systems, understanding within-field yield variations and yield-limiting factors are crucial for optimizing resource investments and financial returns, while avoiding adverse environmental effects. Sensing technologies can collect various crop and soil information, but there is a need to assess whether they reveal within-field yield constraints. Spatial data regarding grain yields, proximal soil sensing data, and topographical and soil properties were collected from 26 maize (*Zea mays* L.) growing fields in the Mid-Atlantic USA. Apparent soil electrical conductivity (ECa) collected by an on-the-go sensor (Veris) was an effective method for estimating subsoil textural variation and water holding capacity in the Coastal Plain region, which was also the best predictor of spatial yield pattern when combined with surface pH and topographic wetness index in a Random Forest (RF) model. In the Piedmont Plateau region, proximal soil sensors showed a lower correlation to measured soil properties, while topographical properties (aspect and slope) were important estimators of spatial yield patterns in an RF model. In locations where the RF model failed to predict yield variation, soil compaction appeared to be limiting crop yields. In conclusion, the application of RF models using ECa sensors and topographical properties was effective in revealing within-field yield constraints, especially in the Coastal Plain region. On the Piedmont Plateau, the calibration of proximal sensor information needs to be improved with a particular focus on soil compaction.

INTRODUCTION

Sustainable intensification of agricultural production is crucial for meeting growing food demands without further degrading Earth's environmental systems (Mueller et al., 2012). Understanding yield potential and yield-limiting factors facilitate optimal resource investments while avoiding adverse environmental effects (Oliver et al., 2010; van Ittersum & Rabbinge, 1997). At the large field-scale,

spatial variation of crop yield within a field is a challenging factor for achieving optimal management (Godfray et al., 2010; Oliver et al., 2010).

Recent technological developments in precision agriculture have enabled farmers to identify the spatial variation of crop yields using grain yield monitor information. They can also adopt variable rate management including fertilizer (Ma et al., 2014), crop protection, and variety selection (Katsvairo et al., 2003). However, the sole use of grain yield information is insufficient for formulating management plans since it does not explicitly identify yield constraining factors including soil properties (Jiang & Thelen, 2004; Miller et al., 1988; Shahandeh et al., 2005), topography (Basso et al., 2009; Jiang & Thelen, 2004), genotype (Yang et al., 2009), fertilizer management (Katsvairo et al., 2003), and their interactions. Variable rate management may therefore not be beneficial if the yield constraints are not effectively addressed (Baveye & Laba, 2015). Many significant yield limiting factors (i.e., water, nutrients, etc.) are dependent on soil quality, and the identification of ameliorable or manageable edaphic yield limiting factors is important for more sustainable crop production.

Yield estimation and identification of yield limiting factors may be performed using multiple linear regression (Kitchen et al., 2003), machine learning (Kaul et al., 2005; Kitchen et al., 2003), and biophysical models (Grassini et al., 2015; Hochman et al., 2012; Oliver et al., 2010) although applications of these for within-field assessment have been limited by the data available to parameterize the models. Existing soil survey information often provides inadequate information for within-field scale management (USDA-NRCS, 2015). In grain production systems of the US, the applications of remote and proximal soil sensing techniques (Kuang et al., 2012) have provided information that could link spatio-temporal variation of crop performance to contributing yield limiting factors (Basso et al., 2009; Jiang & Thelen, 2004). Soil proximal sensing equipment such as the Veris Mobile Sensor Platform (MSP-3; Veris Technologies, Salina, KS) combine apparent electrical

conductivity (ECa; Corwin and Lesch, 2003), optical sensing (Kweon & Maxton, 2013), pH metering (Adamchuk et al., 1999) and high-precision GPS to undertake continuous on-the-go measurement of within-field soil and topographical variations. Correlations of proximal sensor information to soil salinity, water content, soil texture, bulk density, and organic matter have been found depending on the site and timing of measurement (Corwin & Lesch, 2005; Kweon & Maxton, 2013). Yield constraints can be thus inferred from sensing data when such correlations are strong and agronomically sound (Kitchen et al., 2003; Stadler et al., 2015). The majority of past research in identifying within-field yield limiting factors using technologies including grain yield monitor information, sensing equipment, and soil assessment are done in a single field and the applicability of the findings at a larger geographical scale is often unknown (Basso et al., 2009; Miller et al., 1988).

Past reseach has shown that machine learning approaches can perform better compared to linear regression type models to predict within-field yield variation (Kitchen et al., 2003). A Random Forest model (Breiman, 2001) is a decision tree predictor that fundamentally includes regression and classification algorithms and was applied to estimate spatial yield variation. Random Forest has several advantages over other statistical models such as its ability to model high-dimensional non-linear relationships in mixed categorical and continuous predictors (Grimm et al., 2008). The method has been utilized to predict various soil properties at different geographical scales (Grimm et al., 2008; Hengl et al., 2015; Kinoshita, Roupsard, et al., 2016; Matei et al., 2017) as well as crop yields (Everingham et al., 2016).

In the Mid-Atlantic region of the United States, crop yield is known to vary significantly by rainfall (Kaul et al., 2005) as well as soil water holding capacity and land capability class (Bandel and Heger, 1994), while temperature and growing season length have been found to have minor influence on yield (Bandel & Heger, 1994; Crasta & Cox, 1996). Kinoshita et al. (2021) confirmed the importance of rainfall amount and temporal distribution in controlling within-field yield variation for the coastal

low-lying area but not the hilly upland area. In both areas, low-yield areas within a field are hotspots for economic losses (Kinoshita, van Es, et al., 2016) and can cause nutrient losses (Turner et al., 2016) when crop management ignores the spatial variation of crop growth and associated environmental factors. However, there is limited information about within-field variation of yield constraints and whether they can be ameliorated in these regions. Therefore, this study was conducted on 26 arable fields in the Mid-Atlantic region of the USA over three Major Land Resource Areas (MLRAs). The objectives were to (i) assess the spatial variation of within-field surface and sub-surface soil properties across a range of soil types and topography in the region, (ii) assess the relationships between measured soil properties and proximal soil sensor information, (iii) assess the feasibility of modelling within-field yield variation using sensing information, and (iv) determine potential site-specific yield limiting factors using the proximal sensor and measured soil information.

MATERIALS AND METHODS

Site Description

This study was conducted in the Mid-Atlantic region of the USA across the states of Delaware, Maryland, Pennsylvania, Virginia, and West Virginia (Fig. 1). Twenty six fields were selected that were under cropping systems that included maize (*Zea mays* L.), soybean (*Glycine max* L.), wheat (*Triticum aestivum* L.), or barley (*Hordeum vulgare* L.). The fields were located between 75° 33’ 51” and 77° 54’ 49” W, and between 38° 56’ 10” and 39° 50’ 23” N. There are two climate regions, warm temperate climates in the southern part and hot summer continental climates in the northern part (Peel et al., 2007). Three Major Land Resource Areas (MLRA; USDA-NRCS, 2006) and five distinctively different areas of soil characteristics were present (Soil Survey Staff, 2015; Fig. 1). The first MLRA was 153C-Mid-Atlantic Coastal Plain (further Coastal Plain), and the latter two were 148-Northern Piedmont and 130A-Northern Blue Ridge (further Piedmont Plateau). Thirteen fields were selected in

the Coastal Plain of which seven were rainfed (CR1-CR7) and six were irrigated (CI1-CI6). The Coastal Plain has a relatively flat topography with an elevation ranging from 15.6 to 28.0 m for the entire study site (Fig. 2a). The within-field elevation differences ranged from 2.6 to 7.1 m. The soils are mainly associated with Typic Hapludults and predominantly formed on coastal plain deposits (sandy loam) below varying depths (40-to-100 cm depth) of aeolian silt deposits that are very acidic (Simonson, 1982). Thirteen fields were selected in the Piedmont Plateau and all of the fields were rainfed (PR1-PR13).

The Piedmont Plateau has undulating topography with elevations ranging from 98.6 to 283 m (Fig. 2b). The within-field elevation differences ranged from 6.0 to 19.0 m. Four distinct soil characteristic areas are present in the Piedmont Plateau (Fig.1). Hagerstown Limestone Valley is mainly associated with Typic Hapludalfs and the soil pH is neutral to slightly acid and has low rock content; Middletown Valley has associated with Typic Hapludults on moderate to strongly acid soil and has high rock contents; Western Piedmont is associated with Ultic Hapludalfs formed on reddish parent materials of Triassic period and is moderately acidic and extremely rocky, and Piedmont Crystalline Rocks is associated with Typic Hapludults and has thin to intermediate depth (< 75 cm) aeolian silt deposits on the surface (Simonson, 1982; Weaver, 1967) and is moderately acidic and rocky.

Crop Yield Information

Yield monitors on combine harvesters with onboard GPS were used to collect spatially-referenced yield data between 2001 and 2014. Post-processing of the data was done with the Yield Editor 2.0.7 software (Sudduth and Drummond, 2007) to correct for flow delays and slow combine velocity at the beginnings and ends of field passes. The data were then interpolated to a 6x6 m raster using inverse distance weighting within the QGIS environment (QGIS Development Team, 2015). The raster size

was chosen based on the typical combine harvester header width. Also, 18 m of field borders were removed from the analyses where we observed unusually low grain yields due to factors including tree shading, soil compaction in headlands, yield monitor errors from slow combine harvester velocity, and high incidence of pest damage.

Nine fields were selected with three to seven years of available georeferenced yield records and standardized principal component analysis (stdPCA; Eastman and Filk, 1993) was applied using the 'princomp" function of the R statistical computing environment (R Core Team, 2014) as explained in Kinoshita et al. (2021). In short, each year of yield data was standardized to have equal variance followed by transformation into principal component (PC) space. The derived PC scores is the transformed value in the PC space, which can be mapped again using the associated georeferencing information. A score map of the first PC reveals the most temporally consistent yield pattern while the remaining PCs represent successively less important latent yield patterns (Kinoshita et al., 2021).

Proximal Soil Sensor Information

Prior to any soil sampling, proximal sensor information was collected from all sites using a Veris MSP-3 unit (Veris Technologies, Salina, KS). It was equipped with an RTK-GPS (ParaDyme; 2.54 cm horizontal and 5.08 cm vertical accuracy; Ag Leader Technology, Ames, IA), apparent electrical conductivity (ECa) sensors (≈0-45 cm and ≈0-90 cm depth), optic sensors at the surface (dual band, 660 nm and 940 nm), and a pH sensor. All sensors were calibrated before each data collection following the operating instructions (Veris Technologies, 2012) and the measurement was made at 18 m intervals across a field. Recorded ECa, optic sensor, and pH data were interpolated to 6 m by 6 m grids using inverse distance weighting in the QGIS software (Conrad et al., 2015; QGIS Development Team, 2015).

The elevation data were collected by the GPS and then interpolated using a regularized spline function to 6 m by 6 m grids (Mitášová & Hofierka, 1993). Various topographical properties were then calculated such as slope (SLOPE), aspect (ASP), profile curvature (PROF), tangential curvature (TAN), and mean curvature (CURV). We also derived the topographic wetness index (TWI), which generally shows the zones of soil water variation within a landscape (Moore et al., 1991) using the SAGA GIS function (Conrad et al., 2015):

$$TWI = \ln\left(\frac{a}{\tan b}\right) \quad (1)$$

where *a* is the specific (contributing) catchment area (m^2) and *b* is the slope gradient (degrees). Topographic position index (TPI) was also calculated, which is the categorized elevation difference between a point in a landscape and the set size of surrounding neighborhood cells:

$$TPI = z_0 - \bar{z} \quad (2)$$

where z_0 is the elevation at the central point and $\bar{z}$ is the average elevation within a predetermined radius (Gallant and Wilson, 2000). The TPI was determined using two different neighborhood sizes of 50 and 450 m, and used the combination of two TPI categories to classify each cell in 10 different landform classes using the SAGA GIS function (Conrad et al., 2015). These 10 classes were later reclassified into four landform classes: swale, flat, slope, and knoll.

Soil Sampling

Nine fields were selected for deep soil sampling, with at least one field from each of five distinctly different soil characteristic areas (Fig. 1). On the Coastal Plain, four fields were selected for deep soil sampling, and nine locations per field were sampled from the 0-90 cm depth using the JMC Environmentalist's Sub-Soil Probe (JMC; 2.88 cm i.d.; Clements Associates, Newton, IA). On the Piedmont Plateau, five fields were selected, and except for PR3, soil samples were collected from

the 0-60 cm depth using a diamond tipped rotary core (Diamond Core; 9.44 cm i.d.; Lackmond Products, Inc., Marietta, GA) assisted with a gas motor due to high rock contents in the field. Due to the hard soil conditions, six locations per field were selected for sampling. Soil sampling location within a field was selected by subdividing the field into nine or six equal-sized rectangular areas. In each grid, deep ECa (0-90 cm depth; ECdp) values were evaluated and three or two samples each from a low, medium, and high ECdp value were taken while maximizing the distance between the rectangle with the same ECdp level. At each sampling point, three and one subsamples were taken for the JMC and the Diamond Core, respectively from non-traffic inter-row of the previous crop. Each subsample was cut in increments at 0-15, 15-30, 30-45, and 45-60, as well as 60-90 cm (only for the JMC), and composited. In total, 60 sampling locations and 298 samples were collected. The sampling location was recorded using a hand-held GPS unit (eTrex Venture HC; Garmin, Schaffhausen, Switzerland). Soil penetration resistance was measured using a soil compaction meter within 0-45 cm of the soil profile where a deep sample set was taken using a FieldScout SC900 Soil Compaction Meter (Spectrum Technologies, Aurora, IL). At each sampling point, six penetrometer measurements were taken and the data were averaged.

In addition, surface soil samples (0-15 cm depth; shallow sample set) were collected in most of the study fields except at CR5 and CR6. In each field, six soil samples were collected by compositing eight push probe cores (2.06 cm i.d.) adjacent to the optic sensor track. The sampling locations were determined using a combination of collected shallow ECa (0-45 cm depth; ECsh) values and optic sensor information at 940 nm (IR). Each field was divided into six equal-sized grids and two samples were collected from high ECsh and low IR areas; two samples from low ECsh and high IR areas; one sample from high ECsh and high IR area, and one sample from low ECsh and low IR area classified using the Veris Soil Viewer software (Veris Technologies, Salina, KS). In total, 142 soil samples were collected since some soil samples were omitted due to GPS data collection issues.

Soil Analyses

Soil samples were crushed using a metal ring and passed through a 2-mm sieve. Any non-organic materials larger than 2 mm were retained, washed, oven dried at 105 °C, and weighed to determine the mass of coarse fragments. The mass was converted to volume and used to estimate the dry bulk density (ρ_b) for the deep sample set using the known volume of the soil sampling probe. For all soil samples, soil texture was assessed using a rapid method (Kettler et al., 2001). Water contents at -10 kPa, -33 kPa, -100 kPa, and -1500 kPa soil water pressure (θ-10, θ-33, θ-100, and θ-1500, respectively) were assessed gravimetrically using a pressure plate apparatus (pressure plate extractor; Soilmoisture Equipment Corp., Santa Barbara, USA) from saturated soil samples (Topp et al., 1993).

The shallow sample set and the first increment of the deep sample set were subjected to soil organic matter (SOM), soil pH, and soil nutrient analyses by Spectrum Analytic Inc. (Washington Court House, OH). Soil organic matter was analyzed by mass loss on ignition in a muffle furnace at 360°C for two hours (Ball, 1964). Soil pH was measured in 1:1 water slurry. Other soil elements, including P, K, Mg, Ca, Zn, Cu, S, and Al were extracted using Mehlich III extraction and quantified (Soil Survey Staff, 2014).

Exploratory Data Analysis

Due to the non-linear relationships, Spearman's rank correlation coefficients were calculated to assess the relationships among measured soil properties. The correlations of each soil sensor value to measured soil properties were assessed using Pearson correlations after verifying linear associations. For the optic sensor data, both the surface segment (0-15 cm) of the deep sample set and the shallow sample set were used, while the ECa sensor values were assessed against the deep sample set to match the depths of the measurements. The on-the-go pH sensor measured values

were validated against the laboratory values using both the surface (0-15 cm depth) increment of the deep sample set and the shallow sample set.

Random Forest

A Random Forest model was developed using the *randomForest* package in R (Liaw & Wiener, 2002; R Core Team, 2014) to estimate crop yields for fields with available stdPCA scores (n = 9), and then built region-specific models for the Coastal Plain (n = 4) and the Piedmont Plateau (n = 5). For each model, there were two types of model inputs i) topographical properties only, and ii) topographical properties and proximal sensing information. In total, six Random Forest models were built. Before assessing the predictability for each field, there was a need to select the most appropriate model. First, each Random Forest model was trained to the full dataset used to build the model and assessed the variance (%) explained by the model. Subsequently, a 3-fold cross-validation was done and R^2 and Root Mean Square Error (RMSE) for each model were calculated:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(v_{pred.i} - v_{meas.i}\right)^2}{n-1}} \quad (3)$$

where $v_{pred.i}$ is the score value predicted by the model at each 6 m by 6 m cell and $v_{meas.i}$ is the calculated score value using measured yield data (Kinoshita et al., 2021). The model of each field was validated using both full-site cross-validation and independent-site validation. The cross-validation was carried out by removing one field as a validation field, which was iterated until all fields were used for both calibration and validation. Independent-site validation was undertaken using the fields that were not utilized to build the prediction models (n = 11) excluding the irrigated fields. The numbers of years of yield data were not sufficient to calculate stdPCA scores at the independent validation sites, therefore mean yield values from raw yield data were used for independent-site validation. Kinoshita et al. (2021) showed that stdPCA scores for maize on the Coastal Plain

represent the yield patterns under moisture-limited growing conditions, whereas the scores represented yield patterns for all growing seasons for the Piedmont Plateau. Therefore, the measured mean yield values were calculated using only moisture-limited years for the Coastal Plain, whereas all yield records were used for calculating the measured mean yield values for the Piedmont Plateau for the independent-site validation. In order to compare the Random Forest model output (stdPCA scores), and the mean yield values, both values were transformed to standard scores:

$$z_{\pi,xy} = \frac{v_{\pi,xy} - \bar{v}_{\pi}}{s_{\pi}} \quad (4)$$

where $v_{\pi,xy}$ is the value at a specific coordinate xy within a field, and $\bar{v}_{\pi}$ and s_{π} are mean and standard deviation, respectively.

For both cross-validation and independent-site validation, taxonomic distance (d; Davis, 1986) was calculated. This compares the spatial patterns of measured stdPCA scores or mean yield to predicted stdPCA scores. This method fits a trend surface using a moving window of a fixed size that relocates across a field for the two spatial patterns. Then, the differences in the model coefficients of each spatial pattern at each cell are calculated. The window size of 42x42 m (approximately 0.2 ha) was used. This was considered the minimum area for which farmers can effectively undertake site-specific management (M. Twining, personal communication.). Then, a third-degree two-dimensional polynomial function was fit to each window (Van Uffelen et al., 1997):

$$Yld(X,Y) = b_{\pi,0} + b_{\pi,1}(X) + b_{\pi,2}(Y) + b_{\pi,3}(X^2) + b_{\pi,4}(Y^2) + b_{\pi,5}(XY) + b_{\pi,6}(X^3) + b_{\pi,7}(Y^3) + b_{\pi,8}(X^2Y) + b_{\pi,9}(XY^2) + \varepsilon \quad (5)$$

using a least square regression fit. Then, d was determined by:

$$d = \sqrt{\sum_{i=0}^{p-1} \frac{(b_{1,i} - b_{2,i})^2}{p}} \quad (6)$$

where p is the number of polynomials (p = 10), and leading numbers (1 and 2) indicate the two patterns for comparison. The d was determined for every 6 m by 6 m cell and thus can be mapped to assess locations of the agreement of two spatial patterns. A value of d=0 would indicate perfect matching of the RF modelled spatial pattern. There is no statistical assessment for threshold d values (Van Uffelen et al., 1997), but d < 0.3 was used in this study, similar to Gandah et al. (2000). To compare model performance among fields, a mean d value ($\bar{d}$) was calculated for each field.

RESULTS AND DISCUSSION

Spatial Variation of Grain Yields

Maize yields ranged from 3.30 to 15.0 Mg ha^{-1} across all fields (Table S1), while state average maize yields of Delaware and Maryland between 2006 and 2015 ranged from 6.3 Mg ha^{-1} to 11.8 Mg ha^{-1} (mean: 8.76 Mg ha^{-1}; USDA National Agricultural Statistics Service, 2017). Water management regime (rainfed vs. irrigated) and MLRA had substantial impacts on mean yields and CV. Mean yields were expectedly higher for the irrigated fields compared to the rainfed set (10.8 to 13.3 Mg ha^{-1}; Table S1). Irrigation reduced the temporal CV substantially reflecting a decoupling of yield variation from apparently weather-related water stress (Troy et al., 2015). For the rainfed set, the temporal CV of grain yields was substantially higher for the Coastal Plain compared to the Piedmont region and precipitation variation was also higher for the Coastal Plain (mean CV = 54.9 %) than for Piedmont (mean CV = 31.1 %) as well as the difference in soil characteristic areas (Fig. 1).

Soil Characteristics

The measured soil properties strongly reflected the soil forming factors of the two regions especially on soil textural variation. Soil texture for the Coastal Plain was mainly sandy loam, loamy sand, and silt loam textural classes, while the Piedmont Plateau had more samples in the loam class

with higher clay contents (Table 1). In general, field average silt content was higher for the Piedmont compared to the Coastal Plain, but the CV was much higher in the Coastal Plain (Table 1) Due to the effects of eolian silt deposits (Simonson, 1982; Weaver, 1967). For the Coastal Plain, the depth of silt deposition was the thickest in CR5, which agrees with the fact that the loess deposition was thicker on the western side of the peninsula (Foss et al., 1978; Fig. 1). Since loess silt was deposited on the soil surface, the CV of silt content increased at deeper soil depths and is more variable vertically on the Coastal Plain compared to the Piedmont Plateau.

The textural variation had a strong influence on water retention values and therefore plant water availability. Mean Available Water Capacity (AWC) at 0-60 cm depth was lower for the Coastal Plain (0.189 kg kg^{-1}) compared to the Piedmont Plateau fields (0.216 kg kg^{-1}) except for CR5 (0.229 kg kg^{-1}) where silt content was the highest among all fields. Water retention parameters of shallow sample set (0-15 cm depth) for all fields were influenced by soil texture, for example field capacity (θ-10 and θ-33) had the highest correlation with sand content (ρ = 0.83; ρ = 0.86; Table S2) whereas θ-100 and permanent wilting point (θ-1500) had the highest correlation with clay content (ρ = 0.89; ρ = 0.92). Soil organic matter had some influence on the water retention parameters, and the highest correlation was found with θ-100 (ρ = 0.64). For soil chemical properties, surface SOM was higher on the Piedmont Plateau (mean = 1.54 %; Table 1) compared to the Coastal Plain (mean = 1.11 %). The cation exchange capacity (CEC) reflected the SOM results (ρ = 0.66; Table S2) more than the clay content (ρ = 0.45) and the majority of low CEC (< 5 $cmol_c$ kg^{-1}) were found on the Coastal Plain and higher CEC (> 10 $cmol_c$ kg^{-1}) were found on the Piedmont Plateau. Surface soil pH in the shallow sample set showed similar ranges in both of the areas and there were some samples with values lower than 6.0 (Table 1). Phosphorus concentration was variable depending on the field. Very high P concentrations were mainly found on the Coastal Plain but also in PR13. For base cations, K^+

concentrations were relatively similar in both regions, while Mg^{2+} and Ca^{2+}were much higher on the Piedmont Plateau (Table 1).

Factors Estimated by the Proximal Sensors

In this study, ECa sensor data showed stronger correlations to measured soil properties than optic sensor data across the two regions. ECa sensor values in the deep soil sample set (n = 66) reflected the variation in soil particle size and water retention parameters. For both regions combined, ECsh showed the highest correlation to clay content at 0-45 cm (*r* = 0.77; Table 2) and ECdp with θ-1500 at 30-45 cm (*r* = 0.75). For each region, ECsh had the highest correlation with clay content at 0-45 cm (*r* = 0.78) and clay content at 30-45 cm (*r* = 0.67) on the Coastal Plain and Piedmont Plateau, respectively. ECdp showed the highest correlation with clay content at 30-45 cm (*r* = 0.76) for the Coastal Plain which confirmed a better correlation with deeper soil properties. In the Piedmont Plateau region, ECdp had the highest correlation with θ−1500 at 30-45 cm (*r* = 0.65). The ratio of ECsh to ECdp (ECR) had the highest correlation with θ−100 at 60-90 cm (*r* = 0.77) for the Coastal Plain, but the correlation in the Piedmont Plateau was overall low. These findings confirmed past research that found a strong correlation between ECa sensor values to clay (*r* = 0.50; Johnson et al., 2001) and silt (*r* = -0.74; Jung et al., 2005). In addition, Mueller et al. (2003) found correlations with elevation (*r* = -0.65) and clay (*r* = 0.63) for ECsh, and Ca (*r* = 0.58) and Mg (*r* = 0.48) for ECdp across four fields with loess silt layers overlaying limestone residuum.

For optic sensor values, correlations with SOM values were not the highest among the measured soil properties, despite the fact that the sensor is marketed for SOM measurement (Table 3). Better correlations were found with soil P, Ca, and Al availability as well as clay and CEC. In both of the regions combined, Red and Ca content (*r* = -0.41; Table 3) and IR and clay content (*r* = 0.30) were most correlated. The ratio of Red to IR (OMR) was not well correlated to measured soil properties

when the two regions are combined. For Red, the highest correlation on the Coastal Plain was with P (r = -0.51), but with CEC on the Piedmont Plateau (r = -0.49). IR values were less correlated to measured values compared to Red and showed the best correlation on the Coastal Plain with Al (r = -0.44) and with Ca on the Piedmont Plateau (r = -0.47). For OMR, a better correlation was found in the Coastal Plain region with P (r = 0.43) but the correlation for the Piedmont Plateau was low. Kweon and Maxton (2013) showed the R^2 for SOM of 0.79 across six fields in central Kansas in Mollisols where the SOM content is higher compared to our study. Also, soil conditions at the time of measurement including soil moisture and temperature could negatively affect correlations to measured soil properties (Kweon et al., 2013; Kweon & Maxton, 2013), making the use of the optic sensors less beneficial under field conditions.

On-the-go pH measurement was feasible for the entire Coastal Plain region, but only for a few fields on the Piedmont Plateau due to high surface rock contents. A linear regression analysis was carried out to predict the laboratory measured soil pH but found low predictability (R^2 = 0.28), possibly because of the difference in equilibration time between the standard 1:1 water pH and the on-the-go pH measurement. Also, the laboratory samples were collected from nearby locations but were not the same. The short-range spatial variation of pH can be high as found in some past studies which showed 20 % variation at 5-m distance (Webster & Butler, 1976).

Random Forest Modeling

Random Forest models were used to predict within-field spatial patterns of stdPCA scores of yield for each region and identify predictor variables among topographical information and proximal sensor information. Important soil factors affecting within-field spatial yield pattern could be inferred from relationships between measured soil properties and sensor data that were identified and explained in the previous section.

Predictability increased when proximal sensor information was combined with topographical information (Table 4), and improvements were higher for the Coastal Plain. Important topographical properties included Aspect followed by TWI and Slope with an R^2 of 0.59 for all regions combined (Table 4). The predictability was higher for the Piedmont Plateau (R^2 = 0.68) compared to the Coastal Plain (R^2 = 0.60) when only topographical information was used. The aspect was then the most important predictor for both regions. Aspect is an important factor for determining incident solar radiation and crop yield (van Ittersum & Rabbinge, 1997), which is confirmed by these data. Gondwe et al. (2019) also found an effect of aspect on soil C content, CEC, and potato yields in hilly upland fields of northern Japan. Combined with sensor information, predictability was better for the Coastal Plain using ECR, pH, and TWI (R^2 = 0.78), compared to the Piedmont Plateau using aspect, slope, and ECR (R^2 = 0.76). For the Coastal Plain, ECR was the most important predictor in the RF model (Table 4) and was correlated to the change in soil texture at the 60-to-90 cm depth (Table 3). Therefore, within-field yield patterns under moisture-limited growing seasons are presumed to be related to subsoil texture change along with surface pH variation and TWI. The ECa sensor is thus beneficial for soil estimation with high spatial resolution as subsoil textural change is costly to assess, and measuring surface soil texture alone was not informative for yield patterns. Subsoil texture and TWI are not changeable through agricultural management but site-specific crop management such as fertilizer and crop variety selection could utilize this information (Chen et al., 2011). On the Piedmont Plateau, the identification of underlying yield constraints appears to be more challenging due to a lower RF model fit compared to the Coastal Plain (Table 4). The Piedmont Plateau includes four areas of distinctly different soil characteristics and topographies, and overall had high variations in measured soil properties including clay, SOM, pH, P and K (Table 1). Therefore, aspect and slope were more important predictors in this area (Table 4) over the proximal sensor information.

In order to validate yield predictability by the RF model on an independent field, full-site cross validation was necessary. This showed higher predictability for the Coastal Plain ($0.07 < R^2 < 0.28$; Table 5) compared to the Piedmont Plateau ($0.01 < R^2 < 0.28$). A $\bar{d}$<0.3 value was used to establish adequate similarity of two spatial patterns, based on Gandah et al. (2000). The overall $\bar{d}$ value was similar for the Coastal Plain (0.249) and the Piedmont Plateau (0.272), which indicated that the RF models perform well for both regions but somewhat better for the Coastal Plain. The lowest $\bar{d}$ value was found in CR3 (0.209; Table 5; Fig. 3c) and the highest in PR1 (Fig. 3f) and PR4 (0.300) among the cross-validated fields. The threshold d value for adequate pattern similarity should depend on the scope of the work (Van Uffelen et al., 1997). In independent-site validation for the Coastal Plain, good predictability was observed for CR6 ($\bar{d}$ = 0.196; Fig. 4c) and CR7 ($\bar{d}$ = 0.214; Table 5), but not in CR5 ($\bar{d}$ = 0.292; Fig. 4f). For the Piedmont Plateau, the $\bar{d}$ value ranged from 0.197 to 0.292 in independent-site validation and the predictability was variable depending on the field. The calculation of the d-value allowed assignment and visual presentation where the RF model successfully predicted the spatial pattern of yield (Fig. 3c, 3f, 4c, and 4f) and this required the availability of at least one year of yield data. This is significantly less challenging compared to collecting multiple years of yield data for a single crop. Locations where d-values were low had a successful RF model estimation of the spatial yield pattern and areas where the d-values were high suggest further assessment. For example, on the Coastal Plain, CR4 was better predicted than CR5 although they shared the same soil series. There are several possibilities for the mismatch of the model predicted values including (i) errors in the measured yield data (Arslan and Colvin, 2002), (ii) errors with the elevation and proximal sensor measurement (Erskine et al., 2007; Kweon et al., 2013), and (iii) yield constraints that were not represented by the predictors (van Ittersum et al., 2013). The potential errors associated with the yield monitor were minimized using the Yield Editor software and the removal of the 18 m field borders. Errors associated with elevation and proximal

sensing data collection were partly addressed by using the mean centered data for the proximal sensor values and calculating both the ECR and OMR on the mean centered data. Nevertheless, in situ soil moisture variations in each field could affect the results (Corwin & Lesch, 2003). The yield constraints including pests and diseases could affect the results and could be significant because of the limited yield data used for independent-site validation. Also, CR5 had thicker loess silt deposition and relatively higher water holding capacity making water less significant as a yield limiting factor for maize. The mean yield of CR5 was the highest among the rainfed fields on the Coastal Plain and also the within-field yield variation was the lowest (CV = 17.9 %; Table S1).

Site-Specific Yield Limiting Factors

The identification of within-field location specific yield constraints is important because they provide information on whether it is ameliorable (Oliver et al., 2010) and the appropriateness and cost (Plant, 2001). Yield data allowed for the determination of the spatial pattern of characteristic yield for all seasons on the Piedmont Plateau and moisture-limited growing seasons on the Coastal Plain. The extraction of spatial patterns of moisture-limited yield is important because the relationship between soil and topographic properties are known to be more significant under extreme weather conditions (Jiang & Thelen, 2004; Kravchenko & Bullock, 2000). Second, the verification of the correlation between each proximal sensor value and measured soil properties infers the underlying yield constraints where the RF models perform adequately.

Judgment sampling could be carried out in the locations where d-values of the RF model were high and the measured yields were low and samples are assessed for a wide array of soil properties that allow for the identification of soil health constraints (Idowu et al., 2008; Sojka et al., 2003). It is also important to confirm that measured low yields are not affected by pests or diseases through farmer knowledge (Calviño et al., 2003; Oliver et al., 2010) or remote sensing information. In this

study, soil constraints were looked at for sites where a deep sample set was available and had high d-values (d > 0.30) and relatively low yields for each field. The penetrometer measurements were useful in many locations especially on the Piedmont Plateau (Fig. 5). In PR3, PR8, and PR11, the high d-value and low yield locations had very shallow penetrable soil layers (PR3 and PR11) or the highest penetration resistance within the field (PR8). In other Piedmont Plateau fields (PR2 and PR6), the selected locations exceeded 2000 kPa (300 psi) in the subsoils known to restrict root growth (Magdoff & van Es, 2021). On the Piedmont Plateau the RF model used to predict yield variation used topographical properties as important predictor variables while the proximal sensor values were less useful (Table 4). The ECR was selected but it was most indicative of bulk density (Table 2). Limitation in root growth by hard layers appears to be the major yield constraints, which could either be ameliorated or managed through site-specific input management (Kinoshita et al., 2021).

For the Coastal Plain, there were no common soil constraints where d-values were high and yields were low. In CR1, surface hardness was the highest among the sampling locations and exceeded 2000 kPa (Fig. 5) and the soil nutrients in 15-30 cm depth was the lowest among the locations indicating the accumulation of the nutrients at the very surface (data not shown). In CR5, the cause of the low yield at the location was not clear but the RF model did not perform well at this site with only one year of yield data to validate the model (Table 5). A separate RF model might be needed for this region with thick loess layers. In CR 6, the high d-value and low yielding location had low penetrometer readings among the sampling locations - although it exceeded 2000 kPa (Fig. 5) - possibly associated with low SOM levels (0.1 %) and very sandy texture. For the Coastal Plain, the RF model used the proximal sensor values more effectively than the Piedmont Plateau (Tables 2, 4, and 5), and therefore the additional benefits of targeted soil sampling appeared less. A question remains, whether water holding capacity can be altered in these fields. The θ-100 water content is known to be the water potential associated with initial crop drought stress, which also was well correlated to

SOM (Table 3). Soil management strategies to increase SOM content may therefore be effective in low yielding areas not just for the topsoils but also for the subsoils (Kinoshita et al., 2017).

CONCLUSIONS

This study built a framework to identify spatio-temporally variable yield constraints using various sensor technologies in the Mid-Atlantic US. The apparent electrical conductivity sensor (ECa) was the most successful in estimating soil properties. The optical sensors showed low correlation to measured soil properties, and their performance for soil organic matter estimation was poor in some regions. The pH sensor was useful but only applicable in rock-free fields. The ratio of shallow to deep ECa sensor values (ECR) had a strong correlation to subsoil texture change related to soil water holding capacity on the Coastal Plain, which is usually very costly and difficult to measure. The ECR value was also the most important in predicting maize yield variation under moisture-limited conditions in the region when used as part of the random forest (RF) model. For the Piedmont Plateau, aspect and slope information from digital elevation models were important yield predictors in the RF model for all weather conditions but the overall predictability was lower than for the Coastal Plain. Overall, the application of the RF model using ECa sensor and topographical properties was effective in revealing within-field yield constraints, especially on the Coastal Plain where there was one soil characteristic area. A question remains whether soil water holding capacity can be improved in this region. For the Piedmont Plateau, calibration of proximal sensor information could be improved with more sites within each soil characteristic area. Soil compaction or shallow topsoil revealed by a penetrometer appeared to be an important yield limiting factor.

CONFLICT OF INTEREST

The authors declare that the research was conducted in the absensce of any commercial or financial relationships that could be constructed as a potential conflict of interest. Mention of trade names or commercial products inthis publication is solely for the purpose of providing specific information.

ACKNOWLEDGEMENT

We are very grateful to Dave Yannacci, Chris Atkinson, David Hertel, and Nelson Oberholzer who helped in field campaignes. David Rossiter, Jeff Melkonian, Bob Schindelbeck, John Dantinne, and Mike Twinning provided us with invaluable suggestions. This work was supported by the funding provided by Willard Agri-Service of Frederick, Inc. We also acknowledge support from the Joint Japan/World Bank Graduate Scholarship Program. We thank reviewers for their useful suggestions.

AUTHOR CONTRIBUTIONS

Rintaro Kinoshita: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Software; Validation; Visualization; Writing-original draft; Writing-review&editing. Masayuki Tani: Formal analysis; Supervision; Validation; Writing-review&editing. Sonam Sherpa: Conceptualization; Formal analysis; Investigation; Methodology; Validation; Writing-original draft. Afshin Ghahramani: Formal analysis; Supervision; Validation; Writing-review&editing. Harold van Es: Conceptualization; Formal analysis; Funding acquisition; Methodology; Project administration; Resources; Supervision; Validation; Writing-review&editing.

REFERENCES

Adamchuk, V. I., Morgan, M. T., & Ess, D. R. (1999). An automated sampling system for measuring soil pH. *Transactions of the ASAE*, *42*(4), 885–891.

Arslan, S., & Colvin, T. S. (2002). Grain yield mapping: Yield sensing, yield reconstruction, and errors. *Precision Agriculture*, *3*(2), 135–154. https://doi.org/10.1023/A:1013819502827

Ball, D. F. (1964). Loss-on-ignition as estimate of organic matter and organic carbon in non-calcareous soils. *Journal of Soil Science*, *15*(1), 84–92.

Bandel, V. A., & Heger, E. A. (1994). *MASCAP Maryland's agronomic soil capability assessment program*. Agronomy Department and Cooperative Extension Service, University of Maryland.

Basso, B., Cammarano, D., Chen, D., Cafiero, G., Amato, M., Bitella, G., Rossi, R., & Basso, F. (2009). Landscape position and precipitation effects on spatial variability of wheat yield and grain protein in southern Italy. *Journal of Agronomy and Crop Science*, *195*(4), 301–312. https://doi.org/10.1111/j.1439-037X.2008.00351.x

Baveye, P. C., & Laba, M. (2015). Moving away from the geostatistical lamppost: Why, where, and how does the spatial heterogeneity of soils matter? *Ecological Modelling*, *298*, 24–38. https://doi.org/10.1016/j.ecolmodel.2014.03.018

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Calviño, P. A., Andrade, F. H., & Sadras, V. O. (2003). Maize yield as affected by water availability, soil depth, and crop management. *Agronomy Journal*, *95*(2), 275–281.

Chen, X. P., Cui, Z. L., Vitousek, P. M., Cassman, K. G., Matson, P. A., Bai, J. S., Meng, Q. F., Hou, P., Yue, S. C., Römheld, V., & Zhang, F. S. (2011). Integrated soil–crop system management for food security. *Proceedings of the National Academy of Sciences*, *108*(16), 6399–6404. https://doi.org/10.1073/pnas.1101419108

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., & Böhner, J. (2015). System for automated geoscientific analyses (SAGA) v. 2.1.4. *Geoscientific Model Development*, *8*, 1991–2007.

Corwin, D. L., & Lesch, S. M. (2003). Application of soil electrical conductivity to precision agriculture: Theory, principles, and guidelines. *Agronomy Journal*, *95*(3), 455–471.

Corwin, D. L., & Lesch, S. M. (2005). Apparent soil electrical conductivity measurements in agriculture. *Computers and Electronics in Agriculture*, *46*(1–3), 11–43. https://doi.org/10.1016/j.compag.2004.10.005

Crasta, O. R., & Cox, W. J. (1996). Temperature and soil water effects on maize growth, development yield, and forage quality. *Crop Science*, *36*(2), 341–348. https://doi.org/10.2135/cropsci1996.0011183X003600020022x

Davis, J. C. (1986). *Statistics and Data Analysis in Geology* (2nd ed.). Wiley.

Eastman, J. R., & Filk, M. (1993). Long sequence time series evaluation using standardized principal components. *Photogrammetric Engineering and Remote Sensing*, *59*(6), 991–996.

Erskine, R. H., Green, T. R., Ramirez, J. A., & MacDonald, L. H. (2007). Digital elevation accuracy and grid cell size: Effects on estimated terrain attributes. *Soil Science Society of America Journal*, *71*(4), 1371–1380. https://doi.org/10.2136/sssaj2005.0142

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*(2), 27. https://doi.org/10.1007/s13593-016-0364-z

Foss, J. E., Fanning, D. S., Miller, F. P., & Wagner, D. P. (1978). Loess deposits of the Eastern Shore of Maryland. *Soil Science Society of America Journal*, *42*(2), 329–334. https://doi.org/10.2136/sssaj1978.03615995004200020026x

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). Wiley.

Gandah, M., Stein, A., Brouwer, J., & Bouma, J. (2000). Dynamics of spatial variability of millet growth and yields at three sites in Niger, West Africa and implications for precision agriculture research. *Agricultural Systems*, *63*(2), 123–140. https://doi.org/10.1016/S0308-521X(99)00076-1

Godfray, H. C. J., Beddington, J. R., Crute, I. R., Haddad, L., Lawrence, D., Muir, J. F., Pretty, J., Robinson, S., Thomas, S. M., & Toulmin, C. (2010). Food security: The challenge of feeding 9 billion people. *Science*, *327*(5967), 812–818. https://doi.org/10.1126/science.1185383

Gondwe, R. L., Kinoshita, R., Aiuchi, D., Suminoe, T., Palta, J., & Tani, M. (2019). Influence of slope direction on soil properties and potato yield potential in hilly upland fields of Hokkaido. *Pedologist*, *63*(2), 61–72. https://doi.org/10.18920/pedologist.63.2_61

Grassini, P., Torrion, J. A., Yang, H. S., Rees, J., Andersen, D., Cassman, K. G., & Specht, J. E. (2015). Soybean yield gaps and water productivity in the western U.S. Corn Belt. *Field Crops Research*, *179*, 150–163. https://doi.org/10.1016/j.fcr.2015.04.015

Grimm, R., Behrens, T., Märker, M., & Elsenbeer, H. (2008). Soil organic carbon concentrations and stocks on Barro Colorado Island—Digital soil mapping using Random Forests analysis. *Geoderma*, *146*(1–2), 102–113. https://doi.org/10.1016/j.geoderma.2008.05.008

Hengl, T., Heuvelink, G. B. M., Kempen, B., Leenaars, J. G. B., Walsh, M. G., Shepherd, K. D., Sila, A., MacMillan, R. A., Mendes de Jesus, J., Tamene, L., & Tondoh, J. E. (2015). Mapping soil properties of Africa at 250 m resolution: Random Forests significantly improve current predictions. *PLoS ONE*, *10*(6). https://doi.org/10.1371/journal.pone.0125814

Hochman, Z., Gobbett, D., Holzworth, D., McClelland, T., van Rees, H., Marinoni, O., Garcia, J. N., & Horan, H. (2012). Quantifying yield gaps in rainfed cropping systems: A case study of wheat in Australia. *Field Crops Research*, *136*, 85–96. https://doi.org/10.1016/j.fcr.2012.07.008

Idowu, O. J., van Es, H. M., Abawi, G. S., Wolfe, D. W., Ball, J. I., Gugino, B. K., Moebius, B. N., Schindelbeck, R. R., & Bilgili, A. V. (2008). Farmer-oriented assessment of soil quality using field, laboratory, and VNIR spectroscopy methods. *Plant and Soil*, *307*(1–2), 243–253.

Jiang, P., & Thelen, K. D. (2004). Effect of soil and topographic properties on crop yield in a north-central corn–soybean cropping system. *Agronomy Journal*, *96*, 252–258. https://doi.org/10.2134/agronj2004.0252

Johnson, C. K., Doran, J. W., Duke, H. R., Wienhold, B. J., Eskridge, K. M., & Shanahan, J. F. (2001). Field-scale electrical conductivity mapping for delineating soil condition. *Soil Science Society of America Journal*, *65*(6), 1829–1837. https://doi.org/10.2136/sssaj2001.1829

Jung, W. K., Kitchen, N. R., Sudduth, K. A., Kremer, R. J., & Motavalli, P. P. (2005). Relationship of apparent soil electrical conductivity to claypan soil properties. *Soil Science Society of America Journal*, *69*(3), 883–892. https://doi.org/10.2136/sssaj2004.0202

Katsvairo, T. W., Cox, W. J., van Es, H. M., & Glos, M. (2003). Spatial yield response of two corn hybrids at two nitrogen levels. *Agronomy Journal*, *95*(4), 1012–1022. https://doi.org/10.2134/agronj2003.1012

Kaul, M., Hill, R. L., & Walthall, C. (2005). Artificial neural networks for corn and soybean yield prediction. *Agricultural Systems*, *85*(1), 1–18. https://doi.org/10.1016/j.agsy.2004.07.009

Kettler, T. A., Doran, J. W., & Gilbert, T. L. (2001). Simplified method for soil particle-size determination to accompany soil-quality analyses. *Soil Science Society of America Journal*, *65*(3), 849–852.

Kinoshita, R., Rossiter, D., & van Es, H. (2021). Spatio-temporal analysis of yield and weather data for defining site-specific crop management zones. *Precision Agriculture*, *22*, 1952–1972. https://doi.org/10.1007/s11119-021-09820-z

Kinoshita, R., Roupsard, O., Chevallier, T., Albrecht, A., Taugourdeau, S., Ahmed, Z., & van Es, H. M. (2016). Large topsoil organic carbon variability is controlled by Andisol properties and effectively assessed by VNIR spectroscopy in a coffee agroforestry system of Costa Rica. *Geoderma*, *262*, 254–265. https://doi.org/10.1016/j.geoderma.2015.08.026

Kinoshita, R., Schindelbeck, R. R., & van Es, H. M. (2017). Quantitative soil profile-scale assessment of the sustainability of long-term maize residue and tillage management. *Soil and Tillage Research*, *174*, 34–44. https://doi.org/10.1016/j.still.2017.05.010

Kinoshita, R., van Es, H., Dantinne, J., & Twining, M. (2016). Within-Field Profitability Analysis Informs Agronomic Management Decisions in the Mid-Atlantic USA. *Agricultural & Environmental Letters*, *1*(1), 160034. https://doi.org/10.2134/ael2016.09.0034

Kitchen, N. R., Drummond, S. T., Lund, E. D., Sudduth, K. A., & Buchleiter, G. W. (2003). Soil electrical conductivity and topography related to yield for three contrasting soil-crop systems. *Agronomy Journal*, *95*(3), 483–495.

Kravchenko, A. N., & Bullock, D. G. (2000). Correlation of corn and soybean grain yield with topography and soil properties. *Agronomy Journal*, *92*(1), 75–83. https://doi.org/10.2134/agronj2000.92175x

Kuang, B., Mahmood, H. S., Quraishi, M. Z., Hoogmoed, W. B., Mouazen, A. M., & van Henten, J. (2012). Sensing soil properties in the laboratory, in situ, and on-line: A review. *Advances in Agronomy*, *114*, 155–223.

Kweon, G., Lund, E., & Maxton, C. (2013). Soil organic matter and cation-exchange capacity sensing with on-the-go electrical conductivity and optical sensors. *Geoderma*, *199*, 80–89. https://doi.org/10.1016/j.geoderma.2012.11.001

Kweon, G., & Maxton, C. (2013). Soil organic matter sensing with an on-the-go optical sensor. *Biosystems Engineering*, *115*(1), 66–81. https://doi.org/10.1016/j.biosystemseng.2013.02.004

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22.

Ma, B.-L., Wu, T.-Y., & Shang, J. (2014). On-farm comparison of variable rates of nitrogen with uniform application to maize on canopy reflectance, soil nitrate, and grain yield. *Journal of Plant Nutrition and Soil Science*, *177*(2), 216–226. https://doi.org/10.1002/jpln.201200338

Magdoff, F. R., & van Es, H. M. (2021). *Building soils for better crops* (4th ed.). SARE Outreach.

Matei, O., Rusu, T., Petrovan, A., & Mihuţ, G. (2017). A data mining system for real time soil moisture prediction. *Procedia Engineering*, *181*, 837–844. https://doi.org/10.1016/j.proeng.2017.02.475

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*(4), 1133–1141. https://doi.org/10.2136/sssaj1988.03615995005200040045x

Mitášová, H., & Hofierka, J. (1993). Interpolation by regularized spline with tension: II. Application to terrain modeling and surface geometry analysis. *Mathematical Geology*, *25*(6), 657–669. https://doi.org/10.1007/BF00893172

Moore, I. D., Grayson, R. B., & Ladson, A. R. (1991). Digital terrain modeling—A review of hydrological, geomorphological, and biological applications. *Hydrological Processes*, *5*(1), 3–30.

Mueller, N. D., Gerber, J. S., Johnston, M., Ray, D. K., Ramankutty, N., & Foley, J. A. (2012). Closing yield gaps through nutrient and water management. *Nature*, *490*(7419), 254–257. https://doi.org/10.1038/nature11420

Mueller, T. G., Hartsock, N. J., Stombaugh, T. S., Shearer, S. A., Cornelius, P. L., & Barnhisel, R. I. (2003). Soil electrical conductivity map variability in limestone soils overlain by loess. *Agronomy Journal*, *95*(3), 496–507.

Oliver, Y. M., Robertson, M. J., & Wong, M. T. F. (2010). Integrating farmer knowledge, precision agriculture tools, and crop simulation modelling to evaluate management options for poor-performing patches in cropping fields. *European Journal of Agronomy*, *32*(1), 40–50. https://doi.org/10.1016/j.eja.2009.05.002

Peel, M. C., Finlayson, B. L., & McMahon, T. A. (2007). Updated world map of the Koppen-Geiger climate classification. *Hydrology and Earth System Sciences*, *11*, 1633–1644.

Plant, R. E. (2001). Site-specific management: The application of information technology to crop production. *Computers and Electronics in Agriculture*, *30*(1–3), 9–29. https://doi.org/10.1016/S0168-1699(00)00152-6

QGIS Development Team. (2015). *QGIS Geographic Information System. Open Source Geospatial Foundation Project* (2.6.1) [Computer software]. http://qgis.osgeo.org

R Core Team. (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. http://www.R-project.org/

Shahandeh, H., Wright, A. L., Hons, F. M., & Lascano, R. J. (2005). Spatial and temporal variation of soil nitrogen parameters related to soil texture and corn yield. *Agronomy Journal*, *97*(3), 772–782. https://doi.org/10.2134/agronj2004.0287

Simonson, R. W. (1982). Loess in soils of Delaware, Maryland, and northeastern Virginia. *Soil Science*, *133*(5), 167–178.

Soil Survey Staff. (2014). *Kellogg Soil Survey Laboratory Methods Manual* (Burt and Soil Survey Staff, Ed.; Version 5.0.). U.S. Department of Agriculture, Natural Resources Conservation Service.

Soil Survey Staff. (2015). *Official Soil Series Descriptions*. https://soilseries.sc.egov.usda.gov/osdquery.aspx

Sojka, R. E., Upchurch, D. R., & Borlaug, N. E. (2003). Quality soil management or soil quality management: Performance versus semantics. *Advances in Agronomy*, *79*, 1–68.

Stadler, A., Rudolph, S., Kupisch, M., Langensiepen, M., van der Kruk, J., & Ewert, F. (2015). Quantifying the effects of soil variability on crop growth using apparent soil electrical conductivity measurements. *European Journal of Agronomy*, *64*, 8–20. https://doi.org/10.1016/j.eja.2014.12.004

Sudduth, K. A., & Drummond, S. T. (2007). Yield Editor: Software for removing errors from crop yield maps. *Agronomy Journal*, *99*(6), 1471–1482. https://doi.org/10.2134/agronj2006.0326

Topp, G. C., Galganov, Y. T., Ball, B. C., & Carter, M. R. (1993). Soil water desorption curves. In M. R. Carter (Ed.), *Soil sampling and methods of analysis*. Canadian Society of Soil Science, Lewis Publishers.

Troy, T. J., Kipgen, C., & Pal, I. (2015). The impact of climate extremes and irrigation on US crop yields. *Environmental Research Letters*, *10*(5), 054013. https://doi.org/10.1088/1748-9326/10/5/054013

Turner, P. A., Griffis, T. J., Mulla, D. J., Baker, J. M., & Venterea, R. T. (2016). A geostatistical approach to identify and mitigate agricultural nitrous oxide emission hotspots. *Science of The Total Environment*, *572*, 442–449. https://doi.org/10.1016/j.scitotenv.2016.08.094

USDA National Agricultural Statistics Service. (2017). *NASS - Quick Stats*. https://data.nal.usda.gov/dataset/nass-quick-stats

USDA-NRCS. (2006). *Land Resource Regions and Major Land Resource Areas of the United States, the Caribbean, and the Pacific Basin*. U.S. Department of Agriculture Handbook.

USDA-NRCS. (2015). *SSURGO Soil Map Coverage versus the U.S. General Soil Map Coverage*. USDA-NRCS.

http://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/geo/?cid=nrcs142p2_053626#data

van Ittersum, M. K., Cassman, K. G., Aggarwal, P. K., Wolf, J., Tittonell, P., & Hochman, Z. (2013). Yield gap analysis with local to global relevance—A review. *Field Crops Research*, *143*, 4–17. https://doi.org/10.1016/j.fcr.2012.09.009

van Ittersum, M. K., & Rabbinge, R. (1997). Concepts in production ecology for analysis and quantification of agricultural input-output combinations. *Field Crops Research*, *52*(3), 197–208. https://doi.org/10.1016/S0378-4290(97)00037-3

Van Uffelen, C. G. R., Verhagen, J., & Bouma, J. (1997). Comparison of simulated crop yield patterns for site-specific management. *Agricultural Systems*, *54*(2), 207–222.

Veris Technologies. (2012). *Operating Instructions MSP3*. Veris Technologies Inc.

Weaver, K. N. (1967). *Generalized Geologic Map of Maryland*. Maryland Geological Survey.

Webster, R., & Butler, B. E. (1976). Soil classification and survey studies at Ginninderra. *Australian Journal of Soil Research*, *14*(1), 1–24. https://doi.org/10.1071/sr9760001

Yang, R.-C., Crossa, J., Cornelius, P. L., & Burgueño, J. (2009). Biplot analysis of genotype × environment interaction: Proceed with caution. *Crop Science*, *49*(5), 1564–1576. https://doi.org/10.2135/cropsci2008.11.0665

FIGURE CAPTIONS

Fig. 1. Map of the study site with field locations and soil regions.

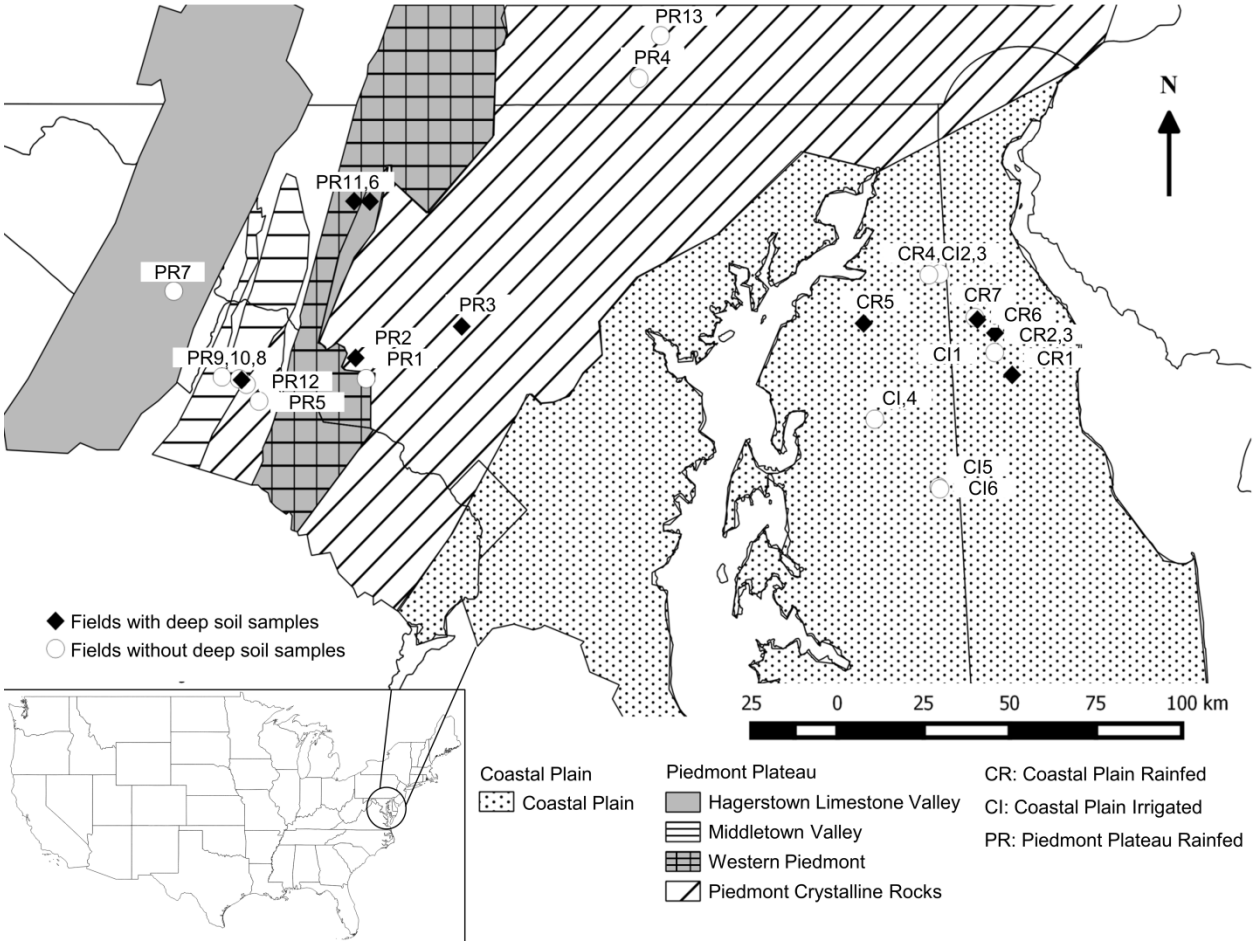


Fig. 2. Typical flat topography of the Coastal Plain (CR1; a) and the undulating topography of the Piedmont Plateau (PR3; b).

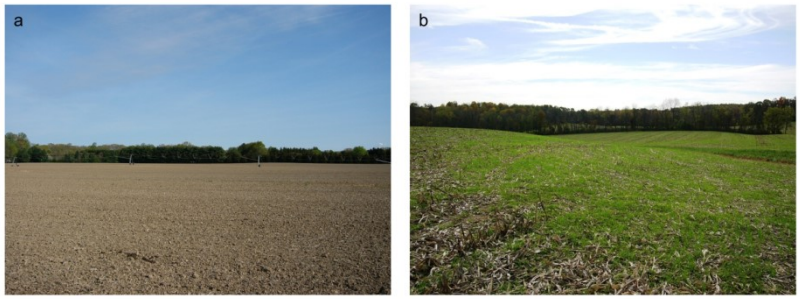


Fig. 3. Maps of a) and d) standardized principal component analysis (stdPCA) derived score values from measured yield data, b and e) stdPCA score values predicted using random forest models, and c and f) taxonomic distance indicating the spatial pattern similarity of the measured and predicted stdPCA scores. The d value of 0.3 was used as a threshold for adequate spatial pattern similarity (Gandah et al., 2000).

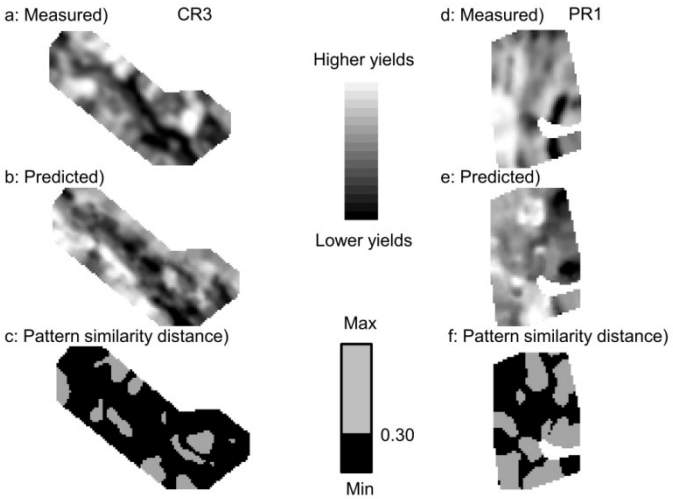


Fig. 4. Maps of a and d) measured yield data, b and e) stdPCA score values predicted using random forest models, and c and f) taxonomic distance indicating the spatial pattern similarity of the measured yield data and predicted stdPCA scores. The d value of 0.3 was used as a threshold for adequate spatial pattern similarity (Gandah et al., 2000).

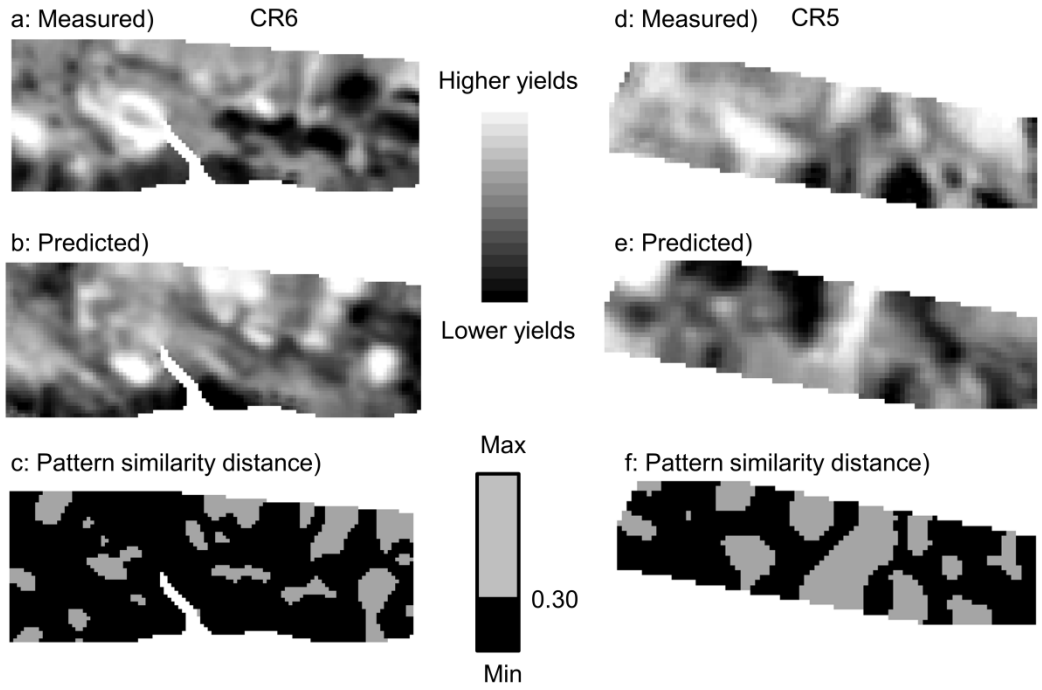
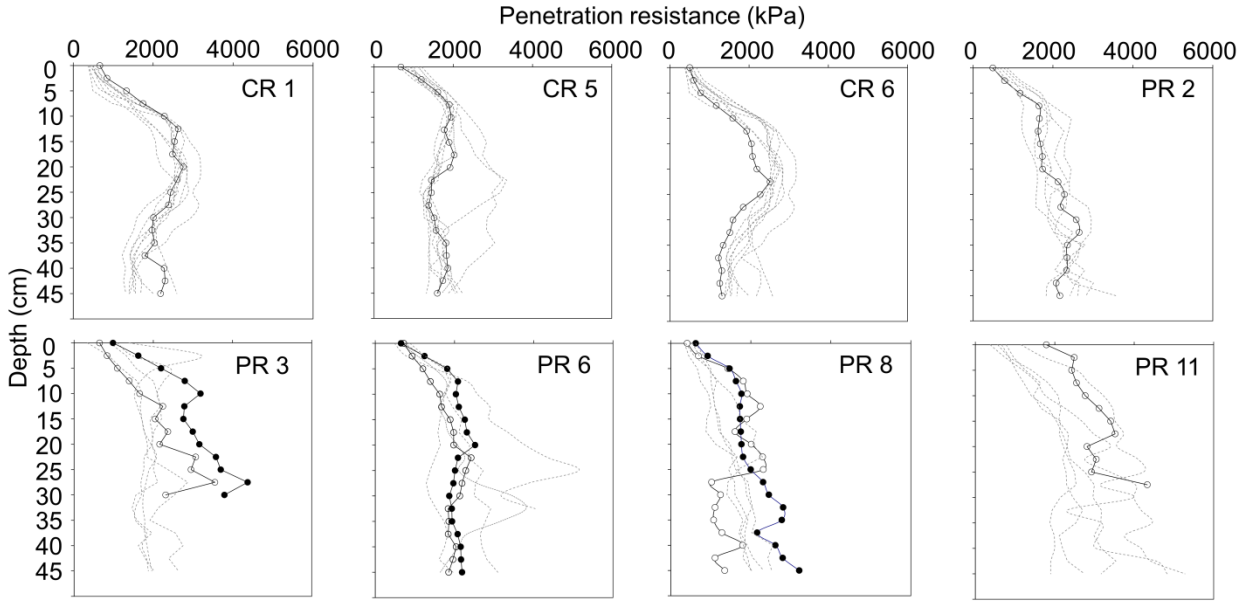


Fig. 5. Penetration resistance measured by a cone penetrometer at 0-45 cm depth for fields with soil profile samples and high d value. The first location is shown in open circle and the second location is shown in closed circle.



TABLES

Table 1. Summary soil statistics of the shallow sample set (0-to-15 cm depth)

	Coastal Plain (n = 64)			Piedmont Plateau (n =78)		
Soil parameter	Mean	Range	CV (%)	Mean	Range	CV (%)
SOM (%)†	1.11	0.300-2.20	37.1	1.54	0.40-3.30	41.9
pH	6.32	5.40-7.10	5.3	6.36	5.10-7.20	7.1
P (mg kg^{-1})	81.1	13.0-226	68.8	59.5	7.00-269	93.9
K (mg kg^{-1})	130	58.0-269	33.3	135	45.0-343	58.1
Mg (mg kg^{-1})	120	67.0-193	25.6	204	77.0-409	35.2
Ca (mg kg^{-1})	685	322-1542	37.8	1148	334-2298	35.9

Al (mg kg^{-1})	806	472-1232	21.7	762	505-1076	15.9
CEC ($cmol_c$ kg^{-1})	4.51	2.10-9.90	37.3	7.69	3.60-14.3	33.9
Clay (%)	7.43	1.27-16.1	40.1	15.8	6.09-27.4	27.7
Silt (%)	46.3	11.9-76.2	36.5	50.9	24.3-69.4	20.8
Sand (%)	46.3	10.8-85.6	41.7	33.3	10.7-64.2	38.4
ρb (Mg m^{-3})	1.29	0.900-1.63	11.2	1.13	0.886-1.52	13.2
θ-10 (kg kg^{-1})	0.243	0.115-0.355	25.5	0.335	0.203-0.475	17
θ-33 (kg kg^{-1})	0.173	0.0680-0.313	33.5	0.25	0.137-0.370	20.8
θ-100 (kg kg^{-1})	0.127	0.0550-0.239	28.9	0.205	0.121-0.315	21.8
θ-1500 (kg kg^{-1})	0.0502	0.0170-0.115	37.1	0.11	0.0580-0.194	23.9
AWC (kg kg^{-1})	0.193	0.0970-0.280	24.2	0.225	0.133-0.309	17.1

† SOM, soil organic matter; CEC, cation exchange capacity; *ρb*, dry bulk density; θ-10, water content at -10 kPa; θ-33, water content at -33 kPa; θ-100, water content at -100 kPa; θ-1500, water content at -1500 kPa; AWC, available water capacity.

Table 2. Pearson correlation coefficients of apparent electrical conductivity values and measured soil samples of the deep sample set.

		ECsh†		ECdp		ECR	
ID	n	*r*	property	*r*	Property	*r*	Property
Coastal Plain	36	0.78	Clay at 0-45cm	0.76	Clay at 30-45 cm	0.77	θ-100 at 60-90cm
		0.78	Clay at 45-60cm	0.76	θ-1500 at 30-45cm	-0.76	Sand at 60-90cm
		0.77	Clay at 15-30cm	0.75	Clay 0-60cm	0.74	Clay at 60-90cm
Piedmont Plateau	30	0.70	Clay at 30-45 cm	0.65	θ-1500 at 30-45 cm	0.33	ρ_b at 30-45 cm
		0.67	Clay at 0-45 cm	0.64	θ-33 at 30-45 cm	0.32	Ca at 0-15 cm
		-0.67	Sand 0-45 cm	0.63	Ca at 0-15 cm	0.32	ρ_b at 0-45 cm
All	66	0.77	Clay at 0-45 cm	0.75	θ-1500 at 30-45 cm	0.31	Silt at 30-45 cm
		0.75	Clay at 30-45 cm	0.71	Clay at 30-45 cm	0.29	Silt at 0-45 cm
		0.74	θ-1500 at 30-45 cm	0.70	θ-33 at 30-45 cm	0.28	Silt at 15-30 cm

† ECsh, shallow apparent electrical conductivity at 0-to-45 cm; ECdp, deep apparent electrical conductivity at 0-to-90 cm; ECR, the ratio of ECsh and ECdp

‡ θ-100: water content at -100 kPa; θ-1500: water content at -1500 kPa; Pen, penetration resistance; θ-33: water content at -33 kPa; θ-10: water content at -10 kPa; Soil moisture, gravimetric soil moisture content; ρ_b, dry bulk density; AWC, available water capacity; SOM, soil organic matter

Table 3. Pearson correlation coefficients of optic sensor values and measured soil properties of the shallow sample set and the surface increment (0-to-15 cm depth) of the deep sample set.

		Red†		IR		OMR	
Area	n	*r*	property	*r*	Property	*r*	Property
Coastal Plain	100	-0.51/-0.47/-0.27	P/Ca/pH	-0.44/0.30/0.30	Al/Sand/SOM	0.43/0.31/-0.28	P/Ca/Al
Piedmont Plateau	108	-0.49/-0.49/0.40	CEC‡/Ca/Clay	-0.47/0.44/-0.40	Ca/Clay/CEC	-0.36/-0.35/-0.33	ρ_b /P/K
All	208	-0.41/-0.35/-0.23	Ca/CEC/SOM	0.30/-0.27/-0.21	Clay/Ca/CEC	-0.22/0.20/0.19	Al/Ca/CEC

† Red, reflectance at 660 nm; IR, infra-red reflectance at 940 nm; OMR, the ratio of Red and IR

‡ CEC, cation exchange capacity; SOM, soil organic matter, ρ_b, dry bulk density

Table 4. Statistical results of the random forest models for the whole region (All), the Coastal Plain, and the Piedmont Plateau with stdPCA scores of crop yield as the response variable, and topographical properties or the combination of topographical properties and proximal sensing information as predictors.

	Topographical properties only					Topographical properties + proximal sensing				
ID	% variance explained	Important variables	%IncMSE†	R^2	RMSE	% variance explained	Important variables	%IncMSE	R^2	RMSE
All	60.4	ASP‡	176	0.59	0.845	75.5	ECR	129	0.74	0.697
		TWI	131				ASP	120		
		Slope	108				OMR	109		
Coastal Plain	61	ASP	172	0.60	0.885	80	ECR	160	0.78	0.68
		TWI	113				pH	110		
		Slope	97.9				TWI	102		
Piedmont Plateau	69.8	ASP	119	0.68	0.67	78.1	ASP	108	0.76	0.59

Slope	102	Slope	96.2
PROF	88.9	ECR	72.2

† %IncMSE, percentage increase in mean square error by dropping one of the important variables; RMSE, root mean square error of prediction

‡ ASP, aspect; TWI, topographic wetness index; PROF, profile curvature; ECR, ratio of shallow and deep apparent electrical conductivity; OMR, ratio of the reflectance at 660 and 940 nm.

Table 5. Validation of random forest model for stdPCA scores for full-site cross validation and comparing predicted stdPCA scores to mean crop yield scores for independent site validation .

	Cross-validation			Independent-validation			
Field ID	R^2	RMSE†	$\bar{d}$	n	R^2	RMSE	$\bar{d}$
CR1	0.17	1.09	0.257	na	na	na	na
CR2	0.07	1.21	0.281	na	na	na	na
CR3	0.28	0.974	0.209	na	na	na	na
CR4	0.15	1.11	0.250	na	na	na	na
CR5	na	na	na	1	0.02	1.32	0.292
CR6	na	na	na	2	0.29	0.957	0.196
CR7	na	na	na	2	0.22	1.03	0.214
PR1	0.02	1.3	0.300	na	na	na	na
PR2	0.03	1.28	0.278	na	na	na	na
PR3	0.09	1.18	0.280	na	na	na	na
PR4	<0.01	1.38	0.300	na	na	na	na
PR5	0.28	0.97	0.200	na	na	na	na
PR6	na	na	na	1	0.17	1.08	0.209
PR7	na	na	na	1	<0.01	1.45	0.292
PR8	na	na	na	2	<0.01	1.43	0.289
PR9	na	na	na	2	0.01	1.35	0.227
PR10	na	na	na	1	0.11	1.16	0.197
PR11	na	na	na	1	0.01	1.34	0.269

PR12	na	na	na	2	0.34	1.28	0.261
PR13	na	na	na	1	0.10	1.17	0.214

† RMSE, root mean square error of prediction; d, mean taxonomic distance; n, number of yield data available