Satellite-based monitoring of crop yields and yield gaps: methods and applications across spatial scales

A dissertation submitted by:
Graham Robert Jeffries

In partial fulfillment of the requirements
for the Degree of
Doctor of Philosophy

Friedman School of Nutrition Science and Policy
Tufts University
Boston, MA

January 2018

Dissertation Committee:
Timothy S. Griffin, PhD (Chair)
David H. Fleisher, PhD
Magaly Koch, PhD
Elena N. Naumova, PhD

This work is dedicated to my wife, Emily.

Table of Contents

Abstract

Background

Meeting the food needs of growing populations consuming more resource-intensive diets will require increasing the magnitude and stability of crop yields. Better data is needed in order to understand the determinants and indicators of crop yields, and to target efforts for yield improvement. Satellite remote sensing imagery has substantial potential for creating low-cost geospatial crop monitoring datasets for research and decision making. This dissertation asks if innovative yield mapping methods with satellite imagery are suitable for three crop monitoring applications: crop yield gap (potential minus actual yield) analysis across farm size groups, yield mapping to inform precision agriculture practices, and in-field characterization of plant traits in advanced crop breeding experiments.

Methods

This project used remote sensing imagery, socio-environmental datasets, and a mixture of biophysical and statistical modeling tools to predict and then analyze crop yields. Objective 1 was to assess variation in Brazilian soybean yields and yield gaps with respect to farm size, using remote sensing imagery to estimate yields and characterize parcel-level features. Objective 2 employed a crop yield prediction algorithm combining remote sensing imagery with crop simulation models to predict mean maize yields at 10 m resolution in Nebraska, USA, and then validated the predictions with harvester yield monitor records. Objective 3 built on the methods and site data in Objective 3 to develop and test a yield prediction algorithm for mapping sub-field maize yields across rainfed and irrigated fields.

Results

Yield gaps in Brazilian soybean production systems were mapped with satellite imagery, revealing significant variations across farms of different size (ranging 20-50,000 ha). Farm size was positively related with soy yield, and inversely related to yield gap size. Objective 2 found that a crop yield prediction model which required no *in situ* data collection was successful in mapping mean maize yields at 10 m resolution (R^2 0.68, RMSE = 0.99 mt ha^{-1}). Objective 3 showed that single-season maize yield prediction accuracy varied with irrigation status, remote sensing imagery source, and algorithm parameters, explaining up to 22.2% of the variation in sub-field yields.

Implications

Crop yield monitoring with satellites and modeling tools has significant potential for applications across scales. Mapping crop yield gaps across Brazil highlights regions and farm types with the highest potential for yield improvement. The methods developed in Objectives 2 and 3 can be applied to lower the cost and increase the availability of sub-field yield maps to support site-specific field management, which can increase farm input use efficiency. The results here suggest that remote sensing will play an important role in identifying pathways leading to higher yielding and more resilient cropping systems.

List of tables and figures

Chapter 2

Chapter 3

Chapter 1: Introduction

Background and Significance

Global food security faces great challenges in the 21st century from growing populations with more resource intensive diets, climate change, and constrained agricultural resources. Food needs are expected to increase 70 to 90% by 2050 due to growing populations and dietary changes related to increasing wealth (Godfray et al., 2010; Godfray and Garnett, 2014). Changes in climate are likely to not only suppress yields globally (Lobell et al., 2011), but also reshape the spatial organization of agriculture (Iizumi and Ramankutty, 2015). Sustainable intensification of agriculture is needed to adapt to social and environmental changes from local to global scales (Godfray and Garnett, 2014).

In response to the significant environmental impact of converting native ecosystems to agriculture, sustainable intensification strategies focus on increasing crop output per unit of area through improved management practices (Garnett et al., 2013). However, the rate of crop yield improvement in recent history is insufficient to meet projected the demands for major crops in 2050 (Ray et al., 2013). Agricultural research must leverage novel data streams for enhancing scientific understanding of the determinants of crop yields across diverse production systems. Identifying best management practices and regions where potential improvement is highest is challenged by the diversity of global growing conditions and paucity of crop yield and management records. Field surveys are expensive and farmer reporting is commonly biased (Burke and Lobell, 2017). Crop monitoring technologies can help to discern the magnitude and causes of underperforming crop yields by combining environmental datasets with modeling tools.

Applications of crop monitoring

Crop monitoring systems can be used to identify strategies for improving crop yields across scales, from farm to region. Three applications of crop monitoring are examined in this dissertation. The first, crop yield gap (potential minus actual yield) analysis, has relevance for targeting investment in regional agricultural infrastructure and technical assistance. The second, mapping sub-field crop yields, supports site-specific management planning at the farmer-level. Finally, the work investigates the suitability of crop monitoring for use with advanced crop breeding research which leverages modern equipment that can programmatically implement agronomic experiments across large areas.

Mapping the location and magnitude of crop yield gaps is valuable for assessing progress toward yield goals and targeting effort for improving yields (Lobell et al., 2009). The process of yield gap mapping begins by estimating potential yield at a given location, which is commonly estimated as either the a) theoretical potential yield from biophysical crop models, or b) attainable yields from the best farmers in a region (van Ittersum et al., 2013). Then, observed yields are subtracted from the potential yield to provide the yield gap. A wide range of methods have been employed to estimate potential and actual yields for yield gap mapping (Burke and Lobell, 2017; Lobell and Azzari, 2017; Sentelhas et al., 2015), but innovations in crop monitoring can improve the resolution and accuracy of yield gap analysis.

Precision agriculture (PA), or site-specific management, combines sensor-based monitoring, data analytics, and variable rate technologies to improve crop productivity by tailoring inputs to match the needs of diverse conditions within a field (Cassman, 1999). PA technologies include yield maps, automated tractor guidance systems (GPS autosteering), and variable rate technologies which change input application rates throughout a field. Many of these technologies have been in development for three or more decades, but adoption in the United States has increased substantially since 2000. Yield mapping was used on more than 30% of US maize (*Zea mays* L.) and soybean (*Glycine max* L.) acres in 2010 and 2012, respectively, while adoption of tractor guidance and variable rate technologies in maize acreage was ~45% and ~22% in 2010 (Schimmelpfennig, 2016). Yield mapping documents the relative performance of crops with respect to their location and management, which can be used by farmers to plan for future seasons. Spatial crop yield maps are valuable for delineating management zones, designing input prescriptions, and tracking crop performance (Zhang et al., 2002). Maps of multi-season mean yields are commonly used to identify areas of persistent high or low yields which may be correlated with localized persistent conditions (e.g. soils and topography) (Schepers et al., 2004). Yield maps are typically created from data collected by harvested yield monitor sensors which record the rate of harvest as the tractor moves through a field. Identifying alternative means of creating yield maps could expand the adoption of yield mapping by making yield maps available for sites and years where yield monitor data was not collected.

High-throughput phenotyping (HTP) is the practice of measuring the characteristics (physiological, biochemical, morphological, etc.) of organisms at rates not feasible with human labor alone (hundreds of plants per day or more) (Fahlgren et al., 2015; Gehan and Kellogg, 2017). The development of sensors for measuring plant features from canopy to DNA has led to a dramatic reduction in the cost and time for characterizing plants, paving the way for larger and more complex experiments. Field-based HTP (FHTP) enable researchers to plant many (tens to potentially thousands) of crop varieties and observe patterns in their growth and development in response to environmental conditions (Araus and Cairns, 2014). Advances in crop breeding through FHTP show potential for increasing efficiency of identifying and breeding for specialized traits such as drought resistance, flavor properties, and nutrient density, among others. Remote sensing products may enable complex FHTP research studies across large areas by providing a uniform source of crop response estimates. In contrast, without remote sensing imagery, crop response data must be collected *in situ* with surveys or harvester yield monitors. Differences in survey procedures, or sensor equipment and calibration, may limit the comparability of records across sites and over time.

Informatics and data science in agriculture

In recent decades, technical advances in remote sensing, modeling, and general computation have improved the quality and availability of data for crop monitoring. Remote sensing (RS) is the practice of observing an object without physical contact, commonly via sensors on satellites, aircraft, or unmanned aerial vehicles. Agricultural applications of RS imagery are founded on the fact that crop canopies reflect, absorb, and transmit electromagnetic radiation from the sun at differing levels depending on canopy properties. For example, leaf chlorophyll levels are correlated with increased reflectance of near-infrared band radiation. Therefore, RS imagery can be used estimate the properties of a crop canopy and relate them to end-of-season yield.

At a global level, satellite imagery has monitored the changing extent and composition of agricultural land (Ramankutty and Foley 1999). At a regional level, RS has been widely used to monitor crop growth, development, and quality throughout the growing season (Bolton and Friedl, 2013; Doraiswamy et al., 2005). At the farm-level, RS imagery has been used to detect pest pressure (Backoulou et al., 2015), estimate canopy properties (Gitelson et al., 2003), and inform management decisions (Hatfield et al., 2008). One prevalent method for estimating yield from RS imagery is using a statistical model relating yield observations from surveys to vegetation indices (VI) from imagery (Delécolle et al., 1992). Recent advances in crop monitoring with RS have combined imagery with other modeling tools, including crop simulation models.

Crop simulation models (CSM) are widely used to examine the relationship between agricultural production factors and food security, for example via the quantification of yield gaps, climate change impacts on agriculture, and the exploration of outcomes from alternative farm management practice (Boote et al., 1996). CSM are simplified, mathematical representations of cultivated plant growth, development, and yield. The models use information about local site conditions (e.g. soil and weather) and management decisions (e.g. crop type and nutrient applications) to simulate changes in crop status over a growing season. Crop models can be used to explore alternate scenarios of crop growth, such as in modeling potential impacts of climate change on farming systems (Asseng et al., 2013), or to provide a tool for investigating the drivers and limits of crop growth in existing fields (Sibley et al., 2014). CSM can be used to estimate crop yields at regional and local scales and may aim to simulate current growing conditions or future scenarios (Resop et al., 2012).

Relative to the established literature base in both crop modeling and RS, research focusing on the integration of crop modeling and RS is young and rapidly evolving. There are multiple methods for incorporating RS and CSM tools, including using imagery-derived products as CSM inputs and calibrating CSM variables or outputs with imagery (Maas, 1988; Moulin et al., 1998). One recent approach combines RS data and CSM outputs to eliminate the need for *in situ* field data for modeling yields. The scalable crop yield mapping (SCYM) algorithm uses CSM simulations to fit a regression model relating in-season crop status to end-of-season yield. The model is then used to predict yields by substituting simulated crop status with proxies computed from RS imagery (Lobell et al., 2015). Recent SCYM research builds on earlier work: combining radar and CSM to estimate sugar beet (*Beta vulgaris* L.) biomass (Bouman, 1991; Bouman, 1992); estimating crop canopy properties from imagery and a sugar beet model (Clevers, 1997); wheat (*Triticum aestivum* L.) yield prediction (Sehgal et al., 2005). Sibley et al. (Sibley et al., 2014) compared three field-level maize yield predictions employing: 1) accumulated photosynthetically active radiation from the Moderate Resolution Imaging Spectroradiometer (MODIS) satellite, 2) CSM parameter calibrations from MODIS and Landsat imagery, and 3) a SCYM-like approach, and demonstrated that the third method performed best. Since it's formalization, a collection of recent work has documented refinements to the SCYM algorithm (Azzari et al., 2017; Burke and Lobell, 2017; Jain et al., 2017; Jin et al., 2017a; Jin et al., 2017b).

Research Objectives

The dissertation is composed of three main objectives, each designed to innovate in crop monitoring methods and applications. The objectives are unified thematically by 1) crop yield monitoring with remote sensing imagery; 2) synthesis of large, multidisciplinary data sources.

The work maps crop yields across a wide range of spatial scales – from Brazil to three fields in Nebraska. The studies focus on two globally important crops, maize and soybean, in order to demonstrate the methods developed herein are not limited to a single crop. The significant differences in growing conditions between tropical and temperate study sites provided opportunity for comparing opportunities and limitations for crop monitoring in both climates.

Objective 1 (Chapter 2)
Characterize the relationship between farm size and soybean yields in Brazil.

- *Sub-objective 1.1*: Design a statistical model for out-of-sample soybean yield prediction with remote sensing imagery.
- *Sub-objective 1.2*: Quantify the spatial and temporal variability of soybean yield ('yield heterogeneity').
- *Sub-objective 1.3*: Develop parcel-scale metrics of farm size – geographic, legal, and economic.
- *Sub-objective 1.4:* Compare soybean yields and yield heterogeneity measures across farm size groups.

Objective 2 (Chapter 3)
Evaluate the accuracy of multi-season maize yield maps created with SCYM algorithm variants for applications in precision agriculture.

- *Sub-objective 2.1*: Link a biophysical model of canopy spectral properties, a radiative transfer model (RTM), with a crop simulation model in order to compare approaches for simulating vegetation index (VI) values.
- *Sub-objective 2.2*: Compare SCYM yield prediction from multiple linear regression and gradient boosted forest models.
- *Sub-objective 2.3*: Characterize bias in yield predictions using Bland-Altman plots and estimate a model for bias-correction.
- *Sub-objective 2.4*: Document the relationship between prediction accuracy and the number of seasons observed.

Objective 3 (Chapter 4)
Assess maize yield prediction accuracy from a scalable crop yield mapping (SCYM) algorithm for suitability in field-based high throughput phenotyping research.

- *Sub-objective 3.1*: Replicate the SCYM algorithm with the DSSAT CERES-Maize crop simulation model
- *Sub-objective 3.2*: Evaluate the sensitivity of a scalable crop yield mapping method to alternative parameterizations (yield vs. biomass)
- *Sub-objective 3.3*: Compare prediction performance across rainfed and irrigated maize systems
- *Sub-objective 3.4*: Document the effect of spatial resolution on yield prediction performance

- *Sub-objective 3.5*: Compare the accuracy of predictions from multiple independent satellite sensors

References

Araus, J.L. and Cairns, J.E., 2014. Field high-throughput phenotyping: the new crop breeding frontier. Trends in Plant Science, 19(1): 52-61.

Asseng, S. et al., 2013. Uncertainty in simulating wheat yields under climate change. Nature Climate Change, 3: 827.

Azzari, G., Jain, M. and Lobell, D.B., 2017. Towards fine resolution global maps of crop yields: Testing multiple methods and satellites in three countries. Remote Sensing of Environment.

Backoulou, G.F., Elliott, N.C., Giles, K.L. and Mirik, M., 2015. Processed multispectral imagery differentiates wheat crop stress caused by greenbug from other causes. Computers and Electronics in Agriculture, 115: 34-39.

Bolton, D.K. and Friedl, M.A., 2013. Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. Agricultural and Forest Meteorology, 173: 74-84.

Boote, K.J., Jones, J.W. and Pickering, N.B., 1996. Potential Uses and Limitations of Crop Models. Agronomy Journal, 88: 704-716.

Bouman, B.A.M., 1991. Linking X-band radar backscattering and optical reflectance with crop growth models, Wageningen Agricultural University, 169 pp.

Bouman, B.A.M., 1992. Linking physical remote sensing models with crop growth simulation models, applied for sugar beet. International Journal of Remote Sensing, 13(14): 2565-2581.

Burke, M. and Lobell, D.B., 2017. Satellite-based assessment of yield variation and its determinants in smallholder African systems. PNAS, 114(9): 2189–2194.

Cassman, K.G., 1999. Ecological intensification of cereal production systems: Yield potential, soil quality, and precision agriculture. Proceedings of the National Academy of Sciences, 96(11): 5952-5959.

Clevers, J.G.P.W., 1997. A simplified approach for yield prediction of sugar beet based on optical remote sensing data. Remote Sensing of Environment, 61(2): 221-228.

Delécolle, R., Maas, S.J., Guérif, M. and Baret, F., 1992. Remote sensing and crop production models: present trends. ISPRS Journal of Photogrammetry and Remote Sensing, 47(2): 145-161.

Doraiswamy, P.C. et al., 2005. Application of MODIS derived parameters for regional crop yield assessment. Remote Sensing of Environment, 97(2): 192-202.

Fahlgren, N., Gehan, M.A. and Baxter, I., 2015. Lights, camera, action: high-throughput plant phenotyping is ready for a close-up. Current Opinion in Plant Biology, 24(Supplement C): 93-99.

Garnett, T. et al., 2013. Sustainable Intensification in Agriculture: Premises and Policies. Science, 341(6141): 33.

Gehan, M.A. and Kellogg, E.A., 2017. High-throughput phenotyping. American Journal of Botany, 104(4): 505-508.

Gitelson, A.A. et al., 2003. Remote estimation of leaf area index and green leaf biomass in maize canopies. Geophysical Research Letters, 30(5): n/a-n/a.

Godfray, H.C.J. et al., 2010. Food Security: The Challenge of Feeding 9 Billion People. Science, 327(5967): 812.

Godfray, H.C.J. and Garnett, T., 2014. Food security and sustainable intensification. Philosophical Transactions of the Royal Society B: Biological Sciences, 369(1639).

Hatfield, J.L., Gitelson, A.A., Schepers, J.S. and Walthall, C.L., 2008. Application of Spectral Remote Sensing for Agronomic Decisions. Agronomy Journal, 100: S-117-S-131.

Iizumi, T. and Ramankutty, N., 2015. How do weather and climate influence cropping area and intensity? Global Food Security, 4(Supplement C): 46-50.

Jain, M. et al., 2017. Using satellite data to identify the causes of and potential solutions for yield gaps in India's Wheat Belt. Environmental Research Letters, 12(9).

Jin, Z., Azzari, G., Burke, M., Aston, S. and Lobell, B.D., 2017a. Mapping Smallholder Yield Heterogeneity at Multiple Scales in Eastern Africa. Remote Sensing, 9(9).

Jin, Z., Azzari, G. and Lobell, D.B., 2017b. Improving the accuracy of satellite-based high-resolution yield estimation: A test of multiple scalable approaches. Agricultural and Forest Meteorology, 247: 207-220.

Lobell, D. and Azzari, G., 2017. Satellite detection of rising maize yield heterogeneity in the U.S. Midwest. Environmental Research Letters, 12.

Lobell, D.B., Cassma, K.G. and Field, C.B., 2009. Crop Yield Gaps: Their Importance, Magnitudes, and Causes. Annual Review of Environment and Resources, 34: 179-204.

Lobell, D.B., Schlenker, W. and Costa-Roberts, J., 2011. Climate Trends and Global Crop Production Since 1980. Science, 333(6042): 616.

Lobell, D.B., Thau, D., Seifert, C., Engle, E. and Little, B., 2015. A scalable satellite-based crop yield mapper. Remote Sensing of Environment, 164: 324-333.

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

Moulin, S., Bondeau, A. and Delecolle, R., 1998. Combining agricultural crop models and satellite observations: From field to regional scales. International Journal of Remote Sensing, 19(6): 1021-1036.

Ray, D.K., Mueller, N.D., West, P.C. and Foley, J.A., 2013. Yield Trends Are Insufficient to Double Global Crop Production by 2050. PLOS ONE, 8(6): e66428.

Resop, J.P., Fleisher, D.H., Wang, Q., Timlin, D.J. and Reddy, V.R., 2012. Combining explanatory crop models with geospatial data for regional analyses of crop yield using field-scale modeling units. Computers and Electronics in Agriculture, 89(Supplement C): 51-61.

Schepers, A.R. et al., 2004. Appropriateness of management zones for characterizing spatial variability of soil properties and irrigated corn yields across years. Agronomy Journal, 96(1): 195-203.

Schimmelpfennig, D., 2016. Farm Profits and Adoption of Precision Agriculture, ERR-217, U.S. Department of Agriculture, Economic Research Service.

Sehgal, V.K., Sastri, C.V.S., Kalra, N. and Dadhwal, V.K., 2005. Farm-Level Yield mapping for precision crop management by linking remote sensing inputs and a crop simulation model. Journal of the Indian Society of Remote Sensing, 33(1): 131-136.

Sentelhas, P.C. et al., 2015. The soybean yield gap in Brazil – magnitude, causes and possible solutions for sustainable production. The Journal of Agricultural Science, 153(8): 1394-1411.

Sibley, A.M., Grassini, P., Thomas, N.E., Cassman, K.G. and Lobell*, D.B., 2014. Testing Remote Sensing Approaches for Assessing Yield Variability among Maize Fields. Agronomy Journal, 106: 24-32.

van Ittersum, M.K. et al., 2013. Yield gap analysis with local to global relevance—A review. Field Crops Research, 143(Supplement C): 4-17.

Zhang, N., Wang, M. and Wang, N., 2002. Precision agriculture—a worldwide overview. Computers and Electronics in Agriculture, 36(2): 113-132.

Chapter 2: Brazilian soybean yield and yield heterogeneity varies with farm size
Prepared for submission to *PNAS*

Authors
Graham R. Jeffries[*]
Avery S. Cohn[†]
Timothy S. Griffin[*]
Arthur Bragança[†, ‡]
David H. Fleisher[§]
Elena N. Naumova[*]
Magaly Koch[¶]

Classification
Social Sciences–Sustainability Science
Biological Sciences–Agricultural Sciences

Short title
Brazilian soybean yields and yield gaps vary with farm size

Author affiliation
[*]Friedman School of Nutrition Science and Policy, Tufts University, Boston, MA 02111; [†]Fletcher School of Law and Diplomacy, Tufts University, Medford, MA 02155; [‡]TRACES, Rio de Janeiro, Brazil; [§]Adaptive Cropping Systems Laboratory, Agricultural Research Service, United States Department of Agriculture, Beltsville, MD 20705; [¶]Center for Remote Sensing, Boston University, Boston, MA 02215

Corresponding author
Graham R. Jeffries
711 Main Street
Rockland, ME 04841
717.649.6907
graham.r.jeffries@gmail.com

Keywords
agriculture, yield gaps, remote sensing, soybeans, Brazil

Abstract

Understanding the relationship between farm size and crop yields and yield gaps may help to improve yields through better targeting investment and programs. A growing literature documents variations in yield gaps, but largely ignores the role of farm size as a factor shaping yield structure (yield magnitude, spatial variation, and stability over time). Research on the inverse farm size-productivity relationship (IR) theory–that small farms are more productive than large ones all else equal–suggests that yield magnitude may vary by farm size. We examined farm size-yield structure relationships for soybeans (*Glycine max* L.) in Brazil for years 2001-2015 by linking remote sensing-based yield estimates with a database of rural property boundaries. Using out-of-sample soybean yield predictions, we estimated 1) spatial heterogeneity in soy yields, and 2) yield stability over time. We found a positive relationship between soy yields and farm size at the national level–a 10% increase in plantable area was associated with a ~1% increase in yields. Soybean yield gaps were inversely related to physical and economic measures of farm size. We identified a relative overall yield heterogeneity of 11.3% for farms with productive area 20-<50 ha, versus ~8% for farms >5,000 ha. The relationship between farm size and yield stability was nonlinear, and mid-sized farms had the most stable yields. The work suggests that farm size is an important factor in understanding yield structure and should be considered when designing efforts to improve soy yields in Brazil.

Significance

Higher crop yields are needed to meet growing global food needs, and identifying where yields are underperforming on some farms relative to others (a 'yield gap') is important for targeting efforts to improve them. Some studies have shown that crop yields can be systematically different among farms of different sizes, even when growing conditions are similar. However, it is unknown whether yield gaps differ across farm size groups. We innovate by combining satellite imagery and a geospatial property boundary dataset, finding that Brazilian soybean yield gaps are higher, and yields lower, among smaller farms.

Background

Understanding the farm size-specific characteristics of crop yields and yield gaps may help to improve yields by enabling better targeting of technical assistance and agricultural development programs (1). A growing literature documents variations in yield gaps (2-4), or the difference between potential yields and actual yields (5), but largely ignores the role of farm size as a factor shaping yield structure (yield magnitude, gaps and heterogeneity, and stability over time). The observation that farm productivity is inversely related to farm size (the *inverse farm size-productivity relationship* theory, IR) – that smaller farms are more productive than large farms all else equal – has been reported, and contested, across many geographic settings (6-8). Systematic differences in crop yields and yield gaps across farm sizes hold implications for the role of farm size in development strategies because increasing the area in small farms through land reform may alter overall productivity. However, the availability of data on crop yields and property characteristics is often limited due to the labor-intensity of manual surveys, which in turn has constrained research on yield structure and farm size relationships. Combining satellite remote sensing imagery, yield surveys, and spatial property boundaries provides a novel lens into the farm size – yield structure relationship across a wide range of farm sizes.

Contemporary work has shown that the IR cannot be fully explained by labor market failures (6, 9), discounting a prominent theory for the IR mechanism (10-12). Whether the IR is a statistical artifact arising from omitted variable bias continues to be disputed–a body of literature shows the IR to be robust to factors such as soil quality (6), but recent work suggests that behavioral mechanisms related to plot geometry may explain away the IR (7). Spatial plot and property boundaries can be used to objectively measure field and farm size, which is preferable to self-reported area estimates which are commonly biased (4, 13, 14). Because most existing studies that employed spatial crop area measurements have collected field boundaries manually with GPS devices, spatial IR research has focused on small (<5 ha) farms and plots. Available data is also commonly limited in spatial and temporal extent, obscuring variation across farming systems, growing regions, and time. No existing research on the relationship between farm size and yield gaps could be identified.

This study innovated by bridging IR and yield gap research with a novel combination of data sources to examine whether Brazilian soybean yield structure varies across a wide range of farm sizes (20 to >10,000 ha, where the minimum size was limited by satellite imagery resolution). The case is relevant because of the global importance of Brazilian soybean production, the wide distribution of farm sizes, recent history of land reform (15), and challenges and opportunities for balancing development with climate resilience (16). Documenting yield heterogeneity and yield stability is relevant to understanding food security risks both regionally, as low yields can reduce farm incomes, and globally as failed crops can contribute to price shocks in soy importing regions (17). We used yield heterogeneity as an indicator of yield gaps, measured as the difference between the 95th percentile of attained yields and mean yields. We used a multi-stage modeling approach which included: (*i*) statistical models to estimate soybean yields with remote sensing (RS) vegetation indices, climate, and environmental variables, (*ii*) a panel of out-of-sample soybean yield predictions linked with geospatial property data, and (*iii*) statistical models relating farm size to variability in soybean yield structure.

Results and Discussion

A direct relationship between farm size and soybean yield was observed

Controlling for yield covariates, we found a direct (i.e. positive, not inverse) relationship between farm size and soybean yields across Brazil (Figure 2-1). Nationally, a 10% increase in plantable area was associated with a ~1% increase in yields. The direction, magnitude, and linearity of the relationship varied across states. We found consistent relationships when comparing several measures of fam size, including physical measures such as total property area and plantable area, and an economic farm size measure which was calculated by adjusting physical area by a measure of county-level expected productivity (FM_{pred}, SI Text: Fiscal module estimates).

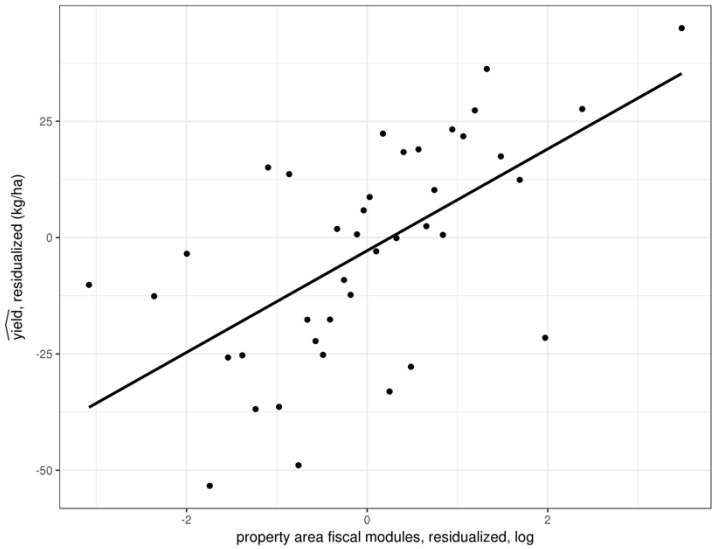
Our analysis of the inverse farm size-productivity relationship built on existing work by 1) introducing satellite remote sensing for yield prediction, 2) assessing variation in yields gaps across farm sizes, 3) plotting non-linear relationships, and 4) studying the IR across space. While a positive relationship was identified at the national level, the direction and strength of the relationship between farm size and yield varied across states. Both of these characteristics were consistent with an analysis of farm size groups and total factor productivity (TFP) in Brazil (8).

In contrast to other work outside of Brazil, the direct relationship was robust to controls for the property perimeter-to-area ratio (Table S5) (7).

The current work adds to the limited pool of scholarship examining the IR across a wide range of farm sizes–most was restricted to farms of a size that could be managed without machinery (6, 7, 14). The analysis was performed using a sample of properties which were predominantly soy over the study period and were >20 ha. Twenty hectares was selected as a cutoff as we 1) wanted to align farm size groups with those reported by the Brazilian government's agricultural statistics (which includes a 20-<50 ha group), and 2) the 500 m MODIS MOD09A1 satellite imagery used in the study provided pixel areas of ~25 ha. We were not able to directly compare analytical results from the study with many other IR papers because the farm sizes examined here do not overlap with those in other studies.

A time-varying fiscal module indicator (SI Text: Fiscal module estimates) successfully captured latent biophysical and social productivity factors in the IR analysis. TFP is preferable to yields especially in regions where total factor returns to scale are not constant, as has been demonstrated in the case of some regions in Brazil (8, 18), but data on factor inputs is limited. A fiscal module indicator may capture aspects of TFP and endogenize them on the right-side of the equation, permitting the use of crop yields as a valid outcome measure.

Figure 2 - 1: Binscatter plot of the effect of changes in farm size (FM_{pred}) on predicted yield.
Note: Points indicate the mean of binned x-axis values. See SI Text: Binscatter plots.
File: 'graphics/ir/yield_hat_fm_parcel_est.png'

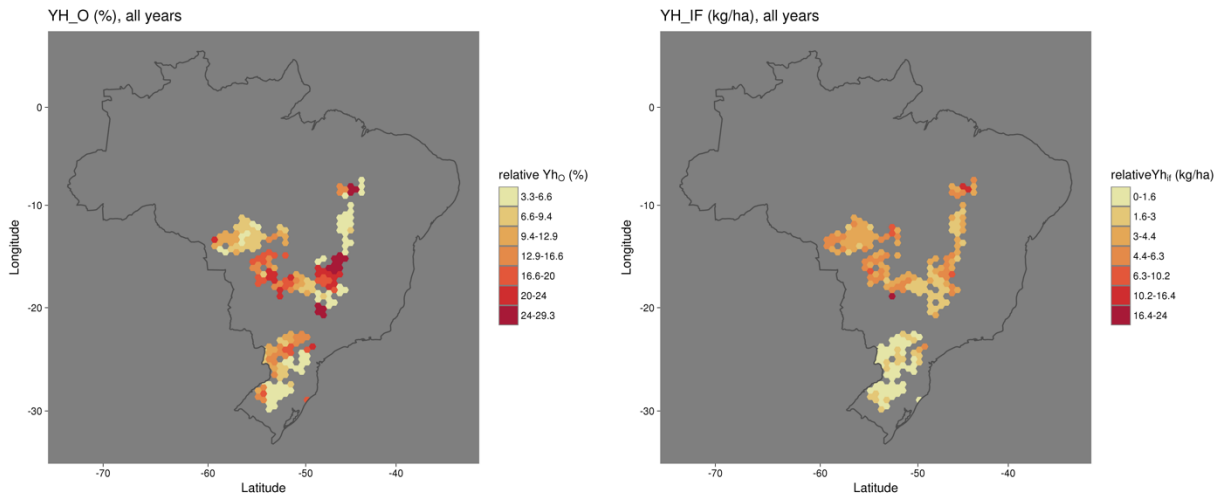


Soy yield models detected crop yield heterogeneity across and within properties.
Out-of-sample yield estimates from a statistical model of soybean yields (SI Text: Soy yield modeling) revealed substantial spatial variation in the overall yield heterogeneity (Yh_O), or the difference between the best farm's yields and the mean yield within each county (Figure 2-2). The average farmer's yield (Y_{am}) in each county was 12.1% lower than the best farmer ($Y_{a,\,95}$),

averaged over years 2001-2016. Within each field, mean yields were 4.6% lower than the highest yields in the field on average (yield heterogeneity within fields, Yh_{if}).

Our satellite-based yield heterogeneity estimates were consistent with existing reports on crop yield gaps due to crop management, or the difference between best farmer's actual yields (Y_{aab}) and the average farmer's actual yields (Y_{aar}). In one study, Brazilian soybean yields gaps due to crop management were estimated using crop simulation models and county-level yield reports (2). The average yield gap due to crop management was estimated as ~13%. The spatial distribution of yield gaps was comparable overall, though the relative yield gap maps disagreed about the magnitude of gaps in southern Mato Grosso and in the Southeast region. Another study of soybean yield heterogeneity in the Midwest United States estimated relative Yh_O for soybeans at 23%, nearly twice of that which we found for Brazil (3). One potential cause of the difference may be differences in reporting unit size–our study reported yield heterogeneity at a coarser scale wherein $Y_{a,\,95}$ may be reduced by more observations near the mean. Yield heterogeneity metrics from out-of-sample soy yield estimates were consistent with those calculated solely from the crop cut survey yields. We found Yh_{if} to be more than one-third of Yh_O, suggesting that field conditions contributed meaningfully to yield variability.

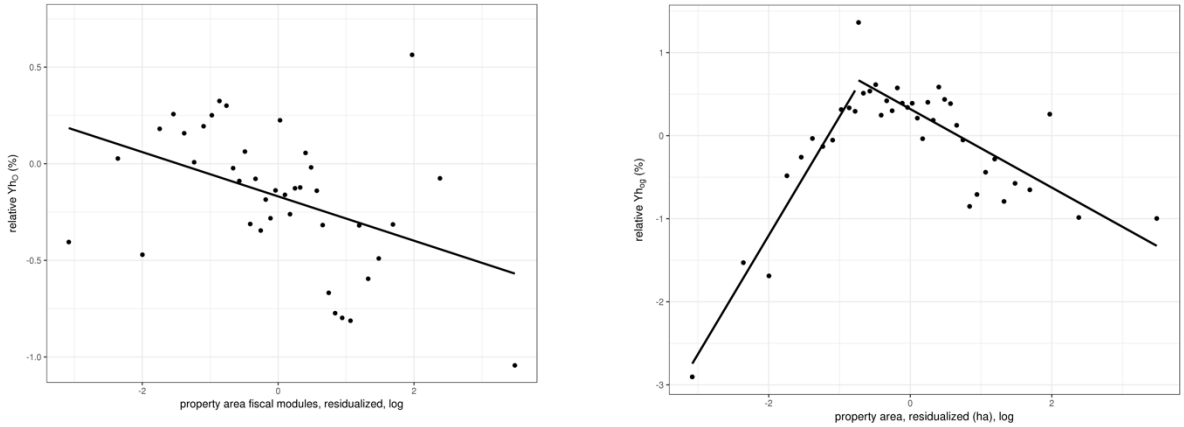
Figure 2 - 2. Relative yield heterogeneity, binned median values, 2001-2016, a) Overall (Yh_O) and b) in-field (Yh_{if})
File: 'graphics/yho_rel.png' and 'graphics/yhif_rel.png'



Soybean yield heterogeneity decreases with increasing farm size
Differences in mean overall yield heterogeneity (Yh_O) across farm size groups were detected (SI Table S5: IR regression table). On average across Brazil, yields gaps decreased with increasing farm size (Fig. 1-3a). The effect was robust to different farm size definitions (SI Text: Farm size definition). The strength and degree of the relationship varied across states with a slight positive relationship between Yh_O and farm size detected more commonly in southern Brazil. When yields were compared with yields from similarly sized farm (Yh_{og}, rather than all farms, Yh_O increased with farm size gaps among smaller farms and decreased among larger farms (Fig. 1-3b). Linear regression methods may be limited in their capacity to capture variation in the

relationship between Yh_{og} and farm size, and semi-parametric models should be considered for future work. Hierarchical modeling methods might also improve performance over linear regression methods by leveraging systematic regional patterns.

Figure 2 - 3. Effect of changes in farm size (FM_{pred}) on yield heterogeneity (Yh_O) after controlling for covariates.
Note: See SI Text: Binscatter plots.
File: 'graphics/ir/yho_rel_fm_parcel_est.png' and 'graphics/ir/yhog_rel_fm_parcel_est.png'
Notes:



Methods

Yield data

A cross-sectional survey of soy fields across central Brazil was carried out each year between 2006 and 2015. The survey, the *Rally da Safra*, was performed by Agroconsult, a private firm in Brazil. Survey teams traveled a prescheduled route through the country designed coincide with expected physiological maturity of the plants at that geography. Survey teams used 1x1 m quadrat sample area to estimate planting density and then clipped to plants in the quadrat to obtain seed number and weight. An expert opinion of crop yield was recorded by the surveyor. Additional information was recorded including growth stage, intensity of pest damage, and, in some cases, management factors such as chemical application frequency. The sample location was recorded with GPS. We applied a data cleaning procedure which removed errant coordinates, and adjusted GPS positions to align with RS imagery pixels classified as soy according to an agricultural land cover classification dataset we developed (see SI Text: Survey data preparation).

An unbalanced panel dataset of annual municipality-level soybean yield estimates was drawn from the Brazilian Institute of Geography and Statistics' Municipal Agricultural Production survey (abbreviated from Brazilian Portuguese as PAM) for the years 2001-2015 (19). PAM yield estimates conducted by the agency were based on county-level soy area estimates and reported production weights.

Image processing

We collected imagery from the moderate resolution imaging spectroradiometer (MODIS) sensor 8-day, 500 m product (MOD09A1) for all available dates between 2001 and 2016. The MODIS internal cloud mask was used to select cloud-free pixels. MODIS pixels were further filtered by selecting only pixels with the highest quality rating for the red (ρ_{red}, 620–670 nm) and near infrared bands (ρ_{NIR}, 841–876 nm). We calculated the Wide Dynamic Range Vegetation Index (WDRVI) which is more responsive to high density crop canopies than NDVI and other vegetation indices (20).

We prepared a daily time series of WDRVI observations for each growing season in order to estimate crop phenology properties which can explain yield variability. We exploited pixel-level collection dates by splitting each 8-day image into daily images in order to gain higher temporal precision for estimating the timing of phenology changes (21). WDRVI observation dates were reindexed to an August 1-July 31 calendar (i.e. August 1 = day of season (dos) 1 and July 31 = dos 365 (366 for leap years) to align with the Brazilian soy growing season which straddles two calendar years. The year in which harvest took place (the 'harvest year') was used to label each growing season. The WDRVI time series for each pixel was filtered and smoothed by first performing linear interpolation for missing values in the daily time series, and then reducing noise and smoothing the time series with the Savitzky-Golay algorithm (22, 23). The resulting smoothed daily WDRVI time series is used for analysis. See SI Text: MODIS data preparation for more information. Phenology metrics were calculated from the smoothed WDRVI time series to approximate crop characteristics related to yield and crop type (SI Text: Phenology and Figure S1). Phenology metrics were calculated solely for the first crop in a double crop sequence, or the only crop in single crop production.

Yield heterogeneity estimation

We computed yield heterogeneity as the difference between the best farmer-attained yields ($Y_{a,\,95}$) and mean (Y_{aa}) or point-level yields. The 95th percentile yield was used to proxy highest attainable yield ($Y_{a,\,95}$) (3). The overall yield heterogeneity (YH_O) was calculated as the difference between $Y_{a,\,95}$ and the average actual yield (Y_{aa}) for each unique year and one decimal degree grid area. Within-field yield heterogeneity (Yh_{if}) was calculated at each point as the difference between $Y_{af,\,95}$ and predicted yield Y_{pred}. Yield stability was estimated as the standard deviation of detrended yields over time at each point. Detrended yields were calculated by 1) regressing Y_{pred} on year of observation, and 2) subtracting the portion of yield explained by time from estimated yields.

IR analysis

Using a database of georeferenced rural property boundaries circa ~2013, the *Cadastro Ambiental Rural* (CAR, See SI Text: Parcel preparation), we constructed a stratified random sample of properties with prevalent soy production where strata included property size class and geographic region (See SI Text: Farm size metrics). A time-varying estimate of parcel fiscal module counts was created as a proxy of expected land productivity (See SI Text Fiscal module estimates). Soybean yields were estimated out-of-sample using a selection of models (SI Text: Soy yield modeling) at sample point locations within each property.

Regression analysis and binscatter plots were used to estimate the linear relationship of farm size measures with soy yield, yield heterogeneity, and management factors. We regressed each of several yield structure metrics on farm size, crop management factors, climate and soil factors, and county-year fixed effects. Binscatter plots were used to demonstrate the non-linear relationship between variables of interest while controlling for covariates.

Agricultural land cover classification

We synthesized two existing agricultural land cover products with the MODIS WDRVI time series described above to classify crop types across Brazil for years 2001-2016. One product (24) provided the desired crop classes, but was limited in spatial extent (greater Mato Grosso) and time (2001-2013), while another (25) identified only soy for a limited spatial extent (Cerrado biome) and time (2001-2013). Both datasets were constructed from ground truth data in combination with remote sensing imagery. We used the soy-only dataset to mask out pixels in the multiple-crop dataset which disagreed about the presence/absence of soybeans. Using a crop-class stratified random sample of points (where for soy only the intersection of soy pixels from the two land cover datasets was sampled), we employed a gradient boosted forest classification model (26) to predict land cover class with a vector of daily WDRVI values, and a vector of WDRVI summary statistics for the season. Repeated cross-validation yielded a classification accuracy of 82.0% and Cohn's kappa of 0.784. See SI Text: Agricultural land cover preparation.

References

1. Lobell DB, Cassma KG, & Field CB (2009) Crop Yield Gaps: Their Importance, Magnitudes, and Causes. *Annual Review of Environment and Resources* 34:179-204.
2. Sentelhas PC*, et al.* (2015) The soybean yield gap in Brazil – magnitude, causes and possible solutions for sustainable production. *The Journal of Agricultural Science* 153(8):1394-1411.
3. Lobell D & Azzari G (2017) Satellite detection of rising maize yield heterogeneity in the U.S. Midwest. *Environmental Research Letters* 12.
4. Burke M & Lobell DB (2017) Satellite-based assessment of yield variation and its determinants in smallholder African systems. *PNAS* 114(9):2189–2194.
5. van Ittersum MK*, et al.* (2013) Yield gap analysis with local to global relevance—A review. *Field Crops Research* 143:4-17.
6. Barrett CB, Bellemare MF, & Hou JY (2010) Reconsidering conventional explanations of the inverse productivity-size relationship. *World Development* 38(1):88-97.
7. Bevis LEM & Barrett CB (2017) Close to the Edge: Do Behavioral Explanations Account for the Inverse Productivity Relationship? in *Pacific Conference for Development Economics* (Riverside, California).
8. Helfand SM & Taylor MPH (2017) The Inverse Relationship between Farm Size and Productivity: Refocusing the Debate. in *Pacific Conference for Development Economics* (Riverside, CA).
9. Carletto C, Savastano S, & Zezza A (2013) Fact or artifact: the impact of measurement errors on the farm size–productivity relationship. *Journal of Development Economics* 103:254–261.
10. Sen AK (1966) Peasants and Dualism with or without Surplus Labor. *Journal of Political Economy* 74(5):425–450.
11. Chayanov A (1926) *The theory of peasant co-operatives* (Richard D. Irwin, Inc, Homewood, IL).
12. Feder G (1985) The Relation between Farm Size and Farm Productivity: The Role of Family Labor, Supervision and Credit Constraints. *Journal of Development Economics* 18:297–313.
13. Carletto C, Gourlay S, & Winters P (2013) From guesstimates to GPStimates: land area measurement and implications for agricultural analysis. in *World Bank Policy Research Working Paper* (World Bank).
14. Gourlay S, Kilic T, & Lobell D (2017) Errors in Farmer-Reported Production and Their Implications for Inverse Scale -Productivity Relationship in Uganda. in *Center for the Study of African Economies (CSAE) Conference* (Oxford, UK).
15. Alston LJ, Libecap GD, & Mueller B (1999) *Titles, Conflict, and Land Use: The Development of Property Rights and Land Reform on the Brazilian Amazon Frontier* (University of Michigan Press, Ann Arbor, MI).
16. Nobre CA*, et al.* (2016) Land-use and climate change risks in the Amazon and the need of a novel sustainable development paradigm. *Proceedings of the National Academy of Sciences* 113(39):10759-10768.
17. Schmidhuber J & Tubiello FN (2007) Global food security under climate change. *PNAS* 104(50):19703–19708.
18. Binswanger HP, Deininger K, & Feder G (1995) Power, distortions, revolt and reform in agricultural land relations. *Handbook of Development Economics* 3:2659-2772.

19. (IBGE) IBdGeE (2017) Producao Agrícola Municipal—Automatic Data Recovery System—SIDRA.
20. Gitelson A (2004) Wide Dynamic Range Vegetation Index for Remote Quantification of Biophysical Characteristics of Vegetation. *Journal of Plant Physiology* 161:165-173.
21. Guindin-Garcia N, Gitelson AA, Arkebauer TJ, Shanahan J, & Weiss A (2012) An evaluation of MODIS 8- and 16-day composite products for monitoring maize green leaf area index. *Agricultural and Forest Meteorology* 161:15-25.
22. Savitzky A & Golay MJE (1964) Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry* 36(8):1627-1639.
23. Chen J*, et al.* (2004) A simple method for reconstructing a high-quality NDVI time-series data set based on the Savitzky–Golay filter. *Remote Sensing of Environment* 91(3-4):332-344.
24. Spera SA*, et al.* (2014) Recent cropping frequency, expansion, and abandonment in Mato Grosso, Brazil had selective land characteristics. *Environmental Research Letters* 9(6).
25. Gibbs HK*, et al.* (2015) Brazil's Soy Moratorium. *Science* 347(6220):377-378.
26. Friedman JH (2001) Greedy function approximation: A gradient boosting machine. *Ann. Statist.* 29(5):1189-1232.
27. Hengl T*, et al.* (2017) SoilGrids250m: Global gridded soil information based on machine learning. *PLOS ONE* 12(2).
28. Farr TG*, et al.* (2007) The Shuttle Radar Topography Mission. *Reviews of Geophysics* 45(2):n/a-n/a.
29. Willmott CJ & Matsuura K (2017) Terrestrial Air Temperature and Precipitation: Monthly and Annual Time Series (1900-2014).
30. Cohn AS*, et al.* (2014) Cattle ranching intensification in Brazil can reduce global greenhouse gas emissions by sparing land from deforestation. *Proceedings of the National Academy of Sciences* 111(20):7236-7241.
31. Maas SJ (1988) Using satellite data to improve model estimates of crop yield. *Agronomy Journal* 80:655–662.
32. Doraiswamy PC, Moulin S, Cook PW, & Stern A (2003) Crop Yield Assessment from Remote Sensing. *Photogrammetric Engineering and Remote Sensing* 69(6):665-674.
33. Funk C & Budde ME (2009) Phenologically-tuned MODIS NDVI-based production anomaly estimates for Zimbabwe. *Remote Sensing of Environment* 113:115–125.
34. Guan K, Li Z, Rao N, & Feng G (in review) Using MODIS-Landsat fused data and ALOS-2/PARSAR-2 to map paddy rice and estimate crop yield for Thai Binh Province in Viet Nam.
35. contributors O (2017) OpenStreetMap. ed OpenStreetMap (https://www.openstreemap.org/).
36. Jönsson P & Eklundh L (2004) TIMESAT—a program for analyzing time-series of satellite sensor data. *Computers & Geosciences* 30(8):833-845.
37. Kuhn M (2017) caret: Classification and Regression Training).
38. Stephan E, Tobias K, Christian L, Matthias B, & Patrick H (2016) Mapping cropland-use intensity across Europe using MODIS NDVI time series. *Environmental Research Letters* 11(2):024015.
39. Pielou EC (1966) Shannon's Formula as a Measure of Specific Diversity: Its Use and Misuse. *The American Naturalist* 100(914):463-465.

40. Team RC (2016) R: A language and environment for statistical computing (R Foundation for Statistical Computing, Vienna, Austria).

Supporting Information

Fiscal module estimates

A fiscal module (FM) is an agrarian land unit which measures area, adjusted for the typical land needs of dominant economic activity in the region. The Brazilian government publishes for each county the number of hectares per fiscal module, last released in 1993. Economic activity and land qualities change over time, but the typical FM measures are time-invariant. We developed model of hectares per fiscal module (hafm) in order to downscale (in space and time) fiscal modules estimates. We regressed county-level hectares per fiscal module on county-level land values in 1997 (the closest year available), and time-invariant biophysical and economic characteristics (the same as those independent variables described in SI Text: Agricultural land values). See Table S1 for regression results. We then predicted hectares per fiscal module ($hafm_{pred}$) out-of-sample for each county year. For each property, we used the $hafm_{pred}$ to estimate property-level predicted fiscal module counts (FM_{pred}).

Table S1. Fiscal module regression results
Note: Values in parentheses are standard errors.
File: 'tables/fm_model.html'

	ha per fiscal module
ag. land value (reais)	-0.003*** (0.0001)
bedrock depth (cm)	-0.002*** (0.0002)
clay (%)	11.372*** (2.094)
silt (%)	11.385*** (2.085)
sand (%)	11.735*** (2.093)
longitude	-1.918*** (0.034)
latitude	1.398*** (0.042)
Constant	-1,166.412*** (209.245)
Observations	5,446
R^2	0.697
Adjusted R^2	0.697
Residual Std. Error	12.705 (df = 5438)
F Statistic	1,791.057*** (df = 7; 5438)
Note:	*p<0.1; **p<0.05; ***p<0.01

Agricultural land values

By developing a 500-m gridded agricultural land value dataset, we aimed to index a wide range of social, economic, and geographic crop yield drivers. We use an unbalanced panel dataset of agricultural land values as the county level (1997-2015) to construct a statistical model relating land values to a collection of biophysical, economic, and climatic factors. Independent variables included 1) 11 soil variables, five of which at seven depths (27), 2) elevation, slope, and aspect at

500 m resolution from the Shuttle Radar Topography Mission (SRTM, (28), 3) coastal and river distance, and 4) 17-year (1980-1996) mean monthly temperature and cumulative precipitation estimates at 0.5 degree resolution (29), 4) farm to market/port soybean transportation costs, gridded at 1 km (30) and 5) year of observation. To estimate a statistical model of land values, all independent variables were averaged at the county-level. A gradient boosted forest model (GBM) with repeated cross-validation was estimated with the municipality-level panel. The best model parameter set had an RMSE of 1,787.15 (R$) and an R^2 of 0.957. The parameterized model was then forced with gridded variables to create a 500 m, spatially explicit agricultural land value dataset for years 1997-2015. In each year, land values at sample locations were binned into quartiles. For each CAR parcel, we calculate the average land value for each year.

Binscatter plots

Binscatter plots convey the effect of one variable on another. Each variable was regressed on a set of control variables and each regression's residuals saved. Residuals were then averaged within equal sized bins defined by the x-axis variable. The slope between x- and y-axis residuals is equal to the beta coefficient for the x-axis variable in a regression of the y-axis variable on the x-axis variable plus controls. We implemented the method in R following the Stata package 'binscatter'.

Soy yield modeling

A suite of soybean yield models compared the relative performance of two classes of models, and three sets of variables. All of the yield models included vegetation index (VI) metrics from remote sensing imagery which have well-documented associations with crop yields (31, 32). We calculated VI metrics for pre- and post-VI peak periods. We compared three yield models: 1) "RS", a model with only RS metrics and no climate variables, 2) "RS + climate", with both RS and climate, and 3) "RS + climate split" with RS metrics only during the pre-peak period and only climate during the post-peak period. Model "RS" serves as a baseline model while, "RS + climate" and "RS + climate split" are motivated by the theory environmental stress metrics (e.g. extreme temperatures) should explain additional variation in yields as VI metrics may not adequately convey plant health during early vegetative and reproductive stages due to sparse canopies or natural yellowing of leaves during grain filling (33, 34).

We found that all three sets of independent variables perform similarly, as measured by adjusted R^2 and root mean squared error (RMSE) from repeated cross-validation (10-folds, 5-repeats). See Table S2. Results were found to be robust to three types of models: ordinary least squares (OLS), generalized linear models (GLM) with Gaussian error distribution, and gradient boosted forest model. GLM models were used due to their prevalence in the literature and comparability with other methods tested. Fixed effects for unique combinations of spatial units and years significantly improved yield performance. Though county-year fixed effects resulted in better model performance (R^2 = .38), they limited the spatial and temporal scope of out-of-sample yield predictions. Predictions could only be made in county-years for which there were adequate observations in the training dataset to determine the fixed effect intercept. Ultimately, region-year fixed effects were chosen so as to maintain the scope of out-of-sample predictions.

Table S 2. Comparison of regression results for three alternative model specifications for estimating surveyed crop yields.
Note: Values provided in parentheses are standard errors.
File: 'tables/yield_model_fill.html'

	Soy yield		
	RS + climate split	RS + climate	RS
season start day	0.003** (0.001)	0.003** (0.001)	-0.00001 (0.001)
greening AUC	-0.00000 (0.00000)	-0.00000 (0.00000)	-0.00000 (0.00000)
peak WDRVI	0.0001 (0.0001)	0.0001 (0.0001)	0.0002 (0.0001)
browning AUC		-0.00000 (0.00000)	0.00000 (0.00000)
greening length (days)	0.004 (0.003)	0.006 (0.013)	0.002 (0.003)
browning length (days)	-0.013*** (0.004)	-0.014*** (0.005)	0.0002 (0.002)
GDH during greening		-0.00000 (0.00000)	
GDH during browning	-0.00000 (0.00000)	-0.00000 (0.00001)	
EDH during greening		0.00003 (0.00004)	
EDH during browning	-0.0001 (0.0001)	-0.0001 (0.0001)	
precip. during greening (mm)		-0.0002 (0.0003)	
precip. during greening 2 (mm)		0.00000 (0.00000)	
precip. during browning (mm)	0.001*** (0.0003)	0.001*** (0.0003)	
precip. during browning 2 (mm)	-0.00000*** (0.00000)	-0.00000*** (0.00000)	
solar rad. during greening (W m-2)		0.00005 (0.0003)	
solar rad. during browning	0.0005*** (0.0002)	0.0004** (0.0002)	
clay (%)	-0.005 (0.005)	-0.005 (0.005)	-0.005 (0.005)
bedrock depth (cm)	-0.015 (0.025)	-0.015 (0.025)	-0.013 (0.025)
avail. water capacity	0.048*** (0.014)	0.048*** (0.014)	0.048*** (0.013)
Constant	3.316 (4.997)	3.214 (5.003)	3.611 (4.998)
Region-year fixed effects	Yes	Yes	Yes
Adj R2, cv	0.229	0.229	0.23
RMSE, cv	0.878	0.878	0.879
Variance, fitted, cv	0.269	0.27	0.27
Variance, obs., cv	1.041	1.041	1.042
Observations	8,166	8,166	8,237

Note: Standard error provided in parentheses $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table S 3. Farm size groups

Farm size groups (ha)
20-<50
50-<100
100-<200
200-<500
500-<1000
1000-<2500
2500-<5000
5000-<7500
7500-<10000
>10000

Survey data preparation

Rally da Safra surveys were conducted field edges and along roads. Locations were marked with GPS. However, due to the relatively coarse resolution of the 500 m MODIS imagery used in study, the MODIS pixel containing the GPS point may have contained several land cover types (e.g. roads, adjacent grassland, crops other than soybean). Heterogeneous land cover within a pixel introduced noise into the WDRVI time series used for yield modeling. We implemented a process to nudge GPS point locations which fall inside pixels which were not classified as soybean in the land cover classification to nearby pixels that are likely to only soybean land. We calculated the distance between the GPS location and the center point of each pixel with a high (≥99%) probability of being soy in the corresponding survey year. We then used OpenStreetMap road lines (35) to classify whether survey points and soy pixel points were on the same side of the road, based on the number of times a line between survey points and pixel centroids transected the roads layer. If the nearest soybean pixel centroid was on the same side of the road as the GPS point, that point was taken as the updated or 'nudged' survey point location. If no candidate soy pixel centroids were present with 1,000 m on the same side of the road, the nearest point was used.

IR sample

We drew a stratified random sample of soy points where strata included year, county, and farm size group (Table S3: Farm size groups).

MODIS data preparation

We collect MOD09A1 imagery from NASA's MOLT server for all tiles which overlap with Brazil, and for all dates available (2/18/2000-2/2/2017). For each image tile and date, we extracted the following bands: red (620–670 nm) and near infrared (841–876 nm, NIR) reflectances, state flags, reflectance band quality control, and day of year. The red and NIR bands were masked to exclude any pixels where either band has less than the highest quality. We excluded all pixels labeled as 'clouds' by the internal cloud algorithm. The wide dynamic range vegetation index (WDRVI) was calculated for all pixels:

$$WDRVI = (\alpha * \rho_{NIR} - \rho_{red}) / (\alpha * \rho_{NIR} + \rho_{red}) + 1$$

where α is a weighting coefficient set at 0.2 which linearizes the relationship between vegetation density and VI score. We modified the index to enhance storage and computational efficiency by adjusting the range from -1 to 1, to 0-2, and then scaling up so that the value can be stored as an integer.
E.g. : $WDRVI_{mod} = (WDRVI + 1) * 1000$

For each growing season, we smoothed each pixel's WDRVI values. We extracted WDRVI values for imagery between 60 days prior to August 1st and 60 days after July 31 of the following calendar year to ensure that the crop growing period was captured. The pixel day of year was extracted for each image the date range. For each pixel, the result as a ~485 day time series with a 1-day interval. We first applied linear interpolation between observations, though non-linear interpolation methods were also considered but rejected due to their tendency to overestimate maximum values. Then the Savitzy-Golay smoothing filter with p = 3, and n = 31 was applied to the time series. Multiple parameter combinations were evaluated and the selected values were those which had best trend fitting without overfitting. Values outside the range of our $WDRVI_{mod}$ (0-2000) were set to their respective minimum and maximum values. We preserved the daily-level smoothed WDRVI estimates as rasters corresponding to each date.

Phenology

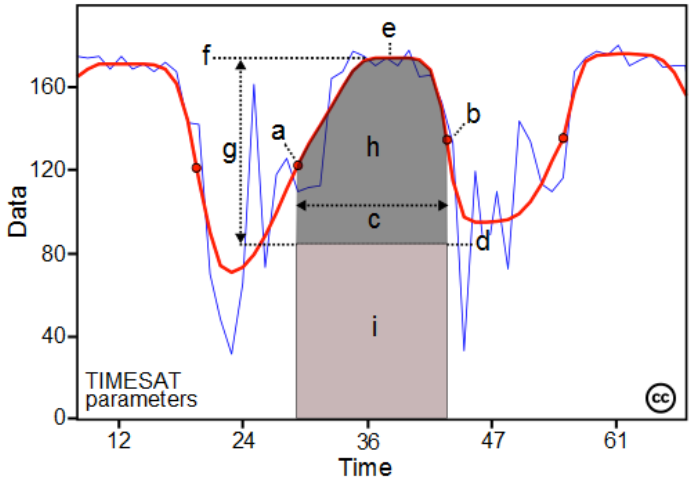
Several metrics were calculated from seasonal WDRVI time series to capture phenological properties of the crop. The VI metrics were calculated within a constrained range of days in the season so as to avoid capturing characteristics of a second crop, if present. Guided by the list of metrics calculated by the widely used TIMESAT software (36), we calculated the following, indicated with letters in Figure S1:

- WDRVI peak value (e, wdrvi_peak)
- WDRVI peak value day of season (e, wdrvi_peak_dos)
- WDRVI amplitude between peak and a "base value" (e – d, wdrvi_amp)

And the following where each was calculated for periods "green" (WDRVI green-start threshold to peak), "brown" (WDRVI peak to brown-stop threshold), and "season" (WDRVI green-start threshold to brown-stop threshold):

- Length of period (a to e = green_len; e to b = brown_len, a to b = season_len)
- Area under curve (AUC, trapezoidal integration)
- AUC large, from 0 to curve (h + i, e.g. green_auc_lg)
- AUC small, from a "base value" to curve (to reduce noise) (h, e.g. green_auc_sm)

Figure S1. An example time series plot with labels indicating features described in the SI Text section “Phenology”. Reproduced from (36).



Parcel preparation

We used a database of Brazilian rural property boundaries to estimate property size and crop area at the locations of soybean yield estimates. Under the 2012 Brazilian Forest Code, property owners are required to register their land in the Rural Environmental Registry (*Cadastro Ambiental Rural*, CAR). We obtained a copy of the registry, containing spatial data for more than three million properties across Brazil, circa ~2013. Associated with each parcel were additional layers documenting the areal boundings of legally reserved for Forest Code compliance, in native vegetation, and reserved for municipal and infrastructural use.

We preprocessed the CAR dataset to correct for undesirable artifacts including partial overlap along parcel edges, complete overlap of parcel by another, and invalid geometries. Geometry issues were resolved with a hierarchical procedure which included a succession of small buffers (0-0.001 m), polygon simplification, and polygonation (using the R package ‘cleangeo’). Partial overlap between parcels was treated by 1) intersecting the polygons with themselves, 2) deletion of one of the duplicate intersected polygons, 3) random assignment of the overlapping piece to one of the original polygons, 4) merger of parcel pieces. There was a small but significant number of instances where one large property polygon completely covered a large number of smaller parcels and in these cases, we sought to preserve the collections of small parcels. We identified cases needing treatment by counting the number of parcels intersecting the each other parcel, and selecting those which had greater than 20 intersections. We then replace the large polygon with the difference between itself and the smaller polygons it contained.

Farm size metrics

We calculated several attributes of each CAR parcel to assist our analysis of soy yields and farm size. Area calculations were performed on data with the South American Albers Equal Area Conic projection. Property area (ha) and productive area (ha) area calculated, where productive area was the property area less the area of land portions reported by the CAR database as: legal reserves, areas of permanent protection, area of restricted use (due to terrain properties), native

vegetation, hydrology, and municipal or infrastructural use. We calculated the number of fiscal modules per property, where a fiscal module is a unit of agrarian land (area scaled by land quality) used for regulatory purposes.

We calculated land cover features for each parcel and year, drawn from the good-confidence land cover product and its associated metrics. For each parcel, we tabulated the mean of pixels whose center points fall within the bounds of the parcel. These properties included intensity and diversity metrics, land value and soil characteristics. We estimated the area of each land cover class in the agricultural land cover dataset, for each year. To improve the precision of the estimates, we first disaggregated the 500 m land cover cells to 50 m pixels before summing the area.

Agricultural land cover preparation

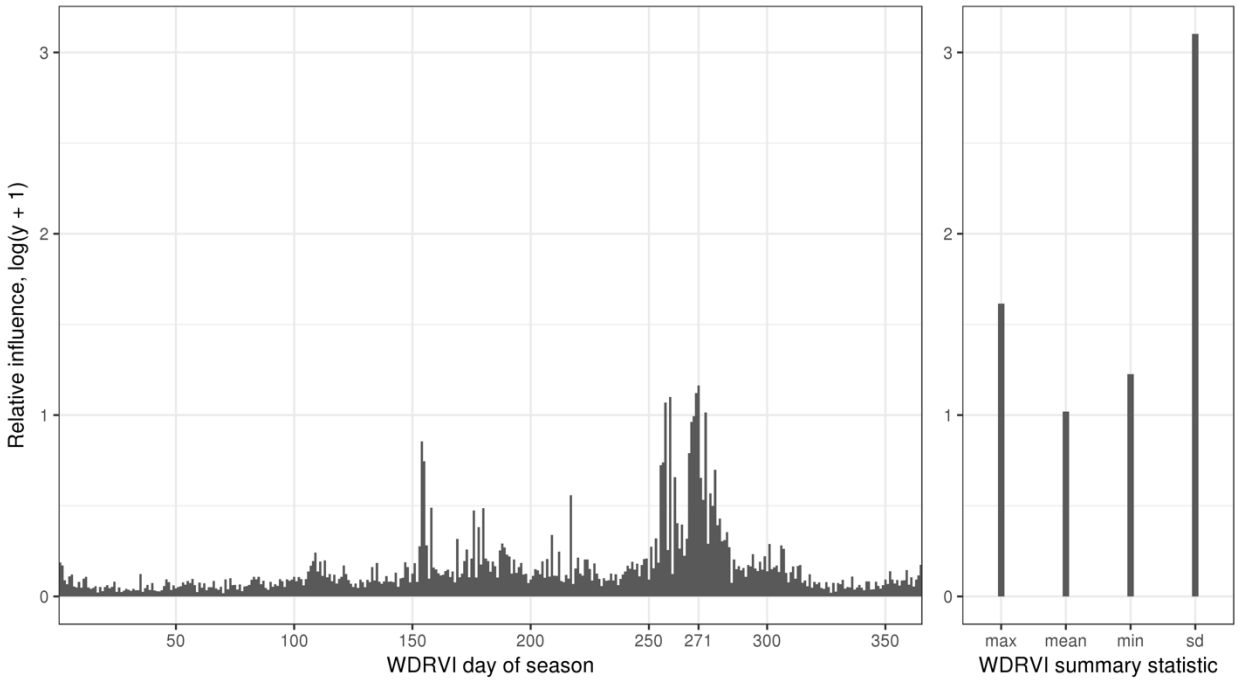
We combined two existing agricultural land cover products available in Brazil in order to predict soybean croplands outside of the region and time period for which these datasets were available. The first, 'Spera' dataset (24), provided 250 m resolution classifications for major crops including soy, soy/corn (i.e., double cropped), soy/cotton, corn, cotton, sugarcane, and irrigated crops. Two additional classes included all land not in mechanical agriculture, and another for unclassifiable pixels. We collapsed the categories 'irrigated', 'sugarcane', and 'unclassifiable' into the 'not mechanical agriculture' category. A second data source, 'Gibbs' dataset (25) was a MODIS-based soybean presence/absence product.

We conducted stratified random sampling across each Spera land cover class (with the aforementioned regrouping) and year, constrained in space to the Spera coverage area (greater Cerrado region). We extracted Gibbs soy presence estimates to the sample points. Then, and only for the soybean classes, we dropped observations where the Spera and Gibbs datasets disagreed on whether or not the pixel was soybean. The level of agreement between the two datasets was recorded. We extracted WDRVI estimates for each day in the respective year for each sample point. For each sample point and year, WDRVI summary statistics including mean, minimum, maximum, and standard deviation were calculated. Land cover classifications were modeled as a function of a vector of daily WDRVI values and the aforementioned summary statistics. A gradient boosted forest classification model was trained and validated on a random stratified sample (n = 10,000) of the labelled training data. Model parameters (interaction depth, number of trees, and shrinkage) were tuned by assessing the performance of models estimated with each unique combination of parameter setting using the R package "caret" (37) (see also SI Text: Software). Cross-validation was used to assess the model performance (10 folds, 1 repetition). In the final model the fraction of land cover classes correctly predicted was ~0.82. See Table S4 and Figure S2.

Table S 4. Agricultural land cover classification confusion matrix
File: 'tables/lc_confusion_matrix.png'

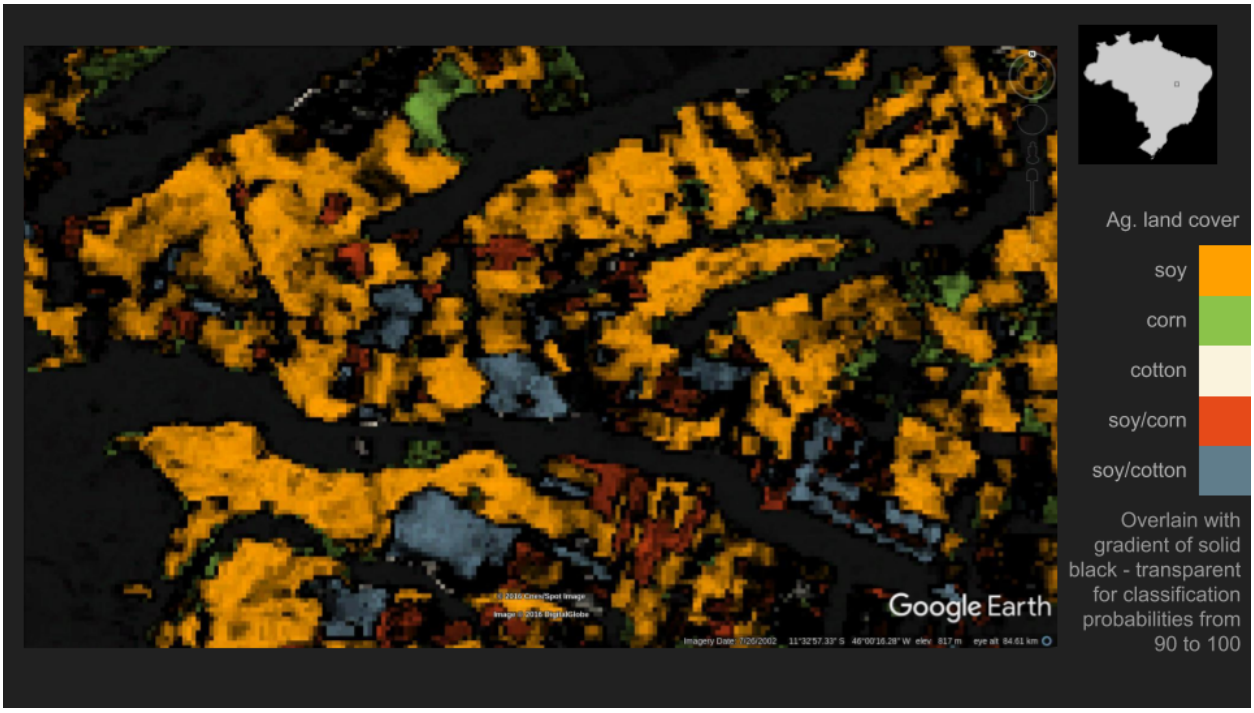
Prediction	Reference not mech. ag.	soybean	corn	cotton	DC soy/cotton	DC soy/corn
not mech. ag.	14.1	0.2	0.9	1.0	0.2	0.2
soybean	0.1	12.4	1.7	0.1	0.2	1.3
corn	1.1	2.3	13.4	0.6	0.1	0.4
cotton	1.0	0.3	0.3	14.1	0.7	0.3
DC soy/cotton	0.1	0.1	0.0	0.6	14.3	0.9
DC soy/corn	0.2	1.5	0.3	0.3	1.3	13.6

Figure S2. Distribution of variable importance for classifying agricultural land cover with gradient boosted forests.
File: 'graphics/lc_model_rel_inf.png'



The model was then used to predict land cover classes for all pixels in Brazil, and for all growing seasons between 2001 and 2016, inclusively. For each pixel, the model estimated the probability of the presence of each land cover class. Recording these probabilities, we created two versions of the land cover dataset. A 'good-confidence' land cover product was constructed where only land classifications with ≥90% probability were retained while all others were marked as 'unclassifiable'. A 'better-confidence' land cover product was created with the same logic, except with a probability threshold of ≥99%. The better-confidence land cover dataset was used for applications where it was important to identify pixels with low within-pixel land cover heterogeneity. See Figure S3.

Figure S3. Example of the agricultural land cover dataset with confidence filter
File: 'graphics/land_cover_example.png'



Land cover intensity and diversity

The intensity of agricultural production was correlated with the suitability of land, skill of farmers, and other factors influencing crop yields and yield variability (38). We calculated several metrics of agricultural land use intensity including 1) fraction of years cropped, 2) fraction of years double cropped, 3) temporal Shannon-Wiener diversity index (39), and 4) a binary indicator of whether or not the previous crop was soy. The intensity metrics were each calculated for two temporal windows–a five-year lag period trailing but excluding the observation year, and a variable length window of all years with available data trailing but excluding the observation year. All intensity metrics were calculated using the land cover dataset described above. See Figure S4.

Figure S4. Example of soybean cultivation intensity map over Mato Grosso.
File: ‘graphics/soy_intensity.png’

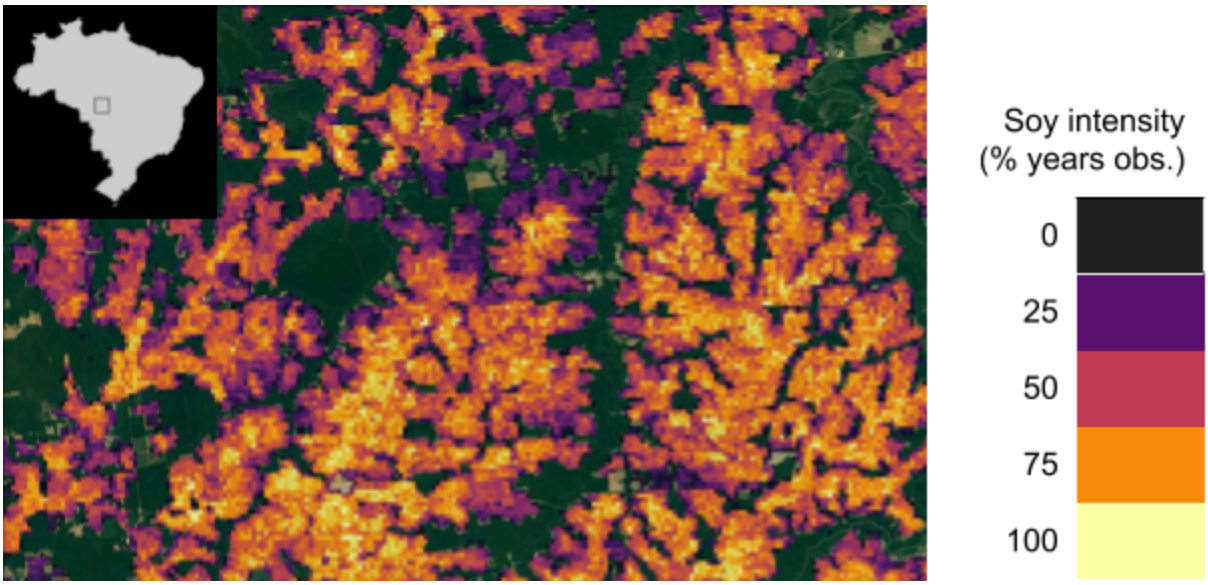


Figure S5. Binscatter plot of the effect of changes in fiscal modules on Yh_{if}
File: 'graphics/ir/yhif_rel_fm_parcel_est.png'

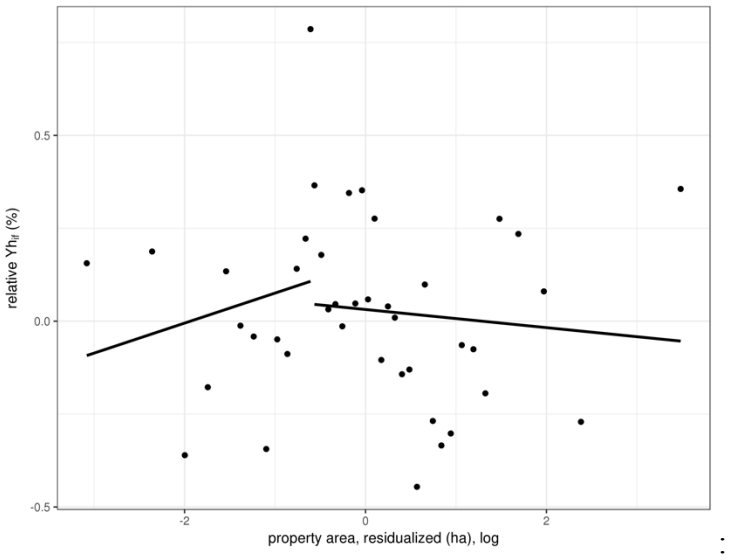


Figure S6. Binscatter plot of the effect of changes in fiscal modules on Yh_{if}
File: 'graphics/ir/yield_stability_fm_parcel_est.png'

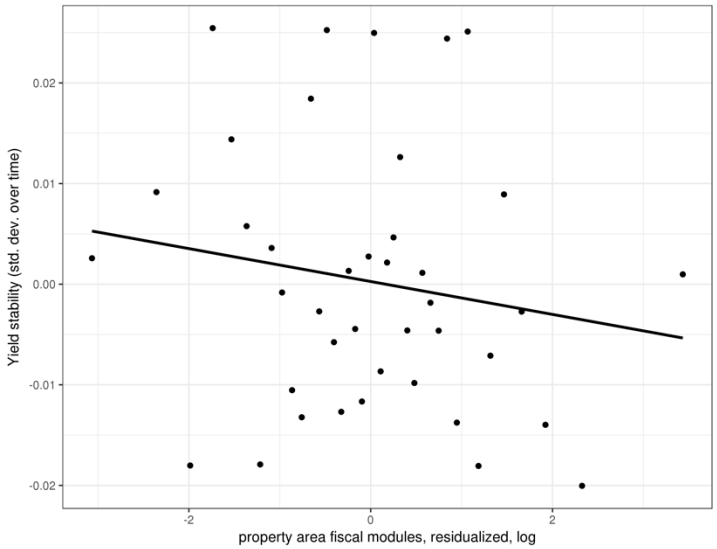


Figure S7. Binscatter plot of the effect of within parcel starting date variability yield
File: 'graphics/ir/season_start_dos_sd_within_yield_hat.png'

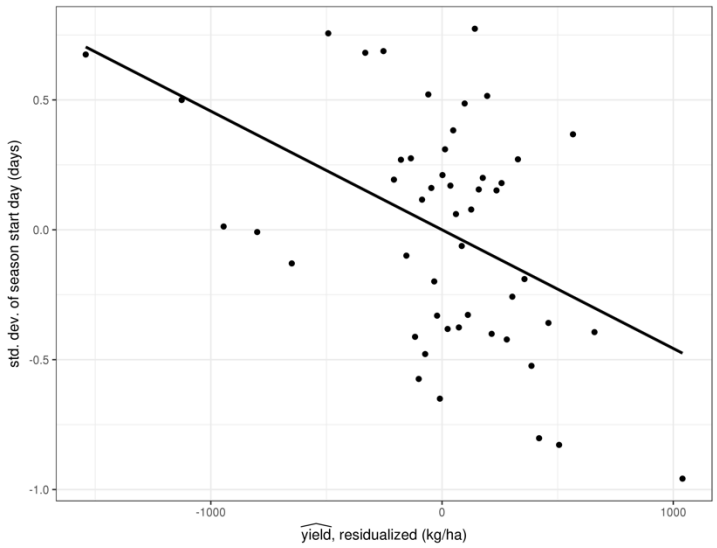


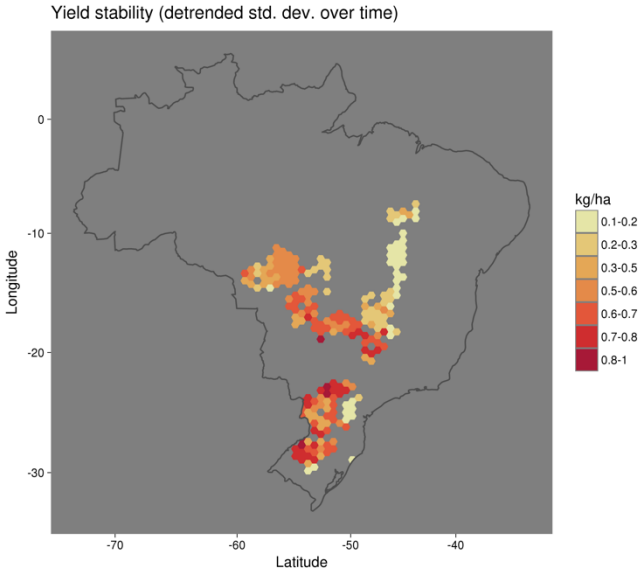
Table S 5. IR regression table
See online data. 'tables/ir_yield.html' and state-level tables with the pattern 'tables/ir_yield_<state name>.html'

Additional results

Nationally, within field yield heterogeneity (Yh_{if}) increases among smaller farms and decreases among larger farms.
The difference between the best yield in a field and the average yield (Yh_{if}) was in weak non-linear relation to farm size. Initially Yh_{if} increases along with farm size, but for larger farms in-field yield heterogeneity decreased (Figure S5).

Soy yields were more stable among medium sized farms than small or large farms
Interannual variation in soybean yields was higher among small and large farms than medium sized farms. Regions in southern Brazil had less stable yields relative to the Center West and Southeast regions of Brazil.

Figure S8: Hexbin plot of predicted yield stability across Brazil.
File: 'graphics/yield_stability.png'



The amount of deviation between optimal and achieved start date was not strongly related to farm size.
The amount of variation in planting date within a parcel was inversely related to yield (Figure S7). However, the deviation from optimal starting date and farm size measures was weak and negative at the national level. State by state analysis showed mostly weak relationships with mixed signs.

From theory, we anticipated that variation in season start dates within each property would be higher among larger farms, as labor and capital constraints limit the amount of land that can be planted in one day (for many large farms, planting events can span multiple weeks). Due to these planting timing constraints, we also expected large farms to more frequently have starting dates which deviated from the optimum. However, only a weak non-linear relationship was observed at the national-level and state-level analysis did not bring the relationship into focus. Our analysis concurred with other work on the relationship between planting date and yield (17).

Because adequate data on planting dates was not available, we developed a proxy 'season start date' as the day of season when green up occurs (SI Text: Phenology). Errors in start date approximation may have been present due to spatial and temporal coarseness in MODIS imagery, and exacerbated by problematic cloud coverage during the growing season (18).

Software

All analyses were performed using the R programming language (40), version 3.3.2, running on Red Hat Enterprise Linux 6.9 in the Tufts University high-performance computing environment. R session information:

```
R version 3.3.2 (2016-10-31)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Red Hat Enterprise Linux Server release 6.9 (Santiago)

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] splines   parallel  stats     graphics  grDevices utils     datasets
[8] methods   base

other attached packages:
 [1] readstata13_0.9.0   randomForest_4.6-12 cleangeo_0.2-2
 [4] maptools_0.9-2      gbm_2.1.3           survival_2.41-3
 [7] reshape_0.8.6       matrixStats_0.52.2  biganalytics_1.1.14
[10] biglm_0.9-1         DBI_0.5-1           signal_0.7-6
[13] stargazer_5.2       viridis_0.4.0       viridisLite_0.2.0
[16] plm_1.6-5           Formula_1.2-1       caret_6.0-76
[19] ggplot2_2.2.1       lattice_0.20-34     openxlsx_4.0.17
[22] gdalUtils_2.0.1.7   rvest_0.3.2         xml2_1.1.1
[25] stringr_1.2.0       doMC_1.3.4          doParallel_1.0.10
[28] iterators_1.0.8     foreach_1.4.3       data.table_1.10.5
[31] futile.logger_1.4.3 bigmemory_4.5.19    bigmemory.sri_0.1.3
[34] raster_2.5-8        foreign_0.8-68      pracma_2.0.4
[37] zoo_1.8-0           dplyr_0.7.2         plyr_1.8.4
[40] reshape2_1.4.2      lubridate_1.6.0     rgeos_0.3-23
[43] rgdal_1.2-7         sp_1.2-5

loaded via a namespace (and not attached):
 [1] httr_1.2.1          R.utils_2.5.0       assertthat_0.2.0
 [4] stats4_3.3.2        quantreg_5.33       glue_1.1.1
```

```
 [7] quadprog_1.5-5       minqa_1.2.4          colorspace_1.3-2
[10] sandwich_2.3-4       Matrix_1.2-8         R.oo_1.21.0
[13] pkgconfig_2.0.1      SparseM_1.77         scales_0.4.1
[16] lme4_1.1-13          MatrixModels_0.4-1   tibble_1.3.3
[19] mgcv_1.8-17          car_2.1-4            pacman_0.4.7
[22] nnet_7.3-12          lazyeval_0.2.0       pbkrtest_0.4-7
[25] magrittr_1.5         R.methodsS3_1.7.1    nlme_3.1-131
[28] MASS_7.3-45          tools_3.3.2          munsell_0.4.3
[31] lambda.r_1.1.9       bindrcpp_0.2         rlang_0.1.1
[34] grid_3.3.2           nloptr_1.0.4         gtable_0.2.0
[37] ModelMetrics_1.1.0   codetools_0.2-15     R6_2.2.2
[40] gridExtra_2.2.1      bdsmatrix_1.3-2      bindr_0.1
[43] futile.options_1.0.0 stringi_1.1.5        Rcpp_0.12.12
[46] lmtest_0.9-35
```

Chapter 3: Assessing the sub-field accuracy of satellite-based, scalable crop yield mapping methods

Prepared in the style of *Precision Agriculture*

Authors

Graham R. Jeffries
Timothy S. Griffin
David H. Fleisher
Elena N. Naumova
Magaly Koch
Brian D. Wardlow

Author Affiliations

G.R. Jeffries, T.S. Griffin, and E.N. Naumova, Friedman School of Nutrition Science and Policy, Tufts University, 150 Harrison Ave, Boston, MA 02111; D.H. Fleisher, Adaptive Cropping Systems Lab., 10300 Baltimore Ave., Beltsville, MD 20705; M. Koch, Center for Remote Sensing, Boston University, 725 Commonwealth Avenue, Boston, MA 02215; B.D. Wardlow, Center for Advanced Land Mgmt. Info. Tech.; 3310 Holdrege St., Lincoln, NE 68583.

Corresponding Author

graham.r.jeffries@gmail.com
717.649.6907

Abstract

Crop yield maps are valuable for many applications in precision agriculture, but are typically only available for sites and years where yields were recorded with a harvester yield monitor. A method for mapping sub-field crop yields from remote sensing imagery could lower the cost and increase the availability of crop yield maps. We tested an algorithm for creating maize (*Zea mays* L.) yield maps (10 m) without *in situ* observations. Crop simulation model outputs were used to fit statistical models which then predicted yield using remote sensing imagery (Landsat and hyperspectral). The method was validated using harvester yield monitor records for 21 site-years in eastern Nebraska, USA, two irrigated and one rainfed. We tested alternative specifications of the prediction algorithm, and the preferred method explained 67.7 percent of the variation in pixel-level yields across all fields (RMSE = 0.99; NRMSE = 0.08). Predictive performance was typically higher for rainfed sites and in locations with more than five observation years. Linear regression models were outperformed by gradient boosted forest models for predicting mean yields, but not for capturing within-field variation. Significant but correctable proportional bias in predictions was detected. Scalable crop yield mapping methods show promise for precision agriculture applications.

Keywords

remote sensing, crop simulation, yield mapping, machine learning, yield monitor

Background

Projected population expansion and dietary changes toward more resource-intensive foods create an imperative for increasing global agricultural output (Godfray and Garnett, 2014). In light of the substantial environmental impact of converting native ecosystems into agriculture, research efforts focus on increasing yield, production per unit area (Garnett et al., 2013). Precision agriculture (PA) practices may help farmers achieve higher and more stable yields with site-specific management (Cassman, 1999). PA management commonly integrates data from multiple sources, collected *in situ* and remotely, to document the quality of growing conditions within each field. Areas with similar growing conditions are grouped together into management zones. For each zone, a tailored 'prescription' of inputs can then designed to optimize returns or achieve other goals. Inputs are applied according to the prescription using variable rate technologies such as planters and fertilizer applicators.

Spatial crop yield maps are valuable for delineating management zones, designing input prescriptions, and tracking crop performance (Zhang et al., 2002). Maps of multi-season mean yields are commonly used to identify areas of persistent high or low yields which may be correlated with stable site characteristics (e.g. soils and topography) (Schepers et al., 2004). Yield variability over time is a lens into the sensitivity of local conditions to interannual biotic and abiotic stressors. Single-season yield maps are used in conjunction with field scouting or other observations to highlight areas with atypical yields, some of which may be treatable if identified in the growing season (e.g. pest pressure).

Crop yield maps are typically created from harvester yield monitor (YM) records, but satellite and airborne remote sensing (RS) imagery can also be used to estimate yields at high spatial resolutions (Yang, 2015). RS-based yield mapping can make yield maps available for sites and years where YM data is unavailable. When characterizing site-specific growing conditions, data over a longer time period can help separate interannual effects from persistent attributes. RS-based methods can make yield map products available for the first time to sites where no YM-equipped harvesters have been used before.

Many methods for estimating yields from RS imagery use reported yields to fit a statistical model with regressors including vegetation indices (VI), VI-derived products, and climate metrics, among others. However, this class of methods is limited by the properties of the reported yield data such as spatiotemporal resolution and extent, and representativeness of the sample distribution. Substituting reported yields from surveys with outputs from crop simulation models (CSM) eliminates the constraints and costs of survey data, greatly enhancing the scalability and transferability of RS-based yield mapping. This study contributes to contemporary work investigating the applications and limits of a satellite-based scalable crop yield mapping (SCYM) algorithm (Lobell et al., 2015).

The SCYM algorithm, described in more detail elsewhere (Lobell et al., 2015), proceeds in four stages: 1) a CSM simulation is run for a range of plausible management and environmental conditions; 2) empirical models of VI response to canopy properties (e.g. leaf area index) estimate pseudo-VI corresponding with each day of the crop simulations; 3) a multivariate linear regression (MLR) model regresses simulated crop yields on pseudo-VI and climate variables;

and 4) yields are predicted out-of-sample by substituting pseudo-VI values with VI from remote sensing imagery.

Other recent studies have tested variants of the SCYM algorithm across several regions. Jin et al. (2017a), showed that maize (*Zea mays* L.) yield predictions were improved when total aboveground biomass was first estimated and then scaled by a fixed harvest index (HI) in stage 3, rather than estimating yield directly with the model. In tandem with others, that study evaluated the sensitivity of the algorithm to the choice of VI (affecting stages 2-4) (Burke and Lobell, 2017; Jin et al., 2017b), and to source of the satellite imagery (Azzari et al., 2017). The algorithm has subsequently been tested at several geographic scales from county to smallholder fields (<0.5 ha) (Burke and Lobell, 2017; Jain et al., 2017).

This study makes several novel contributions by testing the suitability of the SCYM algorithm for PA applications. We validate yield predictions at the pixel-level, rather than at the field-level, using reported yields from a YM aggregated to a 10 m grid. While most SCYM studies, excepting Sibley et al. (2014), have exclusively modeled rainfed yields, both irrigated and rainfed sites were modeled in this study. In contrast to existing work, we incorporate aerial hyperspectral imagery in addition to multispectral imagery.

We replicated some SCYM variations which were shown to perform well elsewhere. In particular, we conduct yield predictions for both 'direct yield' and 'biomass x HI' responses in stage 3 (Jin et al., 2017a; Jin et al., 2017b). Similar to other works, we compare the predictive performance based on several VI. We introduce radiative transfer models (RTM) as a physical modeling alternative to the empirical models use to estimate pseudo-VI from CSM outputs. Finally, we compare the performance of a machine learning model as a substitute for linear regressions in stage 3.

Materials and Methods

Study sites and field data

Datasets from three experimental sites in south eastern Nebraska (Mead, NE), managed by the University of Nebraska–Lincoln's (UNL) Eastern Nebraska Research and Extension Center (ENREC), were used in the study. The region is in Koppen class Dfa (humid continental climate), having cold winters and hot, dry, and humid summers (Peel et al., 2007), and has deep silty clay loams (Suyker and Verma, 2009). Site 1 (44 ha) and Site 2 (45 ha) are adjacent fields irrigated with a center pivot system, while Site 3 (65 ha) is < 5 km from the others and rainfed. In Site 1 maize was grown continuously during the study period (2001-2012) while Sites 2 and 3 rotated between maize and soybeans (*Glycine max* L.). All three fields were managed with conservation tillage and best management practices appropriate to the study region. Seeding rates, irrigation, and herbicide and pesticide applications followed recommended practices (Table 3-1). Each of the three sites hosts sensor equipment for monitoring a wide array of environmental conditions as part of the Ameriflux network (Suyker, 2016). We acquired weather daily records for each site from 2001 to 2013.

Table 3 - 1. Site management information for maize crops during the study site-years.

Site	Year	Variety	Plant population (plants ha^{-1})	Earliest planting date	Earliest harvest date	Median yield (mt ha^{-1})
Site 1, irrigated	2001	Pioneer 33P67	82,000	May 10	October 18	14.29
	2002	Pioneer 33P67	82,000	May 9	November 4	13.24
	2003	Pioneer 33B51	77,000	May 15	October 27	12.32
	2004	Pioneer 33B51	79,800	May 3	October 15	12.45
	2005	Dekalb 63-75 CRW	70,800	May 4	October 13	12.63
	2006	Pioneer 33B53 CRW	82,000	May 5	October 24	11.79
	2008	Pioneer 31N30 HXX	81,500	April 29	November 10	13.10
	2009	Pioneer 32N73	81,500	April 30	November 9	14.65
	2011	DeKalb 65-63 VT3	85,250	May 17	October 26	11.97
	2012	DeKalb 62-97	84,000	April 23	October 10	13.08
Site 2, irrigated	2001	Pioneer 33P67	83,750	May 11	October 22	13.90
	2003	Pioneer 33B51	78,000	May 14	October 23	13.63
	2005	Pioneer 33B51	81,000	May 2	October 17	14.06
	2009	Pioneer 32N72	81,500	April 21	November 10	15.46
	2011	Pioneer 32T88	85,250	May 17	October 26	12.46
	2012	DeKalb 62-97	84,000	April 25	October 9	13.56
Site 3, rainfed	2001	Pioneer 33B51	62,250	May 14	October 29	8.93
	2003	Pioneer 33B51	57,600	May 13	October 13	7.75
	2005	Pioneer 33G68	56,300	April 26	October 17	9.89
	2009	Pioneer 33T57	61,750	April 22	November 11	14.00
	2011	DeKalb 61-69 VT3	56,800	May 2	October 18	9.91

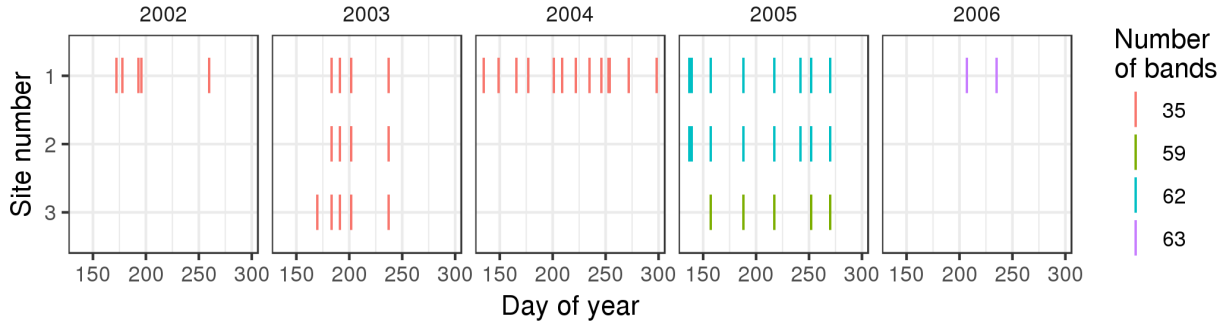
Notes: Yield is the median dry weight yield (15.5% basis) calculated from yield monitor records. We provide the earliest planting and harvest dates, though in some cases field work took multiple days. Refuge plantings with reduced pest control measures were used in a select number of site-years (omitted for brevity), Refuge varieties were of the same maturity class as the primary plantings. Plant population is the average target seeding rate within the field when a refuge variety was planted at a different seeding rate.

Yields were recorded with an Ag Leader PF3000 yield monitor and an Ag Leader 3050 differential global positioning system (GPS). An 8.1 m swath width was used to harvest eight maize rows simultaneously, with a logging interval was one second. YM records were collected by affiliates of the UNL Carbon Sequestration Program (CSP) at ENREC, and then processed by the authors. Yield monitor data was pre-processed using a procedure modified from Simbahan et al. (2004), which was developed and tested on yield monitor data from the UNL study sites. The cleaning process entailed five cases where a point would be removed: 1) data recorded while not harvesting; 2) points at the beginning and end of harvest passes where grain was flowing at a reduced rate; 3) outliers in grain flow, distance traveled, or grain moisture, 4) reported values above or below biologically plausible levels (0 to 22 mt ha^{-1}), 5) co-located points (one random point kept). Unlike the original pre-processing method from Simbahan (2004), we preserved reported yields of zero and local neighborhood yield outliers because we wanted to preserve variation in the yield data and test model performance for predicting aberrant yields.

Remote sensing imagery

Airborne hyperspectral imagery was collected by the staff of the Center for Advanced Land Management Information Technologies at UNL with the SPECIM Aisa sensors (2002-2004: Aisa Classic, 2005-2006: Aisa Eagle) for selected dates between 2002 and 2006. The image collection frequency varied from year to year, with a minimum of two dates per growing season (Figure 3-1). The nominal pixel resolution of the imagery was approximately 1-3 m. The hyperspectral imagery typically includes 35 or 62 bands (varying by vintage) in the 400-990 nm range, a band range suitable for crop characteristic detection (Thenkabail et al., 2002). Image pre-processing was performed with ENVI 4.1 (Exelis Visual Information Solutions) and Caligeo v4b (Gilden Photonics Ltd) software and included the following steps: dark current subtraction, radiometric correction, normalization, and rectification. Atmospheric corrections were made with the Quick Atmospheric Correction (QUAC) algorithm (Bernstein et al., 2005) in ENVI 4.1. All images were collected on cloud-free days. The hyperspectral imagery was resampled to a 10 m grid to a) smooth out image-to-image variations in spatial registration, and b) provide a grid of spatial units large enough to contain multiple YM points over time. The resampling was perfomed by the authors using bilinear interpolation using the R-package 'raster' (Hijmans, 2016).

Figure 3 - 1. Hyperspectral image collection dates by site and year.
Note: Each line denotes the days for which hyperspectral imagery was available for a given site-year. Line color indicates the number of spectral bands provided with that image.



Several spaceborne multispectral sensors are publically available for agricultural applications through the online data stores hosted by the US Geological Survey (USGS) and the National Aeronautics and Space Administration (NASA). Each sensor captures reflected electromagnetic radiation in spectral bands which are defined differently across sensors. The near-infrared band, for example, may include different ranges of the electromagnetic spectrum depending on the sensor. Variation in the band definition may be minor in many cases, but it may affect the value of VI. Whether VI values are equivalent across sensors is relevant to yield mapping because a) some sensors may therefore be more sensitive to crop canopy properties, and thus may be better suited for yield modeling, and b) sensors which measure vegetation similarly could be combined to increase the number of image dates available within a season.

Direct comparison of yield predictions from imagery from each sensor was challenged by mismatches in image timing and spatial resolution. Instead, we use spectral resampling of the hyperspectral imagery (HS) to create synthetic images with spectral properties matching each of four prevalent sensors: Landsat 8 (L8) (Roy et al., 2014), RapidEye (RE) (Planet), Quickbird (QB) (DigitalGlobe), and Sentinal-2 (S2) (Drusch et al., 2012). The spectral resampling procedure integrates reflectances from hyperspectral imagery using the spectral response

functions for each sensor and band. The spectral resample was performed with the R-package 'hsdar' (Lehnert et al., 2015). Synthetic imagery was created for each available hyperspectral image date.

In addition to the hyperspectral imagery, Landsat 7 ETM+ imagery (L7) for years 2001 to 2012 was collected to provide an RS data source for site-years where hyperspectral imagery was unavailable. The L7 imagery also confers the ability to benchmark yield predictive performance against those in other studies where 30 Landsat imagery was employed. The Landsat Ecosystem Disturbance Adaptive Processing System (LEDAPS) (Masek et al., 2013) was used for obtaining atmospherically corrected surface reflections. Cloud-free pixels were identified using the quality flags provided with the imagery.

For each image and sensor we calculate two VI: green chlorophyll vegetation index (GCVI, (Gitelson et al., 2003)),

(Equation 1)

$$GCVI = \frac{\rho_{NIR}}{\rho_{GRN}} - 1$$

with ρ_{NIR} and ρ_{GRN} representing reflectance of near-infrared and green spectral bands, respectively (or bands containing 800 nm and 550 nm); and the MERIS Terrestrial Chlorophyll Index (MTCI, (Dash and Curran, 2004)),

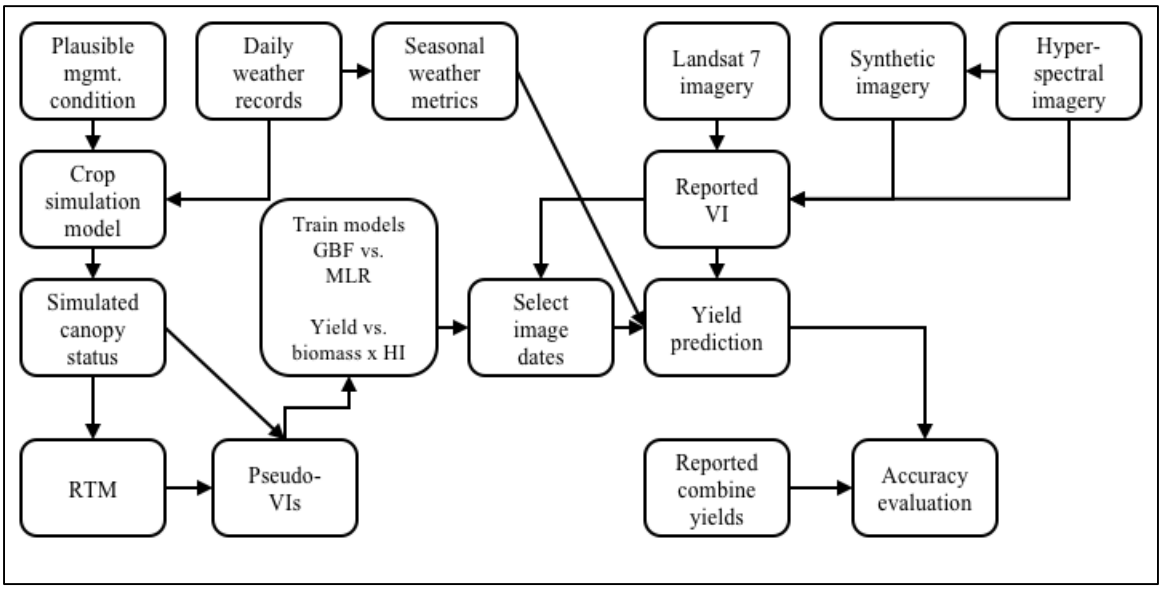
(Equation 2)

$$MTCI = \frac{(\rho_{NIR} - \rho_{RE})}{(\rho_{RE} - \rho_{RED})}$$

where ρ_{RE} is the red edge reflectance and ρ_{RED} is the red band reflectance. The MTCI VI was selected for its reported ability to capture variations in canopy chlorophyll which is related to canopy nitrogen content and, in turn, to yield (Jin et al., 2017a). MTCI was not calculated for synthetic sensors L8 and QB as they both lack a red-edge band. Both MTCI and GCVI have been used in other SCYM applications and were included here to test the replicability of the method across scales and applications.

SCYM models

Figure 3 - 2. Diagram of the modeling workflow.
Note: The left half concerns training a SCYM model to be used with remote sensing inputs depicted in the top right corner. Aggregated weather metrics were used in both training a model and in predicting yields with it.



We simulated maize growth and development with the DSSAT-CERES-Maize (DSSAT 4.6, (Jones et al., 2003)) model. A simulation was run for all combinations of the selected range of management and environmental conditions (Table 3-2). The value ranges for each variable were selected to represent plausible conditions in the region, but no field records were used for calibration. Daily reported weather values were used in the simulations, including minimum and maximum temperature, precipitation, solar radiation, relative humidity, and wind speed. Irrigation applications were not explicitly modeled in the CSM, as precipitation records for the irrigated sites represented the sum of precipitation and irrigation water. Simulations began 100 days before the target planting date to initialize soil moisture and temperature.

The CSM outputs included daily crop growth status measurements and end of season yield. We used simulated leaf and canopy properties to estimate pseudo-VI. Three 'pathways' for converting simulated data into pseudo-VI were tested. All three pathways were used to estimate GCVI and MTCI, but a formula for converting LAI into MTCI was not available (Table 3-3). By linking CSM outputs with an RTM, we estimated pseudo canopy reflectance and then calculated pseudo-VI from reflectances. We employed the widely studied PROSAIL RTM (Jacquemoud et al., 2009). The canopy structure and biochemical variables and formulas used to link the CERES-Maize model are described in Table 3-4.

We used weather observations from each Site not only to perform crop simulations, but also to calculate a set of weather metrics for training statistical models. The selection of climate metrics to capture climate effects on maize yields was consistent other maize yield modeling studies in the Midwest US (Lobell et al., 2015). These variables included: June–August total precipitation ($precip_{JJA}$), June–August mean solar radiation (rad_{JJA}), July mean daytime vapor pressure deficit (vpd_J), and August mean daytime maximum temperature ($tmax_A$).

Table 3 - 2. Crop simulation parameters for fitting statistical models.
Note: All combinations of the selected values were simulated (15,552 total simulations). Three simulated maize varieties were included to represent short, medium, and long season varieties for the region.

Variable	Values	Units	Notes
Year	2002-2013		
Sites	Site 1 (41.1653°, -96.4780°); Site 2 (41.1651°, -96.4689°); Site 3 (41.1787°, -96.4403°)		Ameriflux site identifiers: US-Ne1, US-Ne2, US-Ne3
Fertilizer rate	200, 300	kg of urea N / ha^{-1}	Application one week before planting, and incorporated into soil
Sowing dates	April-24, May-1, May-08, May-15		
Seeding rate	6, 7.5, 9	plants m^{-1}	
Planting depth	5.08	cm	2"
Variety	four varieties of varying maturity length		Descriptions below from DSSAT 4.6 documentation (Hoogenboom et al., 2015)
p1	135, 160, 185	degree days	Thermal time from seedling emergence to the end of the juvenile phase (expressed in degree days above a base temperature of 8 deg.C) during which the plant is not responsive to changes in photoperiod
p2	0.75	days	Extent to which development (expressed as days) is delayed for each hour increase in photoperiod above the longest photoperiod at which development proceeds at a maximum rate (which is considered to be 12.5 hours).
p5	730, 780, 850	degree days	Thermal time from silking to physiological maturity (expressed in degree days above a base temperature of 8 deg.C).
g2	850	kernels $plant^{-1}$	Maximum possible number of kernels per plant.
g3	9.5	mg day^{-1}	Kernel filling rate during the linear grain filling stage and under optimum conditions (mg/day).
phint	38.9	degree days	Phylochron interval; the interval in thermal time (degree days) between successive leaf tip appearances.
Soil types	Estimated soil properties from six nearby locations		Drawn from (HarvestChoice (IFPRI) et al., 2015; Hengl et al., 2014)

Table 3 - 3. CSM to pseudo-VI pathways. Formulas for converting to a pseudo-VI ('VI' column) from a crop simulation model output ('Canopy property').

VI	Canopy property	Formula	Source
GCVI	LAI	$1.4 * LAI^{1.03} + 0.93$	Nguy-Robertson et al. (2012) and Lobell et al. (2015)
GCVI	Canopy nitrogen (CN)	0.8 * CN	Schlemmer et al. (2013) and Jin et al. (2017a)
GCVI	Various	modeled with RTM	-
MTCI	Canopy nitrogen (CN)	0.789 * CN + 3.05	Jin et al. (2017a)
MTCI	Various	modeled with RTM	-

Table 3 - 4. Variables and conversions for linking DSSAT CERES-Maize outputs with the PROSAIL RTM.

Variable	DSSAT variable name	DSSAT unit	PROSAIL variable name	PROSIAL unit	Formula or value	Source
Leaf structure parameter	-	-	N	-	1.5	
$Chlorophyll_{a+b}$	VNAD	kg/ha	Cab	ug/cm2	(((VNAD / 10) - 0.21) / 2.5) * 10	Schlemmer et al.
Carotenoids*	-	-	Car	ug/cm2	2.2828 * exp(0.037474 * (Cab * 0.82))	Efeoglu et al. (2009) and Peneulas et al. (1995)
Equivalent water thickness	-	-	Cw	cm/cm	0.012	
Brown pigment content	-	-	Cbrown	-	0	
Dry matter content	GWAD	kg/ha	Cm	g/cm2	GWAD / 100000	
LAI	LAID	-	LAI	-	LAI	
Soil reflectance parameter	-	-	psoil	-	1.4	
Hot spot parameter	-	-	hspot	-	0.1	
Solar zenith angle	-	-	solar_z	degrees	41	

* $Chlophyll_a$ was first estimated from $Chlorophyll_{a+b}$ and then carotenoid concentration was calculated

Statistical models were constructed for two dependent variables, several alternative independent variables, and two model types. The two response variables tested were maize yield ('yield') and maize aboveground biomass converted to yield with a constant harvest index scalar ('biomass x HI'). Using biomass x HI as the response variable improved model accuracy by reducing dependence on the CSM for accurate energy portioning in some studies (Jin et al., 2017a; Jin et al., 2017b), while the opposite was observed in another (Jeffries et al., 2017). We used each of the five VI pathways as an alternative independent variable in model formulation,

(Equation 3)

$$\begin{aligned}\mathrm{Yield_i} = \beta_0 + \beta_1 \mathrm{irr_i} + \beta_2 \mathrm{VI_{it}} + \beta_3 \mathrm{VI_{it'}} + \beta_4 \mathrm{W_i} + \\ \beta_5(\mathrm{VI_{it}} \times \mathrm{irr_i}) + \beta_6(\mathrm{VI_{it'}} \times \mathrm{irr_i}) + \beta_7(\mathrm{W_i} \times \mathrm{irr_i}) + \\ \beta_8(\mathrm{VI_{it}} \times \mathrm{W_i}) + \beta_9(\mathrm{VI_{it'}} \times \mathrm{W_i}) + \\ \beta_{10}(\mathrm{VI_{it}} \times \mathrm{irr_i} \times \mathrm{W_i}) + \beta_{11}(\mathrm{VI_{it'}} \times \mathrm{irr_i} \times \mathrm{W_i}) + \varepsilon_i\end{aligned}$$

where $Yield_i$ is the maize yield at location i either as response 'yield' or 'biomass x HI', irr_i is a binary variable indicating the presence of irrigation, VI_{it} is the VI value at given day of year, W is a vector of weather variables ($precip_{JJA}$, rad_{JJA}, vpd_J, and $tmax_A$), and ε_i is the error term. We relaxed the constraints of Lobell et al. (2015) on early and late image timing to accommodate irregular hyperspectral image dates, and to allow for the possibility of single date models: $t \geq 140, t' \leq 260, t \leq t'$.

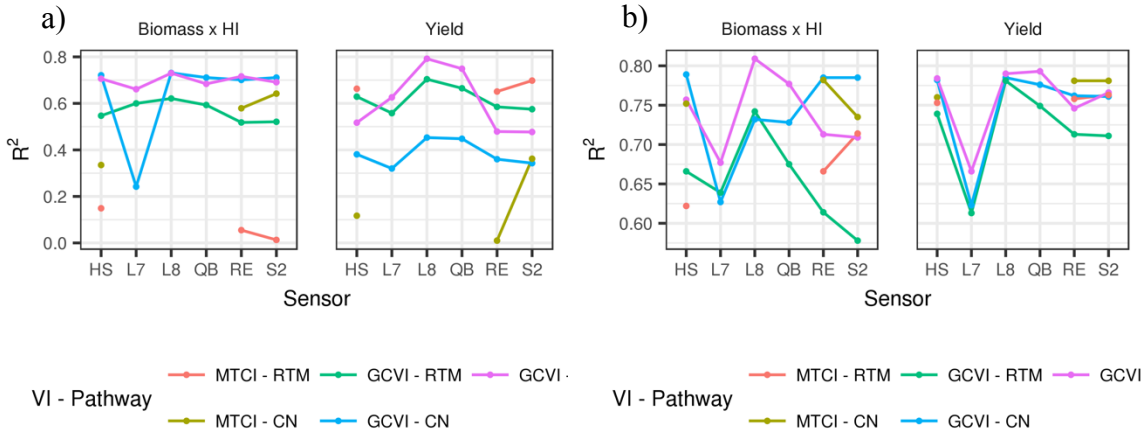
Two statistical model types were used: multiple linear regression (MLR), and gradient boosted forest (GBF). We compare these two approaches because MLR models were predominant in studies using SCYM (Azzari et al., 2017; Jain et al., 2017), but some machine learning methods

have shown to be better in capturing hierarchical relationships and improving prediction quality (Panda et al., 2010). Gradient boosted forest models were selected for their resistance to over-fitting and minimal need for parameter tuning (Friedman, 2001). Variables and interactions in Equation 1, written in terms of the MLR model, were used in the GBF model as well.

We evaluated the utility of SCYM modeling approaches for PA yield mapping by comparing predicted and reported values for several key metrics. A comparison of mean yields over time ($\overline{yield}_i$) demonstrated the ability of the SCYM model to replicate yield maps from YM. Comparison of within-field variability of mean yields ($\overline{yield}_{wfv}$), where *wfv* indicates within-field variability, and the field-level yield standard deviation ($yield_{sd}$) was used to gauge the relevance of the SCYM algorithm for capturing yield spatial variability properties. We documented the performance of several variants of the SCYM algorithm by comparing R, root mean squared error (RMSE), and mean absolute error (MAE). Tukey mean-difference plots, or Bland-Altman plots, were used to inspect biases in the predictions, and identify points which extend beyond the limits of agreement between predicted and YM yields (Bland and Altman, 1999).

Results and Discussion

Figure 3 - 3. Yield predictive performance was sensitive to the choice of imaging sensor. Note: Sub-figure 3-3a shows results from the MLR model, and sub-figure 3-3b the GBF model. R^2 values were calculated for mean yield from predictions and yield monitor records across all sites (10 m resolution). Note that y-axis scales differ between the plots to more clearly depict variability in 3b. Sensors names are abbreviated as: HS, raw hyperspectral image; L7, Landsat 7 ETM+; L8, Landsat 8; QB, Quickbird; RE, RapidEye; S2, Sentinel-2. Line color indicates the vegetation index and canopy-to-VI conversion 'pathway'.



Effect of sensor selection

We assessed the sensitivity of the yield predictive performance to the selection of sensor type. R^2 values were used to compare the amount of yield variability captured by predicted values across all sites with a 10 m resolution. Because the Landsat 7 imagery had a native resolution of 30 m, those images were downscaled with nearest neighbor resampling to 10 m before yield prediction.

Yield model performance was sensitive to the choice of sensors, but the magnitude and direction of the response varied by the response variable, VI, and statistical model (Figure 3-3). L8 most frequency had the best performance, relative to other sensors. L7 was typically among the worst performing sensors.

For the GCVI-based SCYM variants, predictions made with RE and S2 imagery almost always resulted in R^2 values lower than the raw HS imagery. Inversely, RE and S2 outperformed HS in a majority of cases (11 of 16) for MTCI SCYM variants. This performance pattern may be explained in part by how the bands are defined for the RE and S2 sensors. The RE and S2 sensors have dedicated red-edge bands, in addition to a near infrared band, which are designed to capture changes in vegetation related to chlorophyll concentration.

Poor relative performance for L7 was expected due to the coarser spatial resolution imposed by the imagery. In light of that expectation, it was notable that L7 provided yield performance which was comparable to other sensors for the GCVI – RTM and GCVI – LAI pathways in Figure 3-3a.

'Biomass x HI' vs. 'yield'

In two-thirds of the cases examined, the GCVI estimated from LAI resulted in the highest amount of explained yield variations. In the GCVI-LAI case, the 'yield' response variable either explained less variation in reported yields than 'biomass x HI', matched it, or improved it only marginally (mean improvement = 4 percentage points of explained variation). The 'biomass x HI' with GCVI-LAI preformed relatively well and consistently, while the 'yield' response performances had higher maximums that were more volatile across cases. This may be due to the difference in model complexity between the two responses. Unlike the 'yield' response, the 'biomass x HI' response does not depend on the CSM model to accurately represent energy partitioning into plant components including harvestable grain. By instead using a simple scalar (HI), 'biomass x HI' performance was more resilient to cases where simulated crop condition may have been dissimilar to that of the physical crop. Conversely, the 'biomass x HI' response did not benefit from cases where the simulation model was successful in representing yield development.

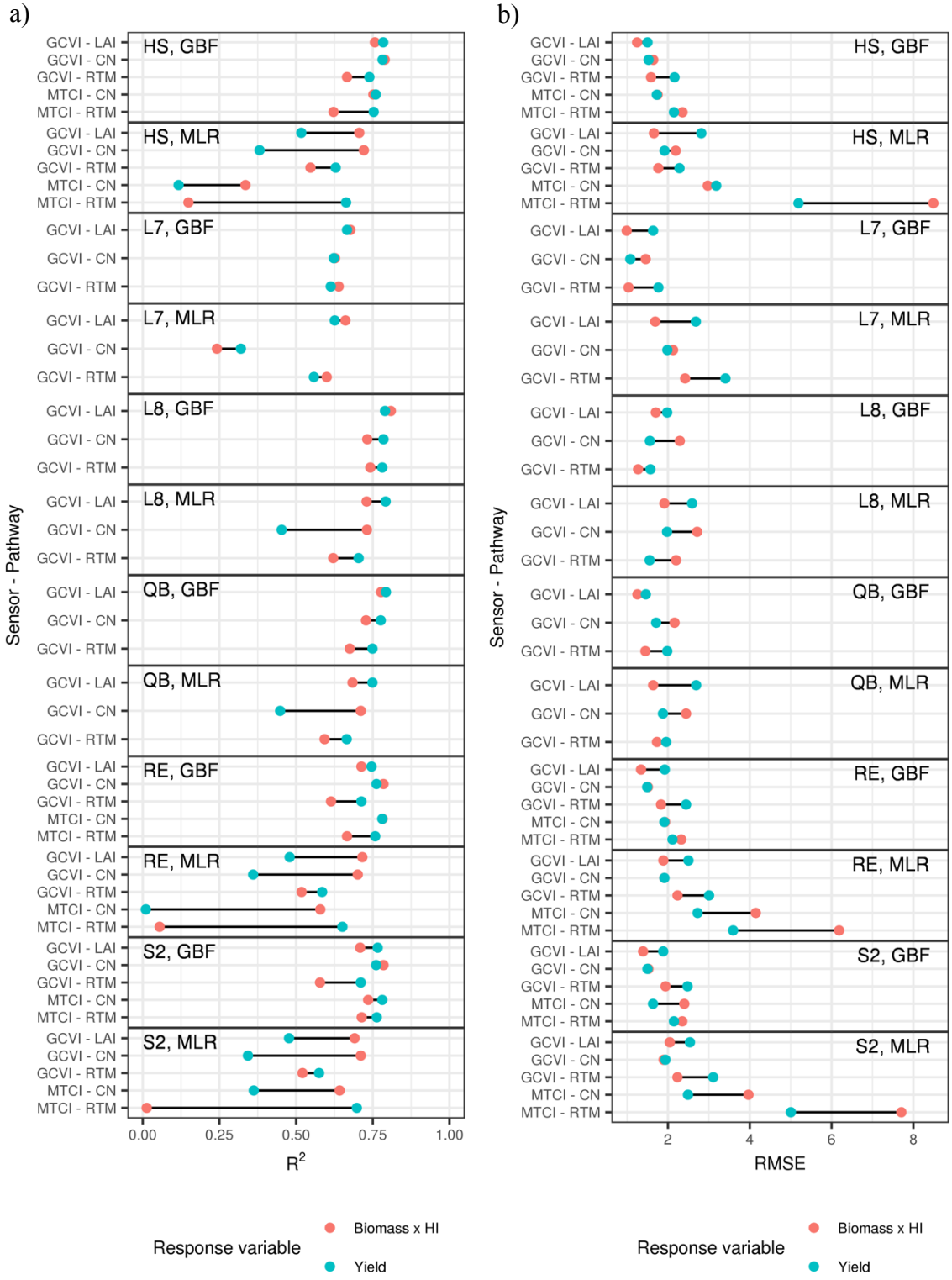
From these results, the 'biomass x HI' approach appears to be the more robust choice for SCYM modeling of those tested. One limitation of estimating yield from biomass and HI was that it was dependent on the selection of an appropriate HI value. This study used a single HI value because of the small spatial extent and the consistency of crop varieties grown across the sites. However, variable HI values may be valuable for applications across large spatial extents, regions with diverse technical levels, or long-time periods. The latter two factors are relevant due to changes in HI through plant breeding over time, and varying in adoption by availability and affordability.

Pathways

We posited that introducing an RTM to model pseudo-VI from simulated canopy conditions would result in more flexible and accurate predictions, relative to predictions using empirical conversion from LAI or CN. In most cases, the results demonstrated that the RTM did not perform as well as the other pathways. The RTM pathway frequently was among the poorest

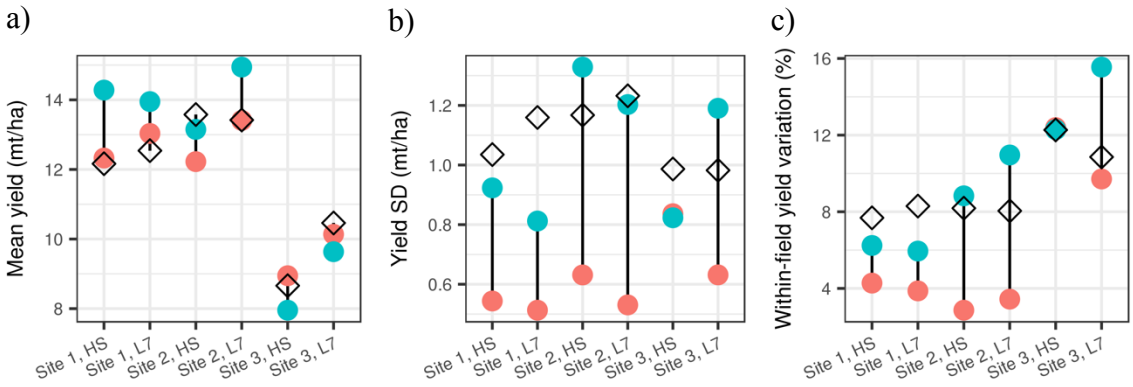
performing methods (Figure 3-3 and 3-4). In 14 of 16 cases, RTM methods explained more yield variation with the ‘yield’ response. Both of the dissenting cases were for L7 imagery. The CN pathway provided typically resulted in better agreement with YM yield means when used in combination with MLR models and the ‘biomass x HI’ response. However, with the GBF model, the CN pathway had mixed results from the two response variables.

Figure 3 - 4. R^2 (4a) and RMSE (4b) of predicted vs. yield monitor mean yields at 10 m resolution across all sites.
Note: Refer to Figure 3-3 description for abbreviation. Labels on each plot indicate the sensor type and statistical model used to make predictions.



GBF vs. MLR

Figure 3 - 5. Comparison of field-level summary statistics for yield monitor yields
Note: yield monitor yields (diamonds) and yield predictions made with the GBF model (orange) and the MLR model (blue). All predictions performed at 10 m resolution with the LAI pathway, GCVI VI, and 'biomass x HI' response. Yield summary statistics include 3-5a) mean yield, 3-5b) standard deviation (SD) of yield, and 3-5c) relative within field yield variation ((1 – (mean yield / 95th percentile yield)) * 100). See Figure 3-3 description for sensor abbreviations.



Predictions from the GBF model had higher R^2 values than their MLR counterparts in nearly all cases (Figure 3-4). The difference in R^2 values between the GBF and MLR models was 0.20 on average, and the difference ranged from 0 to 0.77. GBF performance was less sensitive to sensor, pathway, and response variable than MLR. The standard deviation of R^2 values across all model variants (Figure 3-4) was 0.06 for GBF and 0.21 for MLR.

The better performance of the GBF model relative to MLR was anticipated in light of findings in other comparisons of machine learning and traditional regression models (Panda et al., 2010). The improved yield predictions may result in the machine learning model's ability to identify and model nested variable interactions and hierarchical relationships that were not specified a priori in the MLR model.

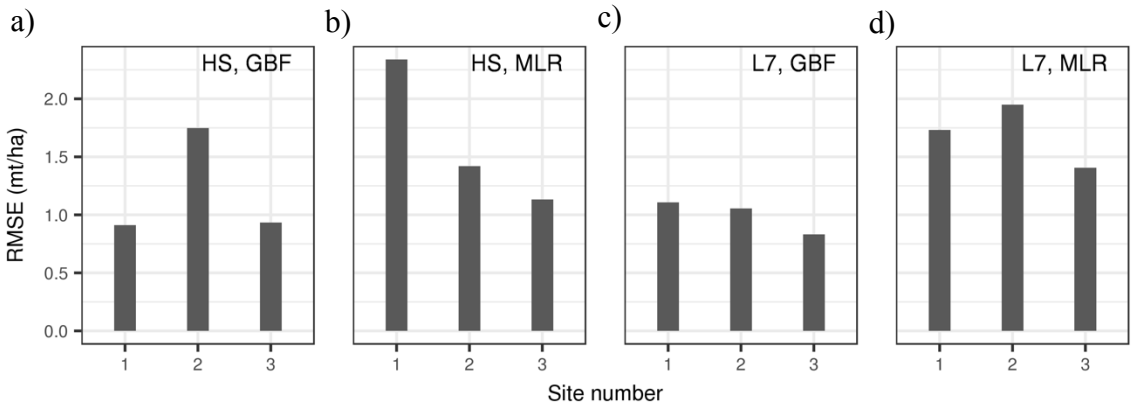
Counterbalancing the GBF SCYM model's strength in predicting mean yields was its tendency to underestimate the variance of the YM yields. The GBF predictions tended to be closer the mean yield for each field (Figure 3-5a), but the standard deviations were typically lower (Figure 3-5b). Poor agreement between yield standard deviations likely resulted in an underestimate of 95th percentile yields, and in turn, a downward bias in the prediction of $\overline{yield}_{wfv}$. Low variance in the predicted values may be explained in part by the character of the crop simulation dataset that was used to train the GBF. If simulated crop conditions do not capture the range of conditions observed in the field, then the GBF models trained with that data not reflect the proper yield response to those conditions.

The MLR model was trained with the same simulation data but exhibits better agreement with the YM yield standard deviation. As a supervised regression model, the GBF should be trained on data which ideally are representative of the mean and variance of the population of interest.

Underestimation of within yield variation, may be explained by inadequate variation in the simulation data used for training the GBF – there were in-field conditions that were not represented by the simulation data. Simulations which include a wider range of soil, weather, and management conditions, potentially refined to match location conditions (Jin et al., 2017b), may result in better performance in future studies. By contrast, the MLR model was able to better capture within field variation, though still imperfectly, because of the linear nature of the model coefficients. Combining the SCYM workflow with machine learning models resulted in more accurate mean yields, but did perform as well as MLR for predicting yield variation. Continued work is needed to determine the potential for machine learning models of various types in the SCYM algorithm.

Rainfed vs. irrigated

Figure 3 - 6. RMSE for predicted vs. reported mean yields across sites.
Note: Sub-figures 3-6a-d depict performance variation across statistical models and sensors, indicated with sub-figure labels. All predictions performed at 10 m resolution with the preferred algorithm. See Figure 3-3 description for abbreviations.

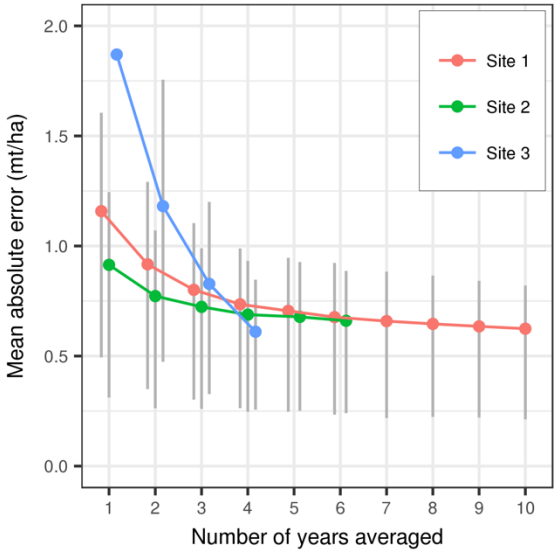


The present study deviated from most earlier SCYM studies by include irrigated, as well as non-irrigated, fields among those modeled. We found that predictions for irrigated fields (Sites 1 and 2) had higher RMSE and lower R^2 values. RMSE values for the rainfed site were lower for the rainfed site than for either irrigated site in the large majority of cases (Figure 3-6). RMSE values for the rainfed site and Site 1 were lower for L7 than HS despite of the 30-10 m downsampling performed on the L7 images. One plausible explanation is that the higher temporal frequency of L7 imagery allowed predictions to be performed with imagery collected on dates where VI were better indicators of yield. However, results from Site 2 possessed the opposite relationship. Prediction accuracy may also be diminished for irrigated sites by: 1) lower VI sensitivity to changes in higher density canopies associated with irrigated management; and 2) distorted surface reflectances caused by irrigation water present on the leaf surface. A larger sample of irrigated and rainfed sites is needed to articulate the best strategies for mapping yields across both types of sites using the SCYM algorithm.

Length of time

Figure 3 - 7. Mean absolute error of yield predictions was reduced by averaging together a greater number of prediction-years.

Note: Predictions show here were created using the preferred algorithm parameters and Landsat 7 imagery downsampled for 10 m resolution. We subset the dataset to include only pixels with data available for all years where maize was grown; ten, six, and four years for Sites 1-3, respectively. For each number of years, one through ten, we sample that many years for each cell and create an average. The reference mean yield was calculated from yield monitor yields across all years available. The vertical grey bar extends between the first and third quartile of the absolute error for pixels within each site.



We found that prediction error decreased as the number of observation years increased (Figure 3-7). For each site, both the mean absolute error (MAE) and the interquartile range of the absolute error were inversely related to the number years used for calculating average yield. Site 3 had a higher marginal response to additional observation years, but the number of years with maize production was limited to four. The marginal response for Sites 1 and 2 decreased over all years, but the rate of decrease stabilized around five years. It was unsurprising that MAE decreased as time-series length increased because averages became less sensitive to interannual variations. The returns to accuracy from longer observation periods found here were consistent to those identified in elsewhere (Asseng et al., 2013). Future work employing a larger panel of site-years is needed to verify the generalizability of the marginal return trends in this study.

Suitability for PA

Table 3 - 5. Linear regression models describing proportional bias in yield predictions. All models, intercepts, and coefficients were statistically significant ($p < 0.001$). Standard errors are provided in parentheses.

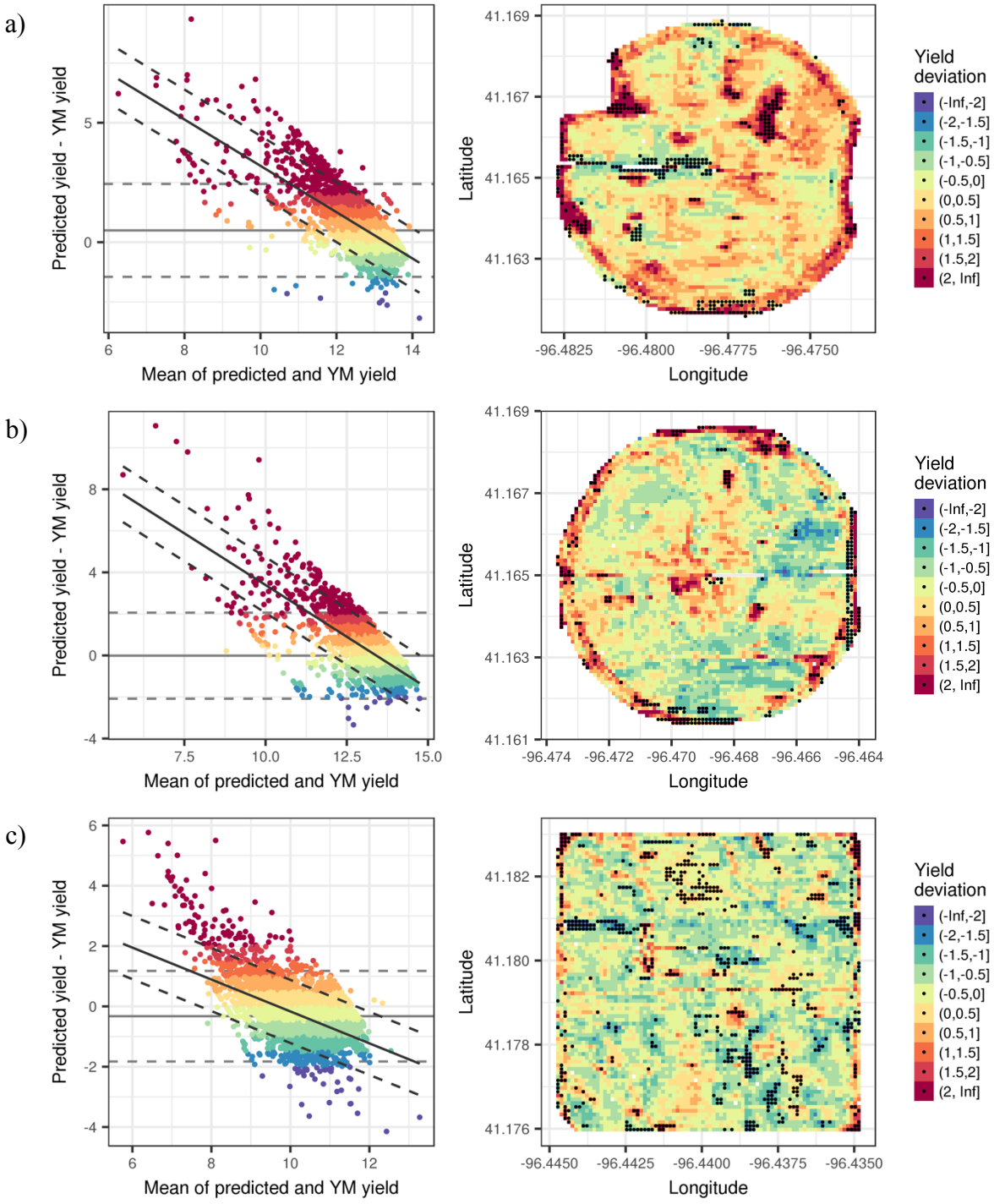
Site	Intercept	Coefficient	n	R^2
1	12.889	-0.969	4,978	0.533
	(0.165)	(0.013)		
2	13.347	-0.996	5,345	0.552
	(0.165)	(0.012)		
3	5.123	-0.529	6,598	0.255
	(0.115)	(0.011)		

We assessed whether the yield predictions were accurate enough for PA yield mapping applications. The Bland-Altman plots with +/-10% limits of agreement suggested that a large majority of 10 m pixel locations in the field were suitable for applications where a 10% error rate was acceptable (Figure 3-8). The level of fixed bias was inconsistent across the three sites; one site was negative, another positive, another near zero. Proportional bias was detected in all three fields, and in each case the bias level was inversely related to mean yield. Linear regression models of the yield difference and mean yields were statistically significant ($p < 0.001$) (Table 3-5).

Mapping the difference between predicted and YM yields suggested that predictions falling outside of the limits of agreement are spatially clustered. In some case the spatial organization can be ascribed to physical features of the field. In Sites 1 and 3 (Figure 3-8a and 3-8c) the location of access roads into the field appear as clusters of points outside of the limits of agreement. Locations near field edges also commonly fell outside of the limits of agreement. Crop which intersect physical features like paths, field edges, test plots may have lower than expected yields because of growing constraints that were not explicitly modeled. For example, crops at field or path edges are more exposed to weather conditions and may have poorer soils. Lower yielding areas were more likely to have higher error, as described by the proportional bias depicted in Figure 3-8. Higher than expected yields may be the result of a) reflectances from soil or non-maize plants being interpreted as the crop; or b) yield-limiting stress events occurred after the images were collected in a season.

We found that linear regression models were successful in capturing variations in error across the range of mean yield values (Table 3-5). Such models may be used to correct for bias in yield predictions. In cases where *in situ* yield data was available for some years, a model of bias may be developed and applied to the yield predictions for after the fact corrections. Whether a generalizable bias correction model exists warrants investigation in a future study.

Figure 3 - 8. Bland-Altman plots with maps of yield deviation (predicted – reported) for sites 1-3, sub-figures 3-8a-c respectively.
Note: Predictions were performed with the preferred model (sensor: HS, model type: GBF, response variable: 'biomass x HI', pathway: GCVI-LAI). Dashed lines indicate the limits of agreement (mean site yield +/- 10%) and solid lines indicate the mean deviation. Proportional bias was detected so both the default limits of agreement (grey) as well as those adjusted with a linear model (black) are shown. Locations falling outside of the adjusted limits of agreement are marked with black points on the maps.



Conclusion

We demonstrated the feasibility of mapping mean maize yields at high spatial resolution (10 m) without *in situ* field records. More than 90% of pixel-level yield predictions were within +/- 10% of yield monitor values. We demonstrated that a machine learning model outperformed linear

regression models for predicting mean yields, and that neither model was able to consistently capture within field yield variability. Predictions at locations with more observation years tended to have better agreement with reported yields.

We contributed to a growing body of technical literature concerned with identifying the best model configurations for the SCYM algorithm. Our findings suggested that predicting biomass and then multiplying by a fixed HI was typically comparable or better than estimating yield directly. The introduction of an RTM for converting simulated crop canopies into pseudo-VI proved was demonstrated to lower yield predictive performance in most cases. A simple empirical approach with the GCVI and simulated LAI frequently worked best. GBF and potentially other machine learning methods have valuable contributions to make in the evolution of SCYM-like yield modeling. We demonstrated that a GBF model was typically better for predicting mean yields than the MLR model, but that the opposite was true for mapping within field yield variability. Variations on the SCYM algorithm promise to provide valuable information about crop health patterns with no *in situ* field records, making it feasible to create high-resolution yield maps for site-specific management.

References

Asseng, S. et al., 2013. Uncertainty in simulating wheat yields under climate change. Nature Climate Change, 3: 827.

Azzari, G., Jain, M. and Lobell, D.B., 2017. Towards fine resolution global maps of crop yields: Testing multiple methods and satellites in three countries. Remote Sensing of Environment.

Bernstein, L.S. et al., 2005. Validation of the QUick atmospheric correction (QUAC) algorithm for VNIR-SWIR multi- and hyperspectral imagery, Defense and Security. SPIE, pp. 11.

Bland, J.M. and Altman, D.G., 1999. Measuring agreement in method comparison studies. Statistical Methods in Medical Research, 8(2): 135-160.

Burke, M. and Lobell, D.B., 2017. Satellite-based assessment of yield variation and its determinants in smallholder African systems. PNAS, 114(9): 2189–2194.

C. Simbahan, G., Dobermann, A. and L. Ping, J., 2004. Screening Yield Monitor Data Improves Grain Yield Maps, 96.

Cassman, K.G., 1999. Ecological intensification of cereal production systems: Yield potential, soil quality, and precision agriculture. Proceedings of the National Academy of Sciences, 96(11): 5952-5959.

Dash, J. and Curran, P.J., 2004. The MERIS terrestrial chlorophyll index. International Journal of Remote Sensing, 25(23): 5403-5413.

Drusch, M. et al., 2012. Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. Remote Sensing of Environment, 120(Supplement C): 25-36.

Efeoğlu, B., Ekmekçi, Y. and Çiçek, N., 2009. Physiological responses of three maize cultivars to drought stress and recovery. South African Journal of Botany, 75(1): 34-42.

Exelis Visual Information Solutions, Boulder, Colorado.

Friedman, J.H., 2001. Greedy function approximation: A gradient boosting machine. Ann. Statist., 29(5): 1189-1232.

Garnett, T. et al., 2013. Sustainable Intensification in Agriculture: Premises and Policies. Science, 341(6141): 33.

Gilden Photonics Ltd, Caligeo v4b, , Glasgow, UK.

Gitelson, A.A. et al., 2003. Remote estimation of leaf area index and green leaf biomass in maize canopies. Geophysical Research Letters, 30(5): n/a-n/a.

Godfray, H.C.J. and Garnett, T., 2014. Food security and sustainable intensification. Philosophical Transactions of the Royal Society B: Biological Sciences, 369(1639).

HarvestChoice (IFPRI), International Research Institute for Climate and Society and Michigan State University, 2015. Global High-Resolution Soil Profile Database for Crop Modeling Applications. Harvard Dataverse.

Hengl, T. et al., 2014. SoilGrids1km — Global Soil Information Based on Automated Mapping. PLOS ONE, 9(8): e105992.

Hijmans, R.J., 2016. raster: Geographic Data Analysis and Modeling (version 2.5-8), pp. R package.

Hoogenboom, G. et al., 2015. Decision Support System for Agrotechnology Transfer (DSSAT) Version 4.6. DSSAT Foundation, Prosser, Washington.

Jacquemoud, S. et al., 2009. PROSPECT+SAIL models: A review of use for vegetation characterization. Remote Sensing of Environment, 113: S56-S66.

Jain, M. et al., 2017. Using satellite data to identify the causes of and potential solutions for yield gaps in India's Wheat Belt. Environmental Research Letters, 12(9).

Jeffries, G.R. et al., 2017. Validation of sub-field maize yield predictions from an algorithm combining crop simulations and remote sensing imagery. The Plant Phenome Journal, in preparation.

Jin, Z., Azzari, G., Burke, M., Aston, S. and Lobell, B.D., 2017a. Mapping Smallholder Yield Heterogeneity at Multiple Scales in Eastern Africa. Remote Sensing, 9(9).

Jin, Z., Azzari, G. and Lobell, D.B., 2017b. Improving the accuracy of satellite-based high-resolution yield estimation: A test of multiple scalable approaches. Agricultural and Forest Meteorology, 247: 207-220.

Jones, J.W. et al., 2003. The DSSAT cropping system model. European Journal of Agronomy, 18(3): 235-265.

Lehnert, L., Meyer, H. and Bendix, J., 2015. Hyperspectral Data Analysis in R - The new hsdar package.

Lobell, D.B., Thau, D., Seifert, C., Engle, E. and Little, B., 2015. A scalable satellite-based crop yield mapper. Remote Sensing of Environment, 164: 324-333.

Masek, J.G. et al., 2013. LEDAPS Calibration, Reflectance, Atmospheric Correction Preprocessing Code, Version 2, ORNL DAAC, Oak Ridge, Tennessee, USA.

Nguy-Robertson, A. et al., 2012. Green Leaf Area Index Estimation in Maize and Soybean: Combining Vegetation Indices to Achieve Maximal Sensitivity. Agronomy Journal, 104: 1336-1347.

Panda, S.S., Ames, D.P. and Panigrahi, S., 2010. Application of Vegetation Indices for Agricultural Crop Yield Prediction Using Neural Network Techniques. Remote Sensing, 2(3).

Peel, M.C., Finlayson, B.L. and McMahon, T.A., 2007. Updated world map of the Köppen-Geiger climate classification. Hydrol. Earth Syst. Sci., 11(5): 1633-1644.

Penuelas, J., Frederic, B. and Filella, I., 1995. Semi-Empirical Indices to Assess Carotenoids/Chlorophyll-a Ratio from Leaf Spectral Reflectance, 31, 221-230 pp.

Roy, D.P. et al., 2014. Landsat-8: Science and product vision for terrestrial global change research. Remote Sensing of Environment, 145(Supplement C): 154-172.

Schepers, A.R. et al., 2004. Appropriateness of management zones for characterizing spatial variability of soil properties and irrigated corn yields across years. Agronomy Journal, 96(1): 195-203.

Schlemmer, M. et al., 2013. Remote estimation of nitrogen and chlorophyll contents in maize at leaf and canopy levels. International Journal of Applied Earth Observation and Geoinformation, 25: 47-54.

Sibley, A.M., Grassini, P., Thomas, N.E., Cassman, K.G. and Lobell*, D.B., 2014. Testing Remote Sensing Approaches for Assessing Yield Variability among Maize Fields. Agronomy Journal, 106: 24-32.

Suyker, A., 2016. AmeriFlux US-Ne1, 2, and 3 Mead, Nebraska, USA, United States.

Suyker, A.E. and Verma, S.B., 2009. Evapotranspiration of irrigated and rainfed maize–soybean cropping systems. Agricultural and Forest Meteorology, 149(3): 443-452.

Thenkabail, P., Smith, R.B. and De Pauw, E., 2002. Evaluation of Narrowband and Broadband Vegetation Indeces for Determining Optimal Hyperspectral Wavebands for Agricultural Crop Characterization, 68, 607-621 pp.

Yang, C., 2015. Hyperspectral Imagery for Mapping Crop Yield for Precision Agriculture. In: B. Park and R. Lu (Editors), Hyperspectral Imaging Technology in Food and Agriculture. Springer New York, New York, NY, pp. 289-304.

Zhang, N., Wang, M. and Wang, N., 2002. Precision agriculture—a worldwide overview. Computers and Electronics in Agriculture, 36(2): 113-132.

Acknowledgements

This work was supported by the National Science Foundation under Grant #0966093, Integrative Graduate Education and Research Traineeship (IGERT) Program on Water Diplomacy at Tufts University. We thank the University of Nebraska – Lincoln's Carbon Sequestration Program for sharing yield monitor data and the Center for Advanced Land Management Information Technologies (CALMIT) for providing AISA hyperspectral imagery.

Chapter 4: Validation of sub-field maize yield predictions from an algorithm combining crop simulations and remote sensing imagery

Prepared in the style of The Plant Phenome Journal

Authors

Graham R. Jeffries
Timothy S. Griffin
David H. Fleisher
Elena N. Naumova
Magaly Koch
Brian D. Wardlow

Author Affiliations

G.R. Jeffries, T.S. Griffin, and E.N. Naumova, Friedman School of Nutrition Science and Policy, Tufts University, 150 Harrison Ave, Boston, MA 02111; D.H. Fleisher, Adaptive Cropping Systems Lab., 10300 Baltimore Ave., Beltsville, MD 20705; M. Koch, Center for Remote Sensing, Boston University, 725 Commonwealth Avenue, Boston, MA 02215; B.D. Wardlow, Center for Advanced Land Mgmt. Info. Tech.; 3310 Holdrege St., Lincoln, NE 68583.

The corresponding author is Graham R. Jeffries (graham.r.jeffries@gmail.com, ORCID ID: https://orcid.org/0000-0001-5333-0976).

Core Ideas

Crop yield maps created from imagery may help to scale out HTP research.

Maize yields created from public data agreed with yield monitor data (max. R^2: 0.73).

Predictions of rainfed maize yields outperformed those for irrigated fields.

Yield estimates were not sensitive to sensor choice, of four synthetic sensors tested.

Modeling yield directly rather than via a biomass proxy generally performed better.

Abstract

Crop yield maps from low-cost data sources contribute to high throughput phenotyping (HTP) research by describing how crop yields vary with respect to location-specific growing conditions across and within fields. Satellite imagery may be a potential alternative to in situ data for creating yield maps. This study evaluated the ability of an algorithm using crop simulation models, remote sensing imagery, and no in situ data to predict sub-field maize (*Zea mays* L.) yields. We predicted yields with imagery from an aerial hyperspectral sensor (10 m resolution), Landsat 7 ETM+ (30 m), and four synthetic sensors which emulate the spectral response of common satellite sensors using hyperspectral imagery (10 m). We validated the predictions with

combine yield monitor records for three sites near Mead, NE. Yield models explained 31.4% of variation on average for single site-years for a rainfed site, and 16.8% for irrigated sites. Across all years and sites, the models explained up to 64.2% and 8.2% of yield variation for rainfed and irrigated sites, respectively. Estimating yield directly, rather than via biomass, provided better results in the majority of cases. Yield performance varied across sensors; hyperspectral imagery performed best, and the difference in performance across synthetic sensors was small. Yield modeling algorithms which combine crop simulations and statistical models show promise for studying variation in yield and crop status at high spatial resolutions.

Abbreviations list

CSM, crop simulation model; ENREC, Eastern Nebraska Research and Extension Center; FHTP, field-based high-throughput phenotyping; FPAR, fraction of absorbed photosynthetically active radiation; GCVI, green chlorophyll vegetation index; HS, hyperspectral, L7, Landsat 7 ETM+; L8, Landsat 8; LAI, leaf area index; PA, precision agriculture; QB, Quickbird; RE, RapidEye; RMSE, root mean squared error; RS, remote sensing; SCYM, scalable crop yield mapper; S2, Sentinel-2; VI, vegetation index; YM, yield monitor

Background

Innovations in cropping systems are needed to increase crop yields in anticipation of population growth, dietary shifts, and environmental stressors in the 21st Century (Godfray and Garnett, 2014). Information about the structure of crop yields, their magnitude and variation over space and time, is valuable to stakeholders across scales – from governments concerned about ensuring food security to producers seeking to improve productivity in their fields. Remote sensing (RS) imagery has been successfully used for estimating crop yields for several decades (Doraiswamy et al., 2003; Maas, 1988), and is a lower cost alternative to surveying yields by other means (Burke and Lobell, 2017). Consistent increases in the spatial resolution of satellite multispectral sensors have made it possible to estimate crop yield variation within fields (Backoulou et al., 2015; Mulla, 2013; Thenkabail, 2003).

Yield maps can reveal within-field variations in growing conditions and crop responses to environmental and management factors. Such maps are relevant to the research and implementation of management practices designed to increase crop yields, such as precision agriculture (PA) and field-based high-throughput phenotyping research (FHTP). PA, or site-specific management, combines sensor-based monitoring, data analytics, and variable rate technologies to improve crop productivity by tailoring inputs to match the needs of diverse conditions within a field (Cassman, 1999). RS imagery is commonly used in PA systems to monitor within-season crop performance and can help inform potential management interventions to improve crop health, e.g. nutrient inputs and pest control (Hatfield et al., 2008). Maps of end-of-season yield document the relative performance of crops with respect to their location and management, which can be used by farmers to plan for future seasons. For FHTP research, yield maps and within-season crop health metrics from imagery can capture crop responses to experimental treatments (e.g. (Basso et al., 2016)). If RS-based measures of crop health and yield are of sufficient accuracy and spatial resolution, they could be used as a compliment or even substitute for in situ measurements. By reducing dependence on ground measurements, it may be possible to reduce the cost and increase the spatial scale of FHTP studies (Araus and Cairns, 2014; Gehan and Kellogg, 2017).

Methods for mapping crop yields from RS imagery have continued to evolve as new, higher-resolution sensors are released. Multispectral satellite imagery has been widely used in conjunction with statistical (linear regression) models to estimate crop yields at various scales from region to field (Bolton and Friedl, 2013; Doraiswamy et al., 2003; Sibley et al., 2014). Statistical yield modeling usually entails regressing surveyed yields on RS-based vegetation index (VI) metrics and environmental and management factors. However, statistical yield estimation from RS imagery requires yield data for calibration, which may not be available. Biophysical models of crop growth, or crop simulation models (CSM), can also be used to estimate yields (Resop et al., 2012). However, crop yield mapping with CSMs is limited by the availability of environmental (e.g. multi-horizon soil records) and management (plant spacing, variety selection, input types and rates, etc.) data inputs for model calibration and yield estimation (Rembold et al., 2013).

New yield prediction approaches have emerged in the past two decades which hybridize RS and CSM approaches to reduce or eliminate the need for calibration data. One class of models assimilates RS imagery by 1) approximating CSM inputs, 2) manipulating simulated crop status to match RS imagery metrics, or 3) or correcting CSM results after-the-fact (Delécolle et al., 1992; Ines et al., 2013). A second class of models combining RS and CSM data eliminates the need for detailed CSM input data. The scalable crop yield mapping (SCYM) algorithm uses CSM simulations to fit a regression model relating in-season crop status to end-of-season yield. The model is then used to predict yields by substituting simulated crop status with proxies computed from RS imagery (Lobell et al., 2015). While the SCYM approach has been tested at the field scale across several continents (Burke and Lobell, 2017; Jain et al., 2017; Lobell and Azzari, 2017), and for small field sizes (Jin et al., 2017a), the performance of the method for sub-field yield estimation has not yet been vetted.

Recent SCYM research builds on foundational work for using CSMs to fit statistical models for prediction with RS imagery. Early examples of the approach combine radar and CSM to estimate sugar beet (*Beta vulgaris* L.) biomass (Bouman, 1991; Bouman, 1992). Clevers (1997), showed that CSM-calibrated statistical model performs better than RS-calibrated CSM estimates. The study estimated the fraction of absorbed photosynthetically active radiation (FPAR) and leaf area index (LAI) from RS images and then 1) calibrated a set of sugar beet model parameters (sowing date, relative growth rate, light use efficiency, maximum leaf area), and 2) trained an empirical model of the yield-FPAR relationship. Normalized root mean squared error (NRMSE) for yield was ~ 19% for yields estimated from the calibrated CSM and ~15% for those estimated from the empirical model fitted with CSM outputs). Sehgal et al. (2005) used empirical estimates of LAI from the normalized difference vegetation index (one image date) in conjunction with wheat (*Triticum aestivum* L.) CSM outputs for a combination of management and environmental conditions to predict wheat yield (NRMSE of ~22.1%, R^2: 0.8, biased upward). Sibley et al. (2014) compared three field-level maize (*Zea mays* L.) yield prediction methods: 1) MODIS-based accumulated photosynthetically active radiation, 2) CSM parameter calibration from MODIS and Landsat imagery, and 3) a SCYM-like approach, and demonstrated that the third method performed best.

Refinements to the SCYM algorithm have continued. Lobell et al. (2015) demonstrated improved model performance with the addition of weather metrics, multiple RS image dates, and

interaction effects between VIs and weather metrics. They formalize the approach into four stages as the SCYM method: 1) CSM runs for combination of plausible management and environment conditions; 2) translation of simulated (CSM) crop status variables (LAI) into pseudo-observations of VI; 3) multiple linear regression of simulated yield on pseudo-VI and climate variables; and 4) yield prediction using observed VI in place of pseudo-VI. Subsequent research has shown that the method is applicable for identifying yield gaps, or the difference between potential and actual yields, for multiple crops (Lobell and Azzari, 2017), and for a variety of farm sizes and geographies (Azzari et al., 2017; Burke and Lobell, 2017; Jain et al., 2017), and sensors (Jin et al., 2017a). Jin et al. (2017a; 2017b) refine the SCYM modeling workflow by introducing alternate methods for estimating pseudo-VIs, and an approach for constraining the range of simulated crop management scenarios to improve the relevance of the empirical model fitted to CSM outputs.

The current study evaluated the performance of the SCYM algorithm for predicting sub-field-level maize yields for use in FHTP research. For the first time in a study using SCYM, yield predictions were validated at a sub-field level using combine yield monitor (YM) records. By contrast, earlier work has validated predictions only at the field- or municipal-level, due to data limitations. We tested the stability of prediction performance across several imagery sources: an aerial hyperspectral sensor (HS, 10 m resolution), Landsat 7 ETM+ (L7, 30 m), and four 'synthetic' sensors which emulate the spectral response of common satellite sensors by downgrading hyperspectral imagery (10 m). While other SCYM studies have performed sensitivity analyses with several imagery sources (Azzari et al., 2017; Jin et al., 2017a), the current study was novel in its use of hyperspectral and synthetic imagery. We compared the relative prediction performance attained with the two different spatial resolutions of the L7 and HS imagery. We also investigated the sensitivity of prediction performance to alternative specifications of the SCYM algorithm. Replicating the a comparison found in two recent studies (Jin et al., 2017a; Jin et al., 2017b), we compared the accuracy of yield predictions made by 1) predicting yield directly ('direct'), and 2) predicting above ground biomass and then estimating yield with a constant harvest index ('biomass x HI'). We revisited this comparison in order to test whether the earlier studies' conclusion that the biomass x HI approach generally performed better was generalizable to the case of sub-field yield predictions.

Materials and Methods

Study sites and field data

The field data used for the study were collected at three experimental sites in Nebraska (Mead, NE) at the Eastern Nebraska Research and Extension Center (ENREC, formerly the Agricultural Research and Development Center). Site 1 (44 ha) managed as continuous maize, and Site 2 (45 ha) was an annual rotation of maize and soy (*Glycine max* L.). Site 3 (65 ha) was an annual rotation of maize and soy during the study period, prior to which the site was managed in smaller sites with conventional tillage. Sites 1 and 2 were irrigated with a center pivot system, while Site 3 was rainfed. The climate of the region is Koppen class Dfa (humid continental climate), having cold winters and hot, dry, and humid summers (Peel et al., 2007). Rainfall and temperatures were consistent across the three sites, as they are co-located within a 5 km area. The soils of all three sites are deep silty clay loams, representative of the eastern Nebraska, comprising: Yutan (fine-silty, mixed, superactive, mesic Mollic Hapludalfs), Tomek (fine, smectitic, mesic Pachic

Argialbolls), Filbert (fine, smectitic, mesic Vertic Argialbolls), and Filmore (fine, smectitic, mesic Vertic Argialbolls) (Suyker and Verma, 2009).

Crop management at three sites followed best management practices for conservation tillage production. Regionally representative crop varieties were grown on all sites (Table 4-1). Seeding rates, irrigation, and herbicide and pesticide applications followed standard practices. Heavy hail caused near total crop failure in 2007 and 2010, so records for these years were omitted from the study.

Table 4 - 1. Site management information for maize crops during the study site-years.
Note: Yield is the median dry weight yield (15.5% basis) calculated from yield monitor records. We provide the earliest planting and harvest dates, though in some cases field work took multiple days. Refuge plantings were used in a select number of site-years (omitted for brevity), Refuge varieties were of the same maturity class as the primary plantings. Plant population is the average target seeding rate within the field when a refuge variety was planted at a different seeding rate. Table replicated from Jeffries et al. (2017).

Site	Year	Variety	Plant population (plants ha^{-1})	Earliest planting date	Earliest harvest date	Median yield (mt ha^{-1})
Site 1, irrigated	2001	Pioneer 33P67	82,000	May 10	October 18	14.29
	2002	Pioneer 33P67	82,000	May 9	November 4	13.24
	2003	Pioneer 33B51	77,000	May 15	October 27	12.32
	2004	Pioneer 33B51	79,800	May 3	October 15	12.45
	2005	Dekalb 63-75 CRW	70,800	May 4	October 13	12.63
	2006	Pioneer 33B53 CRW	82,000	May 5	October 24	11.79
	2008	Pioneer 31N30 HXX	81,500	April 29	November 10	13.10
	2009	Pioneer 32N73	81,500	April 30	November 9	14.65
	2011	DeKalb 65-63 VT3	85,250	May 17	October 26	11.97
	2012	DeKalb 62-97	84,000	April 23	October 10	13.08
Site 2, irrigated	2001	Pioneer 33P67	83,750	May 11	October 22	13.90
	2003	Pioneer 33B51	78,000	May 14	October 23	13.63
	2005	Pioneer 33B51	81,000	May 2	October 17	14.06
	2009	Pioneer 32N72	81,500	April 21	November 10	15.46
	2011	Pioneer 32T88	85,250	May 17	October 26	12.46
	2012	DeKalb 62-97	84,000	April 25	October 9	13.56
Site 3, rainfed	2001	Pioneer 33B51	62,250	May 14	October 29	8.93
	2003	Pioneer 33B51	57,600	May 13	October 13	7.75
	2005	Pioneer 33G68	56,300	April 26	October 17	9.89
	2009	Pioneer 33T57	61,750	April 22	November 11	14.00
	2011	DeKalb 61-69 VT3	56,800	May 2	October 18	9.91

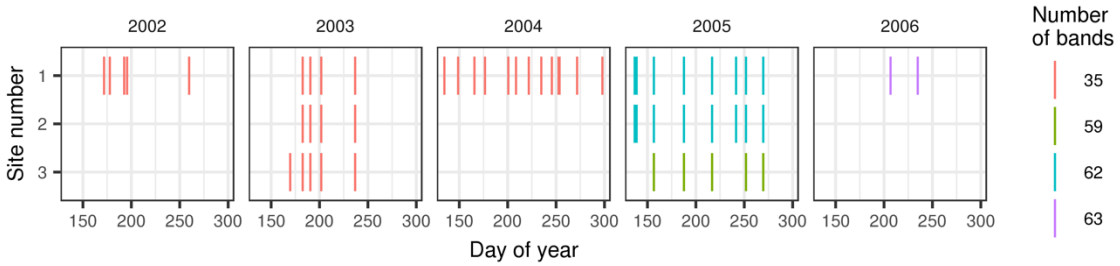
Yields were recorded with an Ag Leader PF3000 yield monitor and an Ag Leader 3050 differential global positioning system (DGPS) by staff at ENREC. An 8.1 meter swath width was used to harvest eight rows a time with a logging interval was one second. Crop yields were calculated by the authors from grain flow rate, grain moisture, and combine speed. Yield monitor data was then edited by the authors to remove misleading data using a procedure modified from that presented in Simbahan et al. (2004). The editing process included five cases where a YM point would be removed: 1) data recorded while not harvesting; 2) points at the beginning and end of harvest passes where grain is flowing at a reduced rate; 3) outliers in grain flow, distance

traveled, or grain moisture, 4) reported values outside of biologically plausible levels (0 to 22 mt ha^{-1}), and 5) co-located points (one random point kept).

Imagery

Airborne hyperspectral imagery was collected by the staff of the Center for Advanced Land Management Information Technologies at the University of Nebraska–Lincoln (UNL) with the SPECIM Aisa sensors (2002-2004: Aisa Classic, 2005-2006: Aisa Eagle) for selected dates between 2002 and 2006. The image collection frequency varies from year to year, with a minimum of two dates per growing season (Figure 4-1). The nominal pixel resolution of the imagery is approximately 1-1.5 meters. The hyperspectral imagery typically includes 35 to 63 bands (varying by vintage and location) between in the 400 to 990 nm range, a band range suitable for crop characteristic detection (Thenkabail et al., 2002). Image pre-processing was performed with ENVI 4.1 (Exelis Visual Information Solutions) and Caligeo v4b (Gilden Photonics Ltd) software and included the following steps: dark current subtraction, radiometric correction, normalization, and rectification. Atmospheric corrections were made with the Quick Atmospheric Correction (QUAC) algorithm (Bernstein et al., 2005) in ENVI. All images were collected on cloud-free days. ENREC staff and researchers collected the imagery and performed pre-processing. All further processing and analysis was performed by the authors. Each image was coregistered in using road intersections (U. S. Census Bureau, 2017) and static landmarks for reference (minimum of 12 points per field). The imagery was resampled to two spatial resolutions (10 and 30 m) with bilinear interpolation using the R 'raster' package (Hijmans, 2016).

Figure 4 - 1. Hyperspectral image collection dates by site and year.
Each line denotes the days for which hyperspectral imagery is available for a given site-year. Line color indicates the number of spectral bands provided with that image. Records for Sites 2 and 3 in years 2002, 2004, and 2006 are absent because no maize was cultivated. Figure replicated from Jeffries et al. (2017).



Combining images from multiple satellite data sources may improve the accuracy of crop yield predictions by increasing the chance that a cloud-free image is available during key stages of crop development. However, due to differences in how spectral bands are defined for each sensor, VI values for each sensor may not agree. In order to characterize the effect of sensor spectral properties on yield prediction performance, we used the hyperspectral (HS) imagery to emulate several prevalent multispectral sensors, creating 'synthetic' imagery for each sensor. Two key advantages of comparing sensor performance from synthetic imagery rather than actual imagery include: 1) synthetic imagery can be created for time periods prior to the launch of the actual sensor; and 2) images for analysis are not restricted to dates where all actual sensors

provided an image. We created synthetic imagery for Landsat 8 (L8) (Roy et al., 2014), RapidEye (RE), Quickbird (QB) (DigitalGlobe, 2001) and Sentinal-2 (S2) (Drusch et al., 2012) with the spectral resampling using the R package 'hsdar' (Lehnert et al., 2015).

We ingested Landsat 7 ETM+ imagery (L7) for years 2001 to 2012 to complement HS imagery. L7 imagery was available for years when where YM yield data was available but HS imagery was not. By including L7 imagery, the study also aimed to provide a benchmark for comparison with other studies which have used L7. Because the L7 imagery was of coarser resolution (30 m), we predicted and validated yield at both 10 m and 30 m resolutions. The Landsat Ecosystem Disturbance Adaptive Processing System (Masek et al., 2013) was used for obtaining atmospherically corrected surface reflections from L7. Cloud-free pixels were identified and removed using the cloud classification flags packaged with the L7 dataset.

The used the green chlorophyll vegetation index (GCVI, (Gitelson et al., 2003)), calculated for each image and sensor, to proxy canopy density and yield-related leaf properties. We chose GCVI over other VIs to enable comparison with other SCYM studies which have used the same index (Jin et al., 2017a; Lobell et al., 2015). GCVI is calculated as

$$GCVI = \rho NIR / \rho GRN - 1 \qquad \text{(Equation 4)}$$

with ρ_{NIR} and ρ_{GRN} representing reflectance of near-infrared and green spectral bands, respectively (or bands containing 800 nm and 550 nm).

Crop simulations

CSM data outputs were used to build a model relating yield to in-season crop conditions. The DSSAT CERES-Maize model (DSSAT 4.6, (Jones et al., 2003)) simulated maize growth and end-of-season yield for a range of environmental and management conditions (Table 4-2). The value ranges for each variable were selected to represent plausible conditions in the region. Weather data for each site-year was drawn from observation stations at each site. The observation stations were managed by ENREC, and accessed through the Ameriflux network (Suyker, 2016). Irrigation applications at the two irrigated sites were reported in aggregate with precipitation, so we did not explicitly model irrigation in the CSM. The start date of each simulation was fixed as 100 days before the target planting date in order to initialize soil moisture and temperature.

Table 4 - 2. Crop simulation parameters for fitting statistical models.
Note: All combinations of the selected values were simulated (15,552 total simulations). Three simulated maize varieties were included to represent short, medium, and long season varieties for the region. Table replicated from Jeffries et al. (2017).

Variable	Values	Units	Notes
Year	2002-2013		
Sites	Site 1 (41.1653°, -96.4780°); Site 2 (41.1651°, -96.4689°); Site 3 (41.1787°, -96.4403°)		Ameriflux site identifiers: US-Ne1, US-Ne2, US-Ne3
Fertilizer rate	200, 300	kg of urea N / ha^{-1}	Application one week before planting, and incorporated into soil

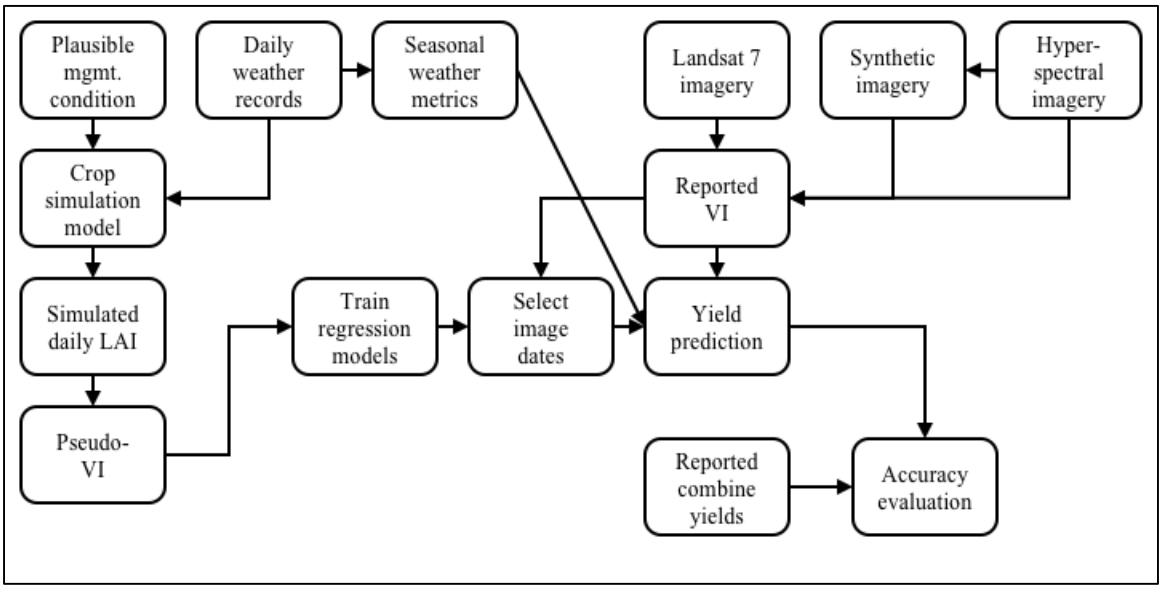
Sowing dates	April-24, May-1, May-08, May-15		
Seeding rate	6, 7.5, 9	plants m^{-1}	
Planting depth	5.08	cm	2"
Variety	four varieties of varying maturity length		Descriptions below from DSSAT 4.6 documentation (Hoogenboom et al., 2015)
p1	135, 160, 185	degree days	Thermal time from seedling emergence to the end of the juvenile phase (expressed in degree days above a base temperature of 8 deg.C) during which the plant is not responsive to changes in photoperiod
p2	0.75	days	Extent to which development (expressed as days) is delayed for each hour increase in photoperiod above the longest photoperiod at which development proceeds at a maximum rate (which is considered to be 12.5 hours).
p5	730, 780, 850	degree days	Thermal time from silking to physiological maturity (expressed in degree days above a base temperature of 8 deg.C).
g2	850	kernels $plant^{-1}$	Maximum possible number of kernels per plant.
g3	9.5	mg day^{-1}	Kernel filling rate during the linear grain filling stage and under optimum conditions (mg/day).
phint	38.9	degree days	Phylochron interval; the interval in thermal time (degree days) between successive leaf tip appearances.
Soil types	Estimated soil properties from six nearby locations		Drawn from (HarvestChoice (IFPRI) et al., 2015; Hengl et al., 2014)

Drawing on Lobell et al. (2015) we estimated pseudo-GCVI for each simulated day of crop growth. By creating a common unit between CSM outputs and VIs from RS imagery, pseudo-VIs can be substituted with imagery VI values in order to predict crop yields at the pixel-level. Pseudo-GCVI was calculated from simulated LAI according to the relationship documented by Nguy-Robertson et al. (2012) for maize:

$$GCVI = 1.4 * LAI^{1.03} + 0.93 \quad \text{(Equation 5)}$$

Statistical models

Figure 4 - 2. Diagram of SCYM algorithm showing the data inputs, and modeling and evaluation stages.



Statistical models were constructed from CSM data outputs and were then used to predict yields by substituting CSM pseudo-VI with VI values from imagery. This study compared the performance of yield predictions for two regression model specifications that other case studies have shown to possess diverging accuracy (Jin et al., 2017a). In the first model, the left-hand response variable was simulated yields, the 'direct' model. In the second, the response variable was simulated total above ground biomass at the time of harvest which was then converted to yield with a constant harvest index scalar (HI, HI = 0.5), the 'biomass x HI' model.

We replicated weather variables used in Lobell et al. (2015), which itself drew on earlier work relating crop yield loss in the Midwest US to climate variables (Lobell et al., 2013; Lobell et al., 2014). The weather variables were calculated using daily weather records from each Site's observation station. The variables included: June–August total precipitation (precip_{JJA}), June–August mean solar radiation (rad_{JJA}), July mean daytime vapor pressure deficit (vpd_{J}), and August mean daytime maximum temperature (tmax_{A}).

As in Lobell et al. (2015), we regressed simulated maize yield or biomass on pseudo-VIs for each pair of two dates in the growing season (t = early and t' = late), climate variables, and interactions between each pseudo-VI observation and each climate variable. We relaxed Lobell et al.'s (2015) constraints on early and late image timing to accommodate irregular hyperspectral image dates, and to allow for the possibility of single date models: $t \geq 140, t' \leq 260, t \leq t'$. We extended the formula from earlier studies to include a dummy variable for irrigation. A binary indicator of the presence of irrigation was known from Site management records in this study, but is readily available at large scales from agricultural land cover classification datasets.

(Equation 6)

$$\text{Response}_i = \beta_0 + \beta_1 \text{irr}_i + \beta_2 \text{VI}_{it} + \beta_3 \text{VI}_{it'} + \beta_4 \text{W}_i + \beta_5(\text{VI}_{it} \times \text{irr}_i) + \beta_6(\text{VI}_{it'} \times \text{irr}_i) + \beta_7(\text{W}_i \times \text{irr}_i) + \beta_8(\text{VI}_{it} \times \text{W}_i) + \beta_9(\text{VI}_{it'} \times \text{W}_i) + \beta_{10}(\text{VI}_{it} \times \text{irr}_i \times \text{W}_i) + \beta_{11}(\text{VI}_{it'} \times \text{irr}_i \times \text{W}_i) + \varepsilon_i$$

where *Response*$_i$ is the maize response at location *i*, either as yield or biomass for models 'direct' and 'biomass x HI', *irr*$_i$ is a binary variable indicating the presence of irrigation, *VI*$_{it}$ is the pseudo- or observed-VI value at given day of year, *W* is a vector of weather variables (precip_{JJA}, rad_{JJA}, vpd_J, and tmax_A), and ε_i is the error term.

After fitting a set of statistical models for each pair of dates and each response variable, yield predictions were created from RS imagery. Because many images were available for each pixel-year, we applied a procedure for selecting the optimal pair of dates. We filtered the statistical model set to include only those where the dates of early and late pseudo-VIs had corresponding RS imagery. We then selected the model which the highest R^2 value. Yield was predicted using the selected model and imagery collected on the dates specified by the model. The procedure was applied for each combination of the six sensors, two response variables, and two spatial resolutions (24 yield prediction sets).

Results and Discussion

Figure 4 - 3. Example maize yield values (reported, predicted, and deviation) for Site 3 in 2005.
Note: From left to right: maize yield reported by yield monitor; maize yields predicted with a SCYM model using raw hyperspectral imagery (HS sensor); deviation of predicted from reported maize yields.

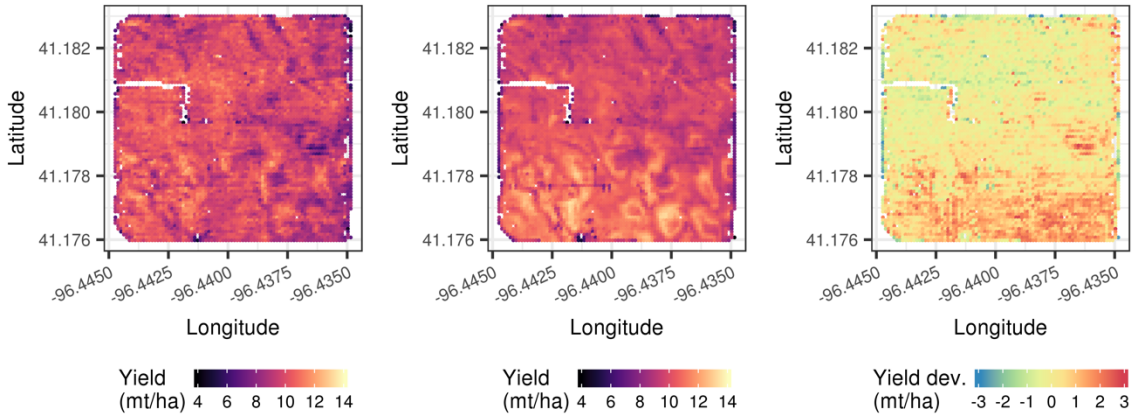
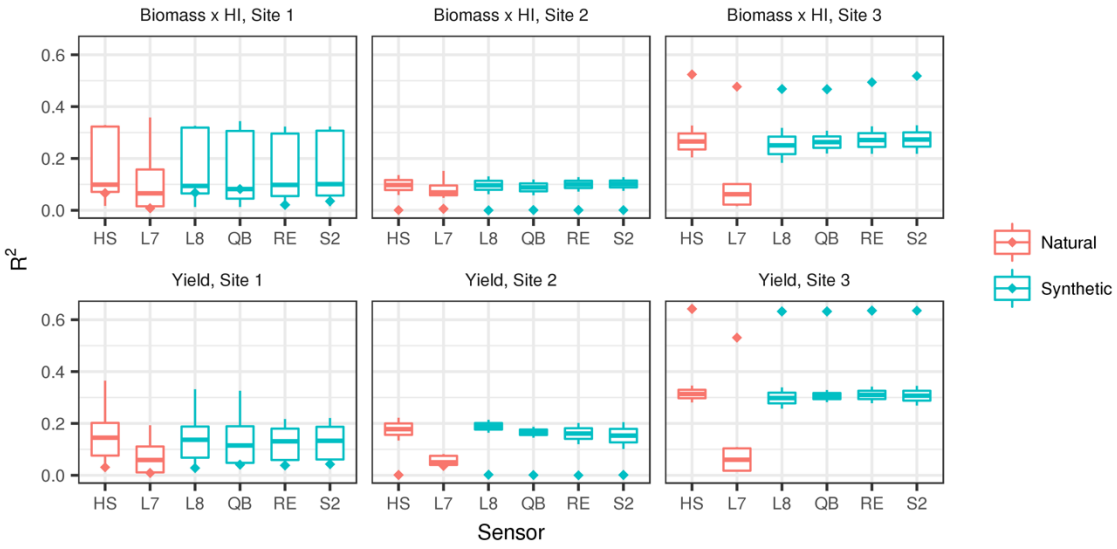


Figure 4 - 4. SCYM model prediction performance by response and sensor.
Note: Sensor abbreviations include: HS, raw hyperspectral image; L7, Landsat 7 ETM+; L8, Landsat 8; QB, Quickbird; RE, RapidEye; S2, Sentinel-2. 'Synthetic' sensors derived from the HS imagery are distinguished with color from 'natural' sensors, or imagery directly from the sensors. Boxplot rectangles indicate the interquartile range and median (thicker line) of site-year R^2 values calculated for a sample of locations (5,000 per site-year). Whiskers indicate outlying values up to 1.5 times the 25th and 75th percentile. Values beyond that range are not shown. Diamonds show the R^2 value of values at sampled locations across all years (n = 45,000 HS and synthetic, and n = 108,000 for L7).



Yield vs. Biomass x HI

In a majority of cases, direct predictions of yield outperformed those from the 'biomass x HI' models (Figure 4). The average variation in yield explained by the model (R^2) for site-years (HS sensor) was 16.8% for 'yield' and 14.8% for 'biomass x HI' for irrigated sites. Similarly, for the rainfed site, 31.4% and 26.6% of yield variation was captured by 'yield' and 'biomass x HI' models respectively. In a limited number of instances 'biomass x HI' models had nominally better performance for certain site and sensor combinations, but otherwise the finding was consistent across sensors at the site-year level. Across all years, the direct 'yield' model explained more point-level yield variation in the rainfed site. Meanwhile, the 'biomass x HI' model performed slightly better for irrigated sites (R^2 0.034 vs. 0.016 for HS sensor), though both performed poorly.

The finding that direct predictions were typically most performant is contrary to the results presented in another SCYM study (Jin et al., 2017b). SCYM models in that study, which incorporated CERES-Maize and two other simulation models, produced predictions of county-level maize yields with higher R^2 and lower RMSE values with 'biomass x HI'. SCYM models from one of the three simulation models (Hybrid-Maize) showed comparable performance across 'yield' and 'biomass x HI' approaches. Differences in spatial scale and extent (three site, point-level vs. multi-state, county-level) may explain contrasting findings. Jin et al. (2017a) performed a similar comparison of the approaches for maize yields in Eastern Africa at level of smallholder fields and sub-county administrative units. Only at coarser spatial scales was the performance better in a majority of cases for the 'biomass x HI' approach.

Rainfed vs. irrigated

Yield predictions for the rainfed site (Site 3) outperformed those for the two irrigated sites (Site 2 and 3) in nearly every case (Figure 4-4). Yields from L7 imagery was the exception – at the level of site-years, R^2 values were comparable across sites, though the rainfed site was more performant across all years. For Site 3, yield predictions with

SCYM attained good performance (2003: R^2 = 0.204, RMSE = 135; 2005: R^2 = 0.327, RMSE = 1.28). The SCYM algorithm explained less variation than some other studies which, unlike the current study, used field data to model yields. One study presented a spatially-adjusted statistical model calibrated with yield monitor data predicted YM maize yields for one rainfed site-year (R^2 = 0.44, RMSE = 1.53) (Peralta et al., 2016), while another study employing YM data and remote sensing captured 63-83% of the yield variability (Yang and Everitt, 2002).

Lower performance of irrigated yields may be explained by RS-based predictions at least two reasons. First, irrigation may bolster against environmental stresses in ways which are not apparent in VIs. For example, water stress during reproductive stages may harm grain filling rates and thus yields, but not substantially alter LAI. Second, irrigation systems commonly used for field crops wet the leaf surface, which can alter canopy reflectance and VI values in turn. The wet canopy effect may translate to poor SCYM model performance because wet canopy VI values are used to predict yields using a model fit with dry canopy VI values. In the present study, the environmental stress effect was likely mitigated by the fact that weather records were site-specific and included irrigation water. Visual examination of VIs from the hyperspectral imagery revealed transient high VI sections of irrigated fields, commonly in the shape of a pie-shaped wedge or semi-circle. These anomalies appeared to corroborate the wet canopy theory. Yield predictions from Landsat 7 images were less affected by irrigation because 1) local irrigation effects were smoothed out in coarser resolution pixels, and 2) each site-year typically included pixels from several dates, dispersing the presence of active-irrigation pixels over time. Future work applying SCYM to irrigated fields should further investigate the role of within-field date diversity in prediction accuracy.

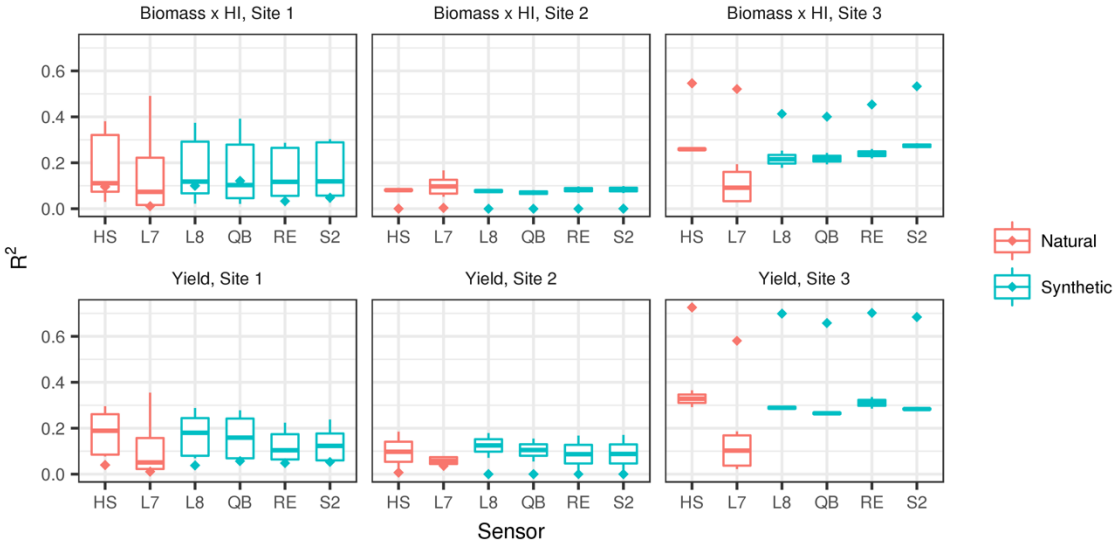
Sensitivity to sensor selection

We predicted yields from VI values calculated with five sensor variants of the hyperspectral imagery: the raw hyperspectral imagery (HS), Landsat 8 (L8), Quickbird (QB), RapidEye (RE), and Sentinel-2 (S2). The distribution of R^2 values across each site and year was similar for all of the hyperspectral-derived sensors. In most cases, performance over all years varied less than 0.02 R^2 points for irrigated fields and 0.04 for rainfed. The most extreme case was a 0.06 R^2 point difference between HS and QB at Site 3 for the 'biomass x HI' model. No single sensor was systematically best or worst in all three sites. Variation in model performance across synthetic sensors suggests that a preference hierarchy for selecting sensor platforms may contribute to improving yield maps. Future work should aim to corroborate these findings with concurrent images from the actual sensors. Multi-sensor comparison highlights opportunities for collating and fusing images from multiple sensors to increase image frequency for SCYM modeling (Azzari et al., 2017).

Performance across spatial resolutions

Figure 4 - 5. SCYM model prediction performance by response and sensor, 30-meter aggregation.

Note: See Figure 4-4 for abbreviations, a description of plot components, and sample procedures. A sample of 500 L7 pixels per site-year was drawn, resulting in 4,500 observations for the HS and synthetic overall R^2 value and 10,800 for L7.



We aggregated predicted and reported yields to a 30-meter grid matching that of the L7 imagery in order to evaluate the performance of the L7 imagery at its native resolution (Figure 4-5). Aggregating yields uniformly improved model performance (Figure 4-4 vs. Figure 4-5). While the mean of site-year R^2 values for all three sites remained below 0.13, the yield model explained up to 58.1% of yield variation across all years in the case of the rainfed site. For sensors other than L7, R^2 values ranged from 0.658 to 0.726 for the direct 'yield' model.

Aggregation of yield monitor records to 30-meter grid cells increased the prediction performance of L7, imagery relative to the point-level predictions. As was the case for point-level prediction performance, L7 model performance at 30 m also lagged behind that of other sensors. The larger size of the L7 pixels increased the area of non-agricultural land covers in pixels containing field edges. Mixed land covers alter the VI values, and, in turn, the yield prediction. All three fields have access paths for flux towers and monitoring equipment, research sub-plots, and, for Sites 1 and 2, center pivot points. Visual inspection of L7 imagery revealed that these features tended to lower the VI values of the occupied pixels which likely affected yield estimates. Errant yield predictions from mixed land cover pixels create wide yield prediction as shown in many of the plots in Figure 4-6. Aggregations of the yield monitor points with yield predictions from hyperspectral data did not suffer the same land cover corruptions because points outside of the field boundaries were omitted in the data preparation stage.

The observation that aggregating improves prediction performance is demonstrated in other cases too. Jin et al. (2017a) for example show that larger spatial units improved prediction R^2 values, and also that pixels containing mixed land covers were worse performers than pure-stands of a crop. The substantial difference between prediction skill

from HS imagery (and synthetic sensor derivatives) and the coarser L7 imagery highlights the value of high-resolution imagery for sub-field yield mapping.

Figure 4 - 6. Predicted vs. reported maize yield, aggregated to 30-meter Landsat grid cells.
Note: Predictions made with direct 'yield' model and Landsat 7 ETM+ imagery for a sample of 500 pixels for each site-year. Solid line is a 1:1 reference line.

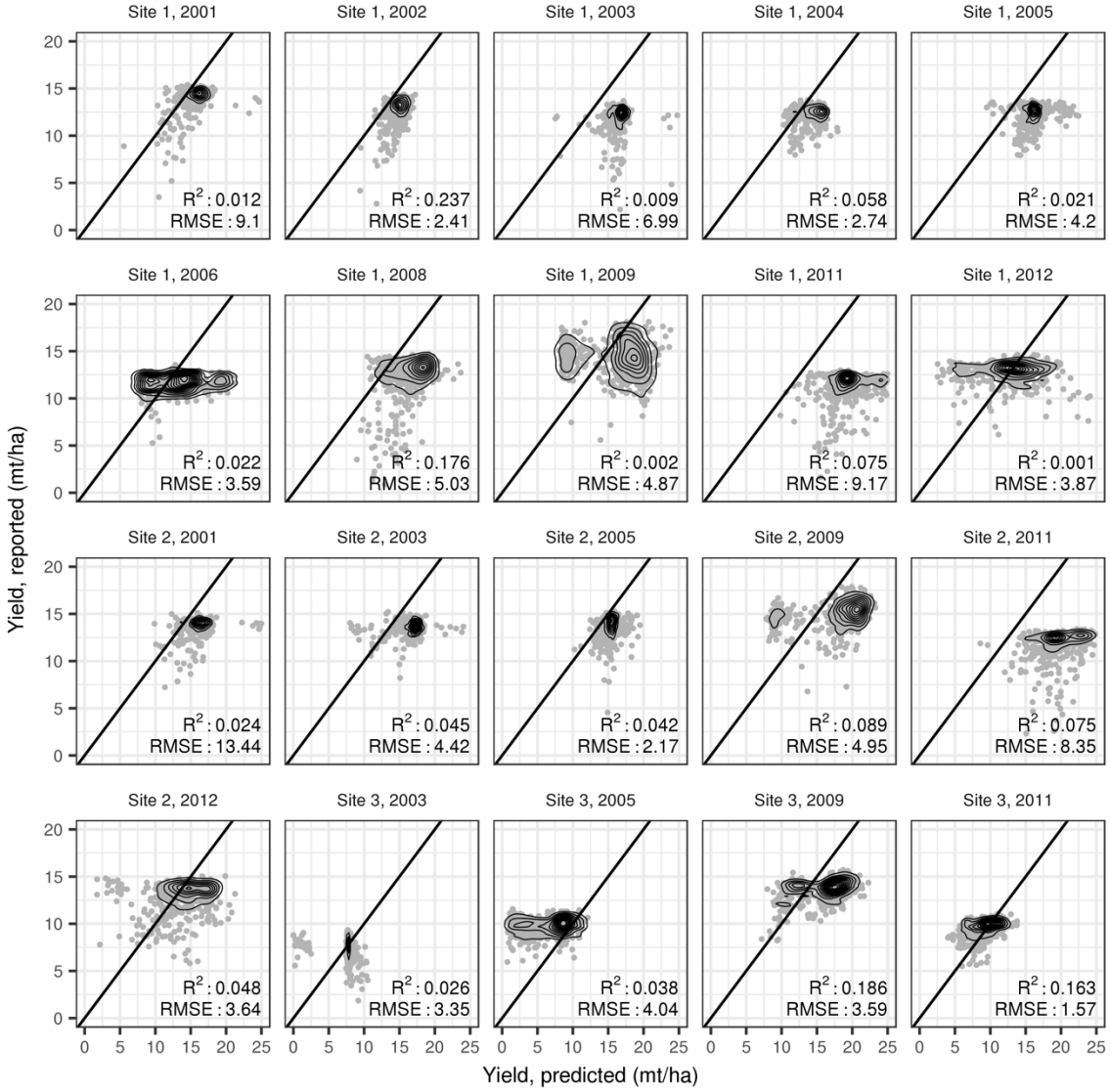


Figure 4 - 7. Predicted versus reported yields by site, all years.
Note: Predictions made with direct 'yield' model and the vegetation indices from the raw hyperspectral images. Solid line is a 1:1 reference line. Sample procedures for the figure

describe in the Figure 4-4 description.

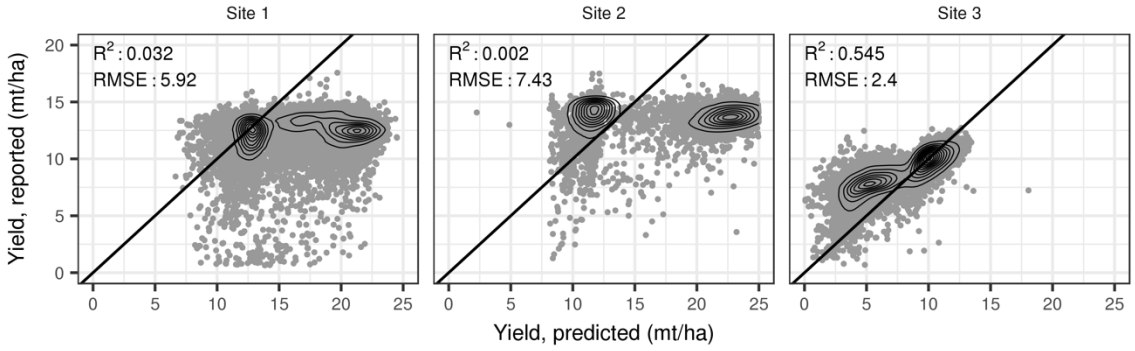
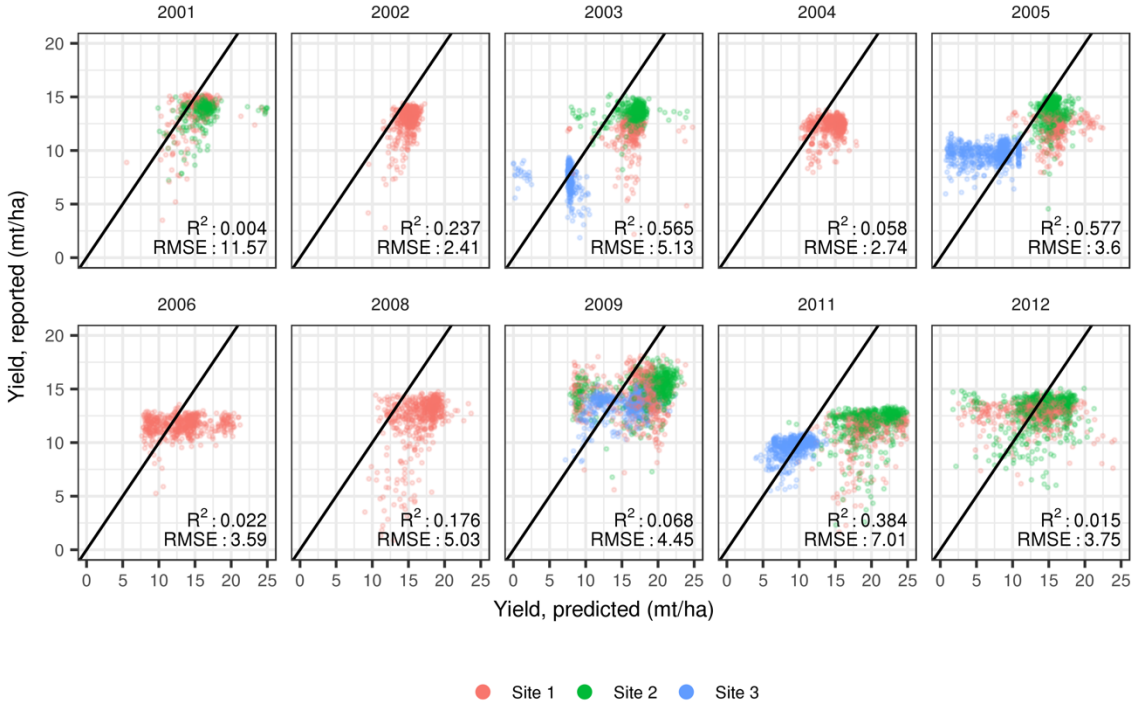


Figure 4 - 8. Predicted versus reported yields by year, aggregated to 30-meter Landsat grid cells.
Note: Predictions made with direct 'yield' model and Landsat 7 ETM+ imagery for a sample of 500 pixels for each site-year. Solid line is a 1:1 reference line.



Performance over time

Yield model performance varied considerably across time, and by site. On whole, Tables 4-6 through 4-8 demonstrate that yield predictions overestimated yields more often than underestimating them, especially for irrigated fields. In some cases, yield predictions were very tightly clustered around the mean and had a low variance relative to the YM reported yields. In other cases, predicted yields possessed appropriate distributions, but were shifted away from the mean. The rainfed field had more consistency across years

than the irrigated fields. In comparison to other studies, the error structure of predicted yields was less consistent across sites and years in this study (Jain et al., 2017; Jin et al., 2017a; Jin et al., 2017b). This may be explained by the fact that other studies coarser spatial units over a wider spatial extent, leading to smoothing of extreme values and possessing more yield heterogeneity due to a greater diversity of cropping environments and management conditions.

Overall performance

Overall performance of the SCYM models was assessed using yield prediction across all sites and years. For the 'yield' response variable with the HS sensor, the model explained 38.7% of the variation (RMSE = 4.498, n = 42,988). QB imagery provided the best overall performance by R^2 (0.437, RMSE = 4.391, n = 47,295) and L8 by RMSE (3.435, R^2 = 0.303, n = 48,033).

Whether or not the performance of the SCYM approaches tested is acceptable will depend on the application and use case. Good performance for the rainfed site overall all years suggests that the approaches tested show potential for location-specific yield qualities (e.g. yields stability and mean yield gaps over time). Maps of mean yields over time can be used by farmers and agronomists to identify the causes of persistent low yields as part of a precision management practice. Whether the performance of SCYM models is sufficiently accurate to be an input dataset for field-based phenotyping studies will depend on the study design and expected yield variation between treatments. In the best cases of site-year point predictions, RMSE values were ~1.25 (mt ha^{-1}). In cases where potentially hundreds or thousands of varieties are planted (White et al., 2012), this error rate may be unacceptably high with the SCYM methods tested here.

Conclusions

This study demonstrated that combing CSM simulations and remote sensing imagery through the scalable crop yield mapper (SCYM) approach has good performance for mapping rainfed maize yield variability. We compared maize yields from a combine yield monitor with yield predictions created with the CERES-Maize CSM, hyperspectral, synthetic, and Landsat 7 imagery, and two variants on the SCYM approach. We tested the hypothesis that predicting yields directly was more performant than predictions of biomass which were then scaled with a fixed harvest index. We found that, in most of the cases tested, direct prediction of yield was more performant. Using synthetic multispectral imagery for four sensors, we found that sensor selection had a nominal effect on prediction performance. Aggregating reported and predicted yields to the 30-meter Landsat pixel grid improved predictive performance. The SCYM approach is scalable to sub-field resolutions while retaining reasonable prediction performance, and, with ongoing improvement, will likely be an applicable technology for HTP research and precision agriculture applications.

References

Araus, J.L. and Cairns, J.E., 2014. Field high-throughput phenotyping: the new crop breeding frontier. Trends in Plant Science, 19(1): 52-61.

Azzari, G., Jain, M. and Lobell, D.B., 2017. Towards fine resolution global maps of crop yields: Testing multiple methods and satellites in three countries. Remote Sensing of Environment.

Backoulou, G.F., Elliott, N.C., Giles, K.L. and Mirik, M., 2015. Processed multispectral imagery differentiates wheat crop stress caused by greenbug from other causes. Computers and Electronics in Agriculture, 115: 34-39.

Basso, B., Fiorentino, C., Cammarano, D. and Schulthess, U., 2016. Variable rate nitrogen fertilizer response in wheat using remote sensing. Precision Agriculture, 17(2): 168-182.

Bernstein, L.S. et al., 2005. Validation of the QUick atmospheric correction (QUAC) algorithm for VNIR-SWIR multi- and hyperspectral imagery, Defense and Security. SPIE, pp. 11.

Bolton, D.K. and Friedl, M.A., 2013. Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. Agricultural and Forest Meteorology, 173: 74-84.

Bouman, B.A.M., 1991. Linking X-band radar backscattering and optical reflectance with crop growth models, Wageningen Agricultural University, 169 pp.

Bouman, B.A.M., 1992. Linking physical remote sensing models with crop growth simulation models, applied for sugar beet. International Journal of Remote Sensing, 13(14): 2565-2581.

Burke, M. and Lobell, D.B., 2017. Satellite-based assessment of yield variation and its determinants in smallholder African systems. PNAS, 114(9): 2189–2194.

C. Simbahan, G., Dobermann, A. and L. Ping, J., 2004. Screening Yield Monitor Data Improves Grain Yield Maps, 96.

Cassman, K.G., 1999. Ecological intensification of cereal production systems: Yield potential, soil quality, and precision agriculture. Proceedings of the National Academy of Sciences, 96(11): 5952-5959.

Clevers, J.G.P.W., 1997. A simplified approach for yield prediction of sugar beet based on optical remote sensing data. Remote Sensing of Environment, 61(2): 221-228.

Delécolle, R., Maas, S.J., Guérif, M. and Baret, F., 1992. Remote sensing and crop production models: present trends. ISPRS Journal of Photogrammetry and Remote Sensing, 47(2): 145-161.

DigitalGlobe, 2001. QuickBird satellite, Longmont, Colorado.

Doraiswamy, P.C., Moulin, S., Cook, P.W. and Stern, A., 2003. Crop Yield Assessment from Remote Sensing. Photogrammetric Engineering and Remote Sensing, 69(6): 665-674.

Drusch, M. et al., 2012. Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. Remote Sensing of Environment, 120(Supplement C): 25-36.

Exelis Visual Information Solutions, Boulder, Colorado.

Gehan, M.A. and Kellogg, E.A., 2017. High-throughput phenotyping. American Journal of Botany, 104(4): 505-508.

Gilden Photonics Ltd, Caligeo v4b, , Glasgow, UK.

Gitelson, A.A. et al., 2003. Remote estimation of leaf area index and green leaf biomass in maize canopies. Geophysical Research Letters, 30(5): n/a-n/a.

Godfray, H.C.J. and Garnett, T., 2014. Food security and sustainable intensification. Philosophical Transactions of the Royal Society B: Biological Sciences, 369(1639).

HarvestChoice (IFPRI), International Research Institute for Climate and Society and Michigan State University, 2015. Global High-Resolution Soil Profile Database for Crop Modeling Applications. Harvard Dataverse.

Hatfield, J.L., Gitelson, A.A., Schepers, J.S. and Walthall, C.L., 2008. Application of Spectral Remote Sensing for Agronomic Decisions. Agronomy Journal, 100: S-117-S-131.

Hengl, T. et al., 2014. SoilGrids1km — Global Soil Information Based on Automated Mapping. PLOS ONE, 9(8): e105992.

Hijmans, R.J., 2016. raster: Geographic Data Analysis and Modeling (version 2.5-8), pp. R package.

Hoogenboom, G. et al., 2015. Decision Support System for Agrotechnology Transfer (DSSAT) Version 4.6. DSSAT Foundation, Prosser, Washington.

Ines, A.V.M., Das, N.N., Hansen, J.W. and Njoku, E.G., 2013. Assimilation of remotely sensed soil moisture and vegetation with a crop simulation model for maize yield prediction. Remote Sensing of Environment, 138: 149-164.

Jain, M. et al., 2017. Using satellite data to identify the causes of and potential solutions for yield gaps in India's Wheat Belt. Environmental Research Letters, 12(9).

Jin, Z., Azzari, G., Burke, M., Aston, S. and Lobell, B.D., 2017a. Mapping Smallholder Yield Heterogeneity at Multiple Scales in Eastern Africa. Remote Sensing, 9(9).

Jin, Z., Azzari, G. and Lobell, D.B., 2017b. Improving the accuracy of satellite-based high-resolution yield estimation: A test of multiple scalable approaches. Agricultural and Forest Meteorology, 247: 207-220.

Jones, J.W. et al., 2003. The DSSAT cropping system model. European Journal of Agronomy, 18(3): 235-265.

Lehnert, L., Meyer, H. and Bendix, J., 2015. Hyperspectral Data Analysis in R - The new hsdar package.

Lobell, D. and Azzari, G., 2017. Satellite detection of rising maize yield heterogeneity in the U.S. Midwest. Environmental Research Letters, 12.

Lobell, D.B. et al., 2013. The critical role of extreme heat for maize production in the United States. Nature Clim. Change, 3(5): 497-501.

Lobell, D.B. et al., 2014. Greater Sensitivity to Drought Accompanies Maize Yield Increase in the U.S. Midwest. Science, 344(6183): 516.

Lobell, D.B., Thau, D., Seifert, C., Engle, E. and Little, B., 2015. A scalable satellite-based crop yield mapper. Remote Sensing of Environment, 164: 324-333.

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

Masek, J.G. et al., 2013. LEDAPS Calibration, Reflectance, Atmospheric Correction Preprocessing Code, Version 2, ORNL DAAC, Oak Ridge, Tennessee, USA.

Mulla, D.J., 2013. Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. Biosystems Engineering, 114(4): 358-371.

Nguy-Robertson, A. et al., 2012. Green Leaf Area Index Estimation in Maize and Soybean: Combining Vegetation Indices to Achieve Maximal Sensitivity. Agronomy Journal, 104: 1336-1347.
Peel, M.C., Finlayson, B.L. and McMahon, T.A., 2007. Updated world map of the Köppen-Geiger climate classification. Hydrol. Earth Syst. Sci., 11(5): 1633-1644.
Peralta, R.N., Assefa, Y., Du, J., Barden, J.C. and Ciampitti, A.I., 2016. Mid-Season High-Resolution Satellite Imagery for Forecasting Site-Specific Corn Yield. Remote Sensing, 8(10).
Rembold, F., Atzberger, C., Savin, I. and Rojas, O., 2013. Using Low Resolution Satellite Imagery for Yield Prediction and Yield Anomaly Detection. Remote Sensing, 5(4).
Resop, J.P., Fleisher, D.H., Wang, Q., Timlin, D.J. and Reddy, V.R., 2012. Combining explanatory crop models with geospatial data for regional analyses of crop yield using field-scale modeling units. Computers and Electronics in Agriculture, 89(Supplement C): 51-61.
Roy, D.P. et al., 2014. Landsat-8: Science and product vision for terrestrial global change research. Remote Sensing of Environment, 145(Supplement C): 154-172.
Sehgal, V.K., Sastri, C.V.S., Kalra, N. and Dadhwal, V.K., 2005. Farm-Level Yield mapping for precision crop management by linking remote sensing inputs and a crop simulation model. Journal of the Indian Society of Remote Sensing, 33(1): 131-136.
Sibley, A.M., Grassini, P., Thomas, N.E., Cassman, K.G. and Lobell*, D.B., 2014. Testing Remote Sensing Approaches for Assessing Yield Variability among Maize Fields. Agronomy Journal, 106: 24-32.
Suyker, A., 2016. AmeriFlux US-Ne1, 2, and 3 Mead, Nebraska, USA, United States.
Suyker, A.E. and Verma, S.B., 2009. Evapotranspiration of irrigated and rainfed maize–soybean cropping systems. Agricultural and Forest Meteorology, 149(3): 443-452.
Thenkabail, P., Smith, R.B. and De Pauw, E., 2002. Evaluation of Narrowband and Broadband Vegetation Indeces for Determining Optimal Hyperspectral Wavebands for Agricultural Crop Characterization, 68, 607-621 pp.
Thenkabail, P.S., 2003. Biophysical and yield information for precision farming from near-real-time and historical Landsat TM images. International Journal of Remote Sensing, 24(14): 2879-2904.
U. S. Census Bureau, 2017. TIGER/Line Shapefiles, All Roads.
White, J.W. et al., 2012. Field-based phenomics for plant genetics research. Field Crops Research, 133(Supplement C): 101-112.
Yang, C. and Everitt, J.H., 2002. Relationships Between Yield Monitor Data and Airborne Multidate Multispectral Digital Imagery for Grain Sorghum. Precision Agriculture, 3(4): 373-388.

Acknowledgements

This work was supported by the National Science Foundation under Grant #0966093, Integrative Graduate Education and Research Traineeship (IGERT) Program on Water Diplomacy at Tufts University. We thank the University of Nebraska – Lincoln's Carbon Sequestration Program for sharing yield monitor data and the Center for Advanced Land

Management Information Technologies (CALMIT) for providing AISA hyperspectral imagery.

Chapter 5: Conclusion

Broad themes of research and results

Remote sensing-based crop monitoring is a significant resource for understanding crop yields and their determinants across spatial scales. The trend toward better quality and frequency of satellite imagery means that crop monitoring research is increasingly limited by how that data is used, rather than by the properties of the data. The research presented in this dissertation was motivated by the need for enhancing crop yield monitoring with innovative algorithms and applications using remote sensing and modeling tools.

Mapping crop yield gaps enables better targeting of interventions for boosting yields, but also provides insight into the drivers of yield loss. It is widely recognized that yield losses can occur from biotic (e.g. pests and disease) and abiotic (e.g. weather, soil quality) stressors, and that managing crops adaptively, in response to experience and expectations, is crucial for guarding yields. Farm structure characteristics (e.g. physical and economic size) may be correlated to certain types of crop stressors, and yield gaps. Objective 1 advanced knowledge about the relationship between farm size (physical and economic) and crop yields and yield gaps in the case of Brazilian soybeans (*Glycine max* L.). The research innovated by synthesizing socio-environmental datasets including remote sensing imagery, property maps, and economic land value. The study also presented several analyses and methods in the Brazilian context for the first time: 1) testing the inverse farm size – productivity hypothesis, 2) satellite-based yield gap assessment, and 3) agricultural land cover mapping with data assimilation and machine learning.

While Objective 1 successfully innovated in data synthesis and yield mapping applications, the work was necessarily limited by time and data constraints. We used medium resolution (500 m) satellite imagery from MODIS for estimating crop yields and agricultural land cover. The selection of this imagery was largely motivated by data processing constraints. Performing computationally intensive signal processing routines on daily MODIS time series data covering the extent of Brazil requires substantial running times, even on large computing cluster nodes. New platforms such as Google Earth Engine and Amazon Earth have abstracted away data storage and processing architecture challenges for MODIS and higher-resolution imagery sources. Future work may port the current data processing pipeline over to one of these platforms. The chapter may be further improved by 1) incorporating yield prediction algorithms developed in Objective 2 and 3, 2) incorporating hierarchical modeling of soybean yield drivers, and 3) quantifying the effect of governmental and international investment in agricultural development.

Existing literature about a scalable crop yield mapping (SCYM) algorithm using satellite data suggested that mapping sub-field crop yield was feasible, but that hypothesis had not been tested until the work presented in Objective 2 and 3 (Lobell et al., 2015). We demonstrated that it was possible to predict maize (*Zea mays* L.) yields at 10 m resolution using no *in situ* observations. The findings in Objective 2 suggest that commercial and research applications of SCYM yield predictions may be feasible. The method could be used to provide yield maps for fields where yield data had not previously been collected.

Yield maps may be used to algorithmically delineate management zones with similar field conditions to simplify the process of planning site-specific management.

Objective 3 extends several themes from Objective 2. The work continues to advance the SCYM algorithm by investigating the relative performance of alternative algorithm specifications. The work also focuses on another facet of modern agriculture, advanced crop breeding. Field-based high throughput phenotyping (FHTP), has potential to increase the pace of development of new varieties possessing traits of interest. Results from Objective 3 suggested that monitoring crop yield response to sub-field conditions in a given season was possible, but prediction accuracy was poor relative to predictions of mean yields over time. The accuracy of yield predictions that is required for use with FHTP research must be informed by the degree of yield variation expected in response to field treatments. In other words, the signal to noise ratio of predictions made in Objective 3 may be too high to detect yield differences across treatment groups where the treatment effect was small. Because the sites were not managed for FHTP research, we were unable to test whether the SCYM algorithm was suitably accurate for an actual experimental case. However, yield predictions are likely suitable for designing FHTP studies, for example in determining optimal locations for matched control-treatment pairs.

In Objectives 1 and 2 we applied machine learning (ML) algorithms to create data inputs as well as final yield predictions. In both cases, we found this class of algorithms to improve modeling performance over standard linear regression approaches. Further research on modeling agricultural systems with machine learning is warranted. More sophisticated methods, such as convolutional neural networks should be examined for inductive modeling to leverage variation in large datasets.

The tremendous variety of field conditions and management decisions makes predictive modeling in agriculture a hard problem regardless of spatial scale. Historically, limited data availability has constrained the extent and type of research that could be performed on the magnitude and drivers of crop yields. Currently, innovations in modeling and data synthesis, such as those presented in this dissertation, are expanding the scope of potential crop yield research. Remote sensing data sources and innovative analytical methods will play a major role in progressing toward global yield targets by targeting investments for closing yield gaps, empowering farmers with sub-field yield data, and accelerating crop breeding research.

Summary of findings

The main findings of this dissertation can be summarized as follows:
Objective 1

1) Farm size and soybean yields were on average positively related across Brazil, but the strength and direction of the relationship varied by state.
2) Soy yield models detected crop yield heterogeneity across and within properties– the average farmer's yield in each county was 12.1% lower than the 95th percentile of county yields.
3) Soybean yield heterogeneity decreases with increasing farm size

Objective 2

1) More than 90% of pixel-level yield predictions were within +/- 10% of the mean yield estimated from yield monitor data.
2) We demonstrated that a machine learning model outperformed linear regression models for predicting mean yields, and that neither model was able to consistently capture within field yield variability.
3) Predictions at locations with more observation years tended to have better agreement with reported yields.
4) Our findings suggested that predicting biomass and then multiplying by a fixed harvest index was typically comparable to or better than estimating yield directly.
5) The introduction of a radiative transfer model for converting simulated crop canopies into pseudo-VIs lowered yield prediction performance in most cases.

Objective 3

1) Crop yield maps created from imagery may help to scale out FHTP research, but broad applicability of the SCYM algorithm for monitoring sub-field single-season crop response will vary by the error tolerance of the field trial design.
2) Maize yields created from public data agreed with yield monitor data (max. R^2: 0.73).
3) Predictions of rainfed maize yields outperformed those for irrigated fields.
4) Yield estimates were not sensitive to sensor choice, of four synthetic sensors tested.
5) Modeling yield directly rather than via a biomass proxy generally provided better prediction.

Directions for future work

Mapping causes of yield loss using remote sensing imagery and crop simulation models

The studies comprising the dissertation demonstrated several methods for predicting crop yields using remote sensing imagery and other datasets. A valuable extension of these methods is to predict not just crop yields but also other characteristics of cropping systems. Mapping management factors, for example, would be valuable for assessing strategies for adapting to changes in interannual climate variability. Factors such as planting date may be possible to infer on the basis of crop emergence observed in satellite imagery, soil conditions, and temperature. Farm management data is typically limited in availability, even in aggregate, so predictive approach may be especially valuable.

Scale out research on sub-field yield prediction

The cases presented in Objectives 2 and 3 in Mead, Nebraska provided valuable illustrations of the potential effectiveness of SCYM yield predictions at sub-field scales. However, yield monitor datasets from a wider geographic scope are needed in order to test the generalizability and robustness of the methods in Objectives 2 and 3. Future work should be conducted on a larger sample of sites in order to investigate, for example, 1)

the bias structure of yield predictions, and 2) differences in prediction accuracy across irrigated and rainfed fields. Though aerial hyperspectral imagery is not available at scale, satellite imagery with spatial resolution <= 30 m is accessible for research through platforms such as Google Earth Engine.

Refine SCYM approaches by localizing plausible growing scenarios for crop simulations

The SCYM algorithm uses crop simulations for a range of plausible growing conditions to work around scarcities of site and management information. Plausible growing scenarios for a region of interest are defined from literature or expert knowledge. A logical progression of the SCYM algorithm is to filter the scope of simulated scenarios to consider when selecting data for training a yield model. Inclusion criteria could be defined based on 1) auxiliary data, such state-level planting date statistics, 2) conditions estimated from remote sensing imagery, as Jin et al. (2017) demonstrated in their approximation of planting date windows, or 3) probabilities assigned from a classification model, trained on crop simulations, linking simulated canopy properties to simulated environment and management conditions. The third approach is a so-called ill-posed problem–a given simulated canopy state could be caused by a multitude of input variable combinations. Model inversion strategies developed for radiative transfer models (RTM) are likely applicable here (Atzberger, 2010).

Investigate the mechanisms contributing to smaller farms having higher yield gaps

Objective 1 showed that mean yield gaps were higher among smaller farms, and discusses potential drivers such as non-uniform access to inputs like fertilizers and recent seed varieties, or information like weather forecasts and agronomic recommendations. Modeling of management factors across the soybeans farms, extending future areas for research described above, could be combined with farm size and yield data to identify causes of yield loss. Characterizing the regions and farm types with greatest potential for closing yield gaps through low cost interventions can help target assistance.

Use sub-field yield predictions for providing site-specific management recommendations

The findings of Objective 2 and 3 concluded that mapping crop yields without *in situ* data was feasible and could be useful to decision making in precision agriculture. Detailed gridded crop yield maps have multiple potential applications for targeted crop management:

1) Provide recommendations for “precision conservation” practices which might include planting perennial grasses in wet, low-yielding part of a field, where the benefits to soil retention and water quality outweigh potential returns on a crop;
2) Provide design and assessment technologies for creating variable rate input prescriptions;
3) Suitability analysis to determine if site-specific management would be beneficial for given fields, especially as a service for farmers considering purchasing equipment for precision agriculture.

References

Atzberger, C., 2010. Inverting the PROSAIL canopy reflectance model using neural nets trained on streamlined databases. Journal of Spectral Imaging, 1(1): a2.

Jin, Z., Azzari, G., Burke, M., Aston, S. and Lobell, B.D., 2017. Mapping Smallholder Yield Heterogeneity at Multiple Scales in Eastern Africa. Remote Sensing, 9(9).

Lobell, D.B., Thau, D., Seifert, C., Engle, E. and Little, B., 2015. A scalable satellite-based crop yield mapper. Remote Sensing of Environment, 164: 324-333.

Appendix: Software packages developed

gghexbin

This package for the R programming language is used to visualize spatial point or raster datasets which are too dense to effectively visualize without aggregation. The package provides methods for algorithmically constructing spatial, hexagonal bins for aggregating variables with a selection of summary statistics. Binned data is displayed with hexagon color and/or size. See figures in Chapter 2 for examples of this library's outputs.

ggbinscatter

Binscatter plots convey the effect of one variable on another. Each variable is regressed on a set of control variables and each regression's residuals saved. Residuals are then averaged within equal sized bins defined by the x-axis variable. The slope between x- and y-axis residuals is equal to the beta coefficient for the x-axis variable in a regression of the y-axis variable on the x-axis variable plus controls. We implemented the method in R following the Stata package 'binscatter'.

pydssat

This Python package provides an API for the DSSAT-CERES-Maize crop simulation model. Development of the package was motivated by the need to simulate maize crops with programmatically defined input variables. Prior to the creation of this package, the crop simulation pipeline required manual configuration of 'experiment' input files which specify management conditions, weather data, and other simulation inputs. Pydssat provides formatting tools for converting to and from specialized file formats required by the simulation model. The package is designed to easily parallelize simulation runs across computing cores or multiple machines.Additional code was produced to interface with python-DSSAT simulations from R.

Appendix: Research internship report

The purpose of this short piece is to describe my three-month internship experience at The Climate Corporation, January 26 - May 1, 2015, highlighting its value in the context of my dissertation work. The internship was conducted in partial fulfillment of requirements for the NSF IGERT Water Diplomacy Fellowship. The Climate Corporation (abbv. Climate) is a technology company founded in San Francisco that provides develops science-driven farm management. The company focuses on interdisciplinary quantitative research including developing and applying methods in machine learning, remote sensing, and crop modeling to the improvement of models for climatological and agronomic forecasting. The science department comprises the climatology, agricultural modeling, and geospatial (focusing on remote sensing) teams. I worked with Dr. John Gates in the agricultural modeling group, but also collaborated with the climatology team.

I led a project improving methods for snow accumulation and ablation modeling in the land surface model developed at Climate. Snow hydrology processes are a major contributor to the timing and intensity of nutrient flux in agricultural soils. Snow cover affects nutrient flux rates and timings by modulating soil temperature and water saturation across the soil profile. Better nutrient flux forecasting enables farmers to more precisely target the amount and timing of nutrient applications - lowering wasted input costs and reducing nutrient runoff into waterways. My project contributed to the improvement of Climate's Nitrogen Advisor service which is a web tools where farmers can optimize nitrogen applications in their fields with respect to a probability distribution of future weather scenarios.

I entered the internship with an inclination to work in industry after finishing my PhD, and that interest was affirmed by my time at Climate. The experience served as an informative case study in translating academic research into products which address "pain points" and add value to farm operations. Research project cycles reflected a high degree of academic rigor with formal project writings peer-reviewed at each stage including: a one-page research brief by a team lead, a 15-30 page research proposal by the researcher, a 15-50 page research report, and presentations to team and department-wide audiences. Twice-weekly small team check-ins and weekly seminars were forums for ongoing dialog. Research code also underwent frequent peer-review. The peer-review process was supported by excellent research infrastructure which provided tailored dataset, code, and document management. Many features of their research infrastructure alleviated common pains in academic research that I've encountered related to data sharing, formatting, and versioning; rapid access to scalable computing resources; and multi-party cross-site meeting scheduling. I hope to replicate some of these resources in my dissertation work and in future organizations that I join.

My snow hydrology project gave me an opportunity to develop a deeper understanding of terrestrial simulation models and to acquaint myself with modeling methods that were new to me. I developed a physical energy-balance model for snow accumulation and ablation use a multi-layer representation of snowpack. This approach challenged me to learn about and apply knowledge of the thermodynamics of snow and soil hydrology. I

compared the performance of the snow energy-balance model and an extant statistical model with several validation datasets. I then evaluated the sensitivity of Climate's nitrogen flux model to changes in the snow simulation submodule.

In addition to expanding my modeling skillsets, working with experienced software engineers led me to grow as a software developer. Even within the science teams, the level of software development professionalism was high - modern tools and software best practices were the norm. While the research process is not intended to yield production-ready code, it is designed to produce code that can readily be translated into production, and this informs the quality of documentation, the selection of algorithms, and the modular design of tools. In my time at Tufts, my software development efforts have been principally solo endeavors - from the choice of the programming languages to the writing, revision, and review of the code. I rapidly grew as a developer when immersed in collaborative projects at Climate. It was reassuring to find that the programming language that I've invested in, Python, R, and Latex, though uncommon at Tufts, were some of the most commonplace at Climate. Many of the software design principles I learned during my internship have informed the design of the technical systems which support my dissertation research.

In addition to the technical experience I gained at Climate, I also drew lessons from the project management strategies I witnessed. While I was there a short time, I observed the cycling of customer research, research plan definition and execution, product development, and customer feedback stages for projects outside of my own. Some challenges in this cycle are unique to industry work: evaluating research priorities based on expected revenue returns, strategically coordinating research timelines with product releases, and communicating uncertainty in complicated models to end users. Working in an environment where I and others around me were faced with these challenges provided me with research and project management skills that will tangibly benefit my current and future work.