An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the United States

David M. Johnson *

United States Department of Agriculture, National Agricultural Statistics Service, 3251 Old Lee Highway, suite 305, Fairfax, VA 22030, United States

ABSTRACT

Four timely and broadly available remotely sensed datasets were assessed for inclusion into county-level corn and soybean yield forecasting efforts focused on the Corn Belt region of the central United States (US). Those datasets were the (1) Normalized Difference Vegetation Index (NDVI) as derived from the Terra satellite's Moderate Resolution Imaging Spectroradiometer (MODIS), (2) daytime and (3) nighttime land surface temperature (LST) as derived from Aqua satellite's MODIS, and (4) precipitation from the National Weather Service (NWS) Nexrad-based gridded data product. The originating MODIS data utilized were the globally produced 8-day, clear sky composited science products (MOD09Q1 and MYD11A2), while the US-wide NWS data were manipulated to mesh with the MODIS imagery both spatially and temporally by regridding and summing the otherwise daily measurements. The crop growing seasons of 2006–2011 were analyzed with each year bounded by 32 8-day periods from mid-February through late October. Land cover classifications known as the Cropland Data Layer as produced annually by the National Agricultural Statistics Service (NASS) were used to isolate the input dataset pixels as to corn and soybeans for each of the corresponding years. The relevant pixels were then averaged by crop and time period to produce a county-level estimate of NDVI, the LSTs, and precipitation. They in turn were related to official annual NASS county level yield statistics. For the Corn Belt region as a whole, both corn and soybean yields were found to be positively correlated with NDVI in the middle of the summer and negatively correlated to daytime LST at that same time. Nighttime LST and precipitation showed no correlations to yield, regardless of the time prior or during the growing season. There was also slight suggestion of low NDVI and high daytime LST in the spring being positively related to final yields, again for both crops. Taking only NDVI and daytime LST as inputs from the 2006–2011 dataset, regression tree-based models were built and county-level, within-sample coefficients of determination (R^2) of 0.93 were found for both crops. Limiting the models by systematically removing late season data showed the model performance to remain strong even at mid-season and still viable even earlier. Finally, the derived models were used to predict out-of-sample for the 2012 season, which ended up having an anomalous drought. Yet, the county-level results compared reasonably well against official statistics with $R^2 = 0.77$ for corn and 0.71 for soybeans. The root-mean-square errors were 1.26 and 0.42 metric tons per hectare, respectively.

© 2013 Elsevier Inc. All rights reserved.

1. Introduction

1.1. Crop yield statistics

Accurate and timely estimation of local and regional crop yield statistics is important for a variety of reasons. On the macroeconomics level they allow societies to understand the food and fiber supply which in turn helps the demand side plan for and better utilize the finite crop resources. In the most developed countries this is manifested through futures contract markets which are most efficient and fair for price discovery when transparent and current statistics are available. Local, direct to consumer markets work similarly in that statistics help both parties understand the value of the crop. From a management

standpoint, yield information gives a farmer a baseline of what is typically expected to be produced and thus can be used to best establish risk, insurance premiums or the value of input costs. Established yield information also highlights the impact to crops from natural events such as severe weather or changing climatic conditions. Likewise, regional yield statistics help quantify how strategies such as planting methodologies, irrigation, fertilizer and pesticide use are playing out in aggregation and can identify regions that are chronically underperforming, or have a "yield gap."

The United States Department of Agriculture (USDA) spends considerable effort in determining United States (US) crop yields in service to the agricultural community. The statistical arm of the USDA, the National Agricultural Statistics Service (NASS), conducts two large panel surveys (USDA, 2012) that are annually ongoing throughout the growing season (USDA, 2010) to establish state- and national-level yield estimates. The first is known as the Agricultural Yield Survey

* Tel.: +1 703 877 8000x169.
E-mail address: dave.johnson@nass.usda.gov.

which is based on a maintained "list frame" of farmers and the results are directly reliant on the information they provide. Each year thousands of those farmers are randomly selected, contacted monthly by phone during the growing season and asked to report expected yields for their crops grown. Information from all those sampled is then combined and summarized to derive a set of regional yield "indications." Run in parallel is the Objective Yield Survey which derives an independent set of indications through biophysical crop measurements. For it, hundreds of small plots are randomly sampled from fields throughout the major growing areas and visited by an enumerator a few times during the crop season. Attributes collected include plant counts per unit area, grain size, grain weight, etc. The information from all of the plot-level data is ultimately aggregated into a model to derive this second set of yield indications. The Objective Yield Survey is more limited in scope over the Agricultural Yield Survey in that it only focuses on the dominant commodity crops like corn, soybeans, wheat, potatoes and cotton. Ultimately, the results from both surveys, along with any relevant ancillary information, are analyzed by the NASS Agricultural Statistics Board (ASB) to establish the monthly published yield forecasts.

After the season is complete late in the fall, an additional widely cast survey is undertaken which documents agricultural production statistics down to the county-level. For it questionnaires are sent to a much larger sample of producers asking for responses on many agricultural facets of their operation including estimates of their crop yields. Finally, these county-level statistics are assessed and published to reconcile with the previously established ASB national- and state-level yields.

Any further independent, error assessable and cost effective measures of crop yield indications that can be provided to the ASB are welcome. Real-time measuring of crop yields from remote sensing technologies has been promoted as a feasible methodology but has been fairly limited in implementation (Allen, Hanuschak, & Craig, 2002; Baruth, Royer, Klisch, & Genovese, 2008; Reynolds et al., 2000; Rojas, 2007). Reasons for lack of uptake are likely many but probably lead by a perception that results are not being seen as accurate, timely, or objective enough. Furthermore, remote sensing estimation of crop yields has potentially been hindered due to the unknown availability, cost and capacity of future imagery data combined with the highly specialized nature of the work for which it may be hard to find skilled and experienced labor.

In terms of US crop statistics themselves, corn and soybeans are the two largest commodities grown by land area and the planted acreage has steadily expanded in reach by about 25% over the last couple of decades (USDA/NASS Quick Stats). Yield trends for these crops have been increasing at a similar rate but see more relative variability year to year. Corn and soybeans from the US are high value and significant commodities on global export markets and of late they have been volatile in pricing. This suggests, at least in part, that the true amount produced has not been fully understood at all times.

1.2. Remote sensing of crop yields

Monitoring crops via satellite remote sensing is not a new idea or one in which there is a lack of research. Funk and Budde (2009) showed a summary of the work in a variety of sensor, location, and crop type contexts. Gallego, Carfagna, and Baruth (2010) also presented a history targeted specifically to crop production estimation. Even with all this aggregated work, assimilating the results to summarize to a best practice is confounding because the research has targeted different ecoregions of the globe, does not use the exact same type of input datasets or has varying methodologies. Furthermore, the specific crop type of focus has varied across the studies making the outcomes further difficult to compare. However, in general there has been more emphasis on corn, soybeans and wheat. Reasons why these three crops in particular have been the most studied are unknown but likely because they are found wide spread and in large quantities geographically. An alternative reason could be that there has been found better remote

sensing yield estimation success with them versus other potential crops (and failures tend not to get published).

Regardless of crop being investigated, a common and central theme of this type of remote sensing research involves the reduction of the sensor's multispectral channels into a single metric known as the Normalized Difference Vegetation Index (NDVI), analyzing its response throughout the crop growing season and then relating it again to in situ collected crop information. NDVI is calculated from the red and near-infrared (NIR) spectral channels as

$$NDVI = (NIR{-}red)/(NIR + red).$$

NDVI exploits the large difference seen between the red and NIR bands for heavily vegetated land cover types and has been shown to be strongly correlated with plant productivity in both in situ (Hatfield, 1983; Shanahan et al., 2001; Viña et al., 2004) and remote sensing applications (Basnyat, McConkey, Lafond, Moulin, & Pelcat, 2004; Tucker, 1979). NDVI is often preferred over the independent use of the red and NIR channel in that it simplifies data analysis into a single metric while at the same time it is a normalization which helps to reduce data errors due to poor viewing geometry or hazy atmospherics. The normalization also allows for easier comparison across different sensors.

Many sensors with the ability to provide NDVI have been utilized throughout the past few decades. Long-term, ubiquitous, and freely available US satellite assets are reviewed here. The program with the longest lineage is that of the Advanced Very High Resolution Radiometer (AVHRR) sensor. Variants of it have been aboard over a dozen operational polar orbiter meteorological satellites that were first launched in the late 1970s. The last was placed into orbit in 2009. AVHRR is decently suited from monitoring vegetation dynamics given its daily revisit rate and reasonable spatial resolve of a little over 1 km which is fine enough for monitoring relatively homogenous crop areas. Assessment of AVHRR NDVI phenology and the general relation local crop yields has been performed (Ferencz et al., 2004; Maselli & Rembold, 2001) in addition to analysis targeted toward the specific commodities of corn, soybeans or wheat (Benedetti & Rossini, 1993; Hays & Decker, 1996; Mkhabela, Mkhabela, & Mashinini, 2005; Salazar, Kogan, & Roytman, 2007; Wall, Larocque, & Léger, 2008). Modeling results have been shown reasonable for all but there are limitations on the yield estimation precision. Errors may be a function of the native coarse pixel size of AVHRR, which is larger than most fields even in heavily mechanized agricultural regions, or due to the sensor itself which may not be able to provide adequate spectral information with low noise.

Technology has progressed and a newer and more sophisticated sensor called the Moderate Resolution Imaging Spectroradiometer (MODIS) improves on AVHRR (Fensholt & Sandholt, 2005; Huete et al., 2002) particularly in terms of spectral response, spatial resolution, and having more emphasis placed on land related observations (Justice et al., 2002). MODIS is aboard two earth science research oriented satellites, Terra and Aqua, which were launched in 1999 and 2002, respectively. MODIS carries a total of 36 spectral bands with most having a nadir ground resolution of about 1 km, which is similar to AVHRR. However, two of the key bands for land observations, the red and NIR, have a much finer resolution of about 250 m. MODIS derived crop and productivity yield work has again often relied on NDVI phenology with a focus on wheat (Becker-Reshef, Vermote, Lindeman, & Justice, 2010; Mkhabela, Bullock, Raj, Wang, & Yang, 2011; Reeves, Zhao, & Running, 2005) and corn or soybeans (Bolton & Friedl, 2013; Doraiswamy et al., 2004; Doraiswamy et al., 2005; Funk & Budde, 2009; Guindin-Garcia, Gitelson, Arkebauer, Shanahan, & Weiss, 2012; Sakamoto, Gitelson, & Arkebauer, 2013). Yield research which use MODIS data have proven better results than those that use AVHRR data. There are no identical MODIS follow-on mission planned but a similar meteorology focused polar orbiting sensor called the Visible Infrared Imaging Radiometer Suite (VIIRS) program has begun. A prototype VIIRS instrument was launched in 2011 aboard a satellite platform

named Suomi National Polar-orbiting Partnership. VIIRS was ultimately designed to replace the AVHRR program but has characteristics more similar to MODIS.

Improving on spatial resolution further are the Landsat series of earth observation satellites. They have also provided an opportunity for assessing crop yields from space. Landsat has carried a few optical and thermal sensor variants since the first launch in 1972 with steady performance and calibration improvements on successive missions to the present. The original sensor was the Multispectral Scanner (MSS) which had four or five spectral channels and a spatial resolution of about 60 m. More contemporary Landsats have carried variants of a sensor called Thematic Mapper (TM). It has even more spectral bands than the MSS and a ground sample resolution of 30 m. The eighth Landsat mission, having eleven bands, was launched early in 2013. While the spatial resolution from all Landsat sensors is vastly superior to AVHRR and MODIS, Landsat has been hindered for use in monitoring crop yields given its relatively narrow swath of about 185 km which provides a temporal revisit rate of only 16 days. This severely limits the number of observations within a typical growing season, particularly when factoring in the likelihood of clouds. However, yield and productivity work utilizing TM data have been prototyped for different crop types (Doraiswamy, Moulin, Cook, & Stern, 2003; Gitelson et al., 2012; Liu et al., 2010; Lobell, Ortiz-Monasterio, Asner, Naylor, & Falcon, 2005) and it has been shown possible, albeit for relatively small analysis areas only.

Weather variables have had a much longer history of use for understanding crop yields compared to NDVI which was spawned from multispectral remote sensing technologies. Temperature and precipitation are the ubiquitous choice for yield predictor variables. Ample, and too lengthy to list, research going back a century ago has shown the negative correlation of heat and positive correlation to precipitation for corn yields (Smith, 1914; Wallace, 1920). More recent analysis has been increasingly sophisticated and refined (Kauffman & Snell, 1997; Lee, Phil Kenkel, & Brorsen, 2013; Tannura, Irwin, & Good, 2008) as more detailed climate and weather datasets in addition to wide-scale and historical yield validation information have become available. Most recent analysis is also integrating future climate variability into crop models (Dixon, Hollinger, Garcia, & Tirupattur, 1994; Schlenker & Roberts, 2009). Ultimately, all work converges around the notion that too much summer heat reduces crop yields while decent rainfall amounts help increase it.

While there has been a lot of agro-meteorology research, little of it has used remotely sensed datasets as input for predicting crop yields. But now, unlike in the past, there exists the ability to collect spatially detailed proxy parameters for agro-meteorology variables via remote sensing. Land surface temperature (LST) is a similar, but not exactly the same, measurement as more commonly collected air temperature. The two variables are strongly related, though, with LST having larger temperature extremes and being locally dependent on the land cover type (Mildrexler, Zhao, & Running, 2011; Prihodko & Goward, 1997; Wan, 2008). In terms of satellite LST and the relation to yield no known direct research has been done. However, there have been related studies that have used LST to monitor agricultural drought (Feddema & Egbert, 2005; Patel, Parida, Venus, Saha, & Dadhwal, 2012). Some have also integrated NDVI alongside LST for the analysis (Goetz, 1997; Prasad, Chai, Singh, & Kafatos, 2006; Wan, Wang, & Li, 2004).

Research using remotely sensing derived precipitation in the context of agricultural yields would also seem prudent given that rainfall rates can vary dramatically by location, and thus interpolated measurements from the limited weather stations may not be providing sufficient spatial detail. To counter this problem, a dedicated satellite for measuring rainfall amounts was indeed developed called the Tropical Rainfall Measuring Mission (TRMM). Unfortunately, its steeply inclined polar orbit around the earth is only focused on the equatorial Tropics and thus it provides no imagery farther than 35° north and south. So, TRMM can only monitor the southern portion of the US and thus misses the Corn Belt, and most other major growing areas around the globe, completely. Within the US however there is an alternative option. The National Weather Service (NWS) part of the National Oceanic and Atmospheric Administration (NOAA) operates a ground based remotely sensed dataset with coverage being provided by a large array of Doppler radars from which precipitation can be estimated (Seo, 1998). Up to this point few applications of the data beyond pure rainfall and hydrological monitoring have been published. Reason why are unknown but are likely a combination of it being a new dataset and one that is not obviously available to those in the remote sensing community. It could also be seen as providing too much spatial and temporal information that is not obvious how to exploit.

1.3. Objective

This applied-leaning research here attempts to provide an increased understanding of the corn and soybean production in the US through two primary objectives:

1. *To firmly understand the relationship between US county-level average corn and soybean yields* versus *relatively common variables collected throughout the entire crop growing season* via *remote sensing*. The variables assessed were a time series of the NDVI as derived by the Terra MODIS sensor, daytime and nighttime LST as derived by the Aqua MODIS sensor and precipitation estimates produced from the NWS Nexrad Doppler weather radar system. The Aqua LST and Nexrad precipitation datasets were both considered novel in regard to what has been researched previously for predicting crop yields. Also, equal importance to corn and soybeans was given whereas in many studies only a single crop type is analyzed.
2. *To build a county-level yield forecasting model, from those remotely sensed variables deemed important, that is not only accurate but also timely and relatively simple to run in an operational environment.* A large and robust cross-sectional panel of historical county-level yield information from NASS was used as the basis of the forecasting and implemented with regression tree, machine learning software. Rulequest Cubist was chosen for the modeling solution and is viewed as both pragmatic and unique.

2. Materials and methods

2.1. Study area

The crop yield work focused on 12 states which sit in the central and north US and are known collectively as the Corn Belt (Fig. 1). They are Arkansas, Illinois, Indiana, Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota and Wisconsin. The states make up over 75% of the US corn and soybean production and thus are focused upon by the NASS Objective and Agricultural Yield Surveys in making national crop production estimates. Note, corn does not include Arkansas or North Dakota in the Objective Yield Survey and soybean does not include Wisconsin. The 12 states are geographically contiguous, relatively flat and level, and have very fertile soils. Winters are usually cold and snowy and summers warm and humid. Prior to transitioning to wide-scale farming in the middle and late 1800 s the area was covered by native tall grasses. Most of the states are now dominated by the two crops of soybeans and corn which are often rotated symbiotically in the same field alternatively from year to year. The states towards the west, ebbing into the drier Great Plains, do however have a higher diversity of crop types and in particular small grains crops like wheat and barley. Recent and overall trends have shown corn and soybeans to be slowly accounting for more and more of the cropland area and reducing the area of other crops (USDA/ NASS, Quick Stats).
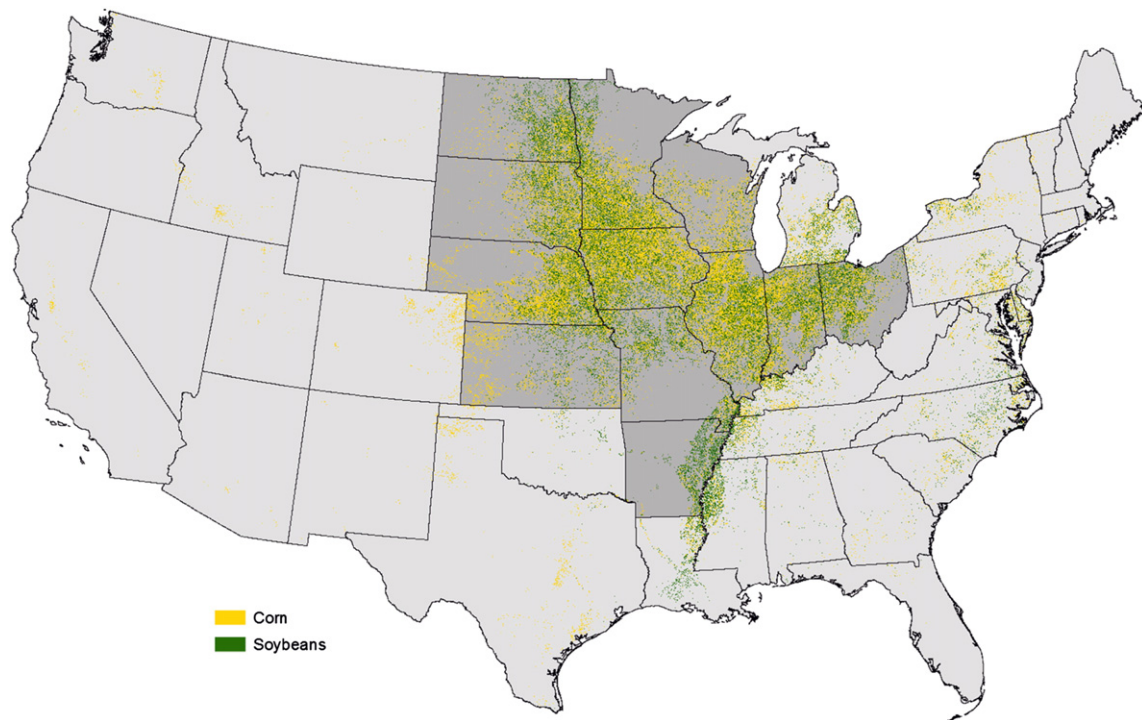
Fig. 1. Distribution of corn and soybean within the US as depicted from the 2012 USDA Cropland Data Layer. Deeply shaded states are those that were used in analysis.

2.2. Datasets

Four grid-based, map referenced datasets were explored in terms of their relationship to county level crop yields and for possible inclusion into corn and soybean yield forecasting models. They were:

1. *NDVI* — calculated from the 8-day composited surface reflectance bands product (Vermote, Kotchenova, & Ray, 2011) from the Terra MODIS sensor (termed MOD09Q1),
2. *Daytime LST* — produced from the 8-day composited thermal product (Wan, 2007) from the Aqua MODIS sensor (termed MYD11A2),
3. *Nighttime LST* — also produced from the 8-day composited Aqua thermal MODIS sensor, and
4. *Precipitation* — derived from the NOAA/NWS Nexrad ground based system (Seo, 1998; Seo, Breidenbach, & Johnson, 1999).

These particular variables were chosen for several reasons. First and foremost, they all are suggestive of being related to crop yields via simple conventional wisdom and more importantly, agronomy type research. Secondly, each of the datasets is spatially detailed and allows for modeling at relatively fine geographic scales approaching, particularly for NDVI, the crop field level. As incorporated here the NDVI resolution is about 250 m, the LST is 1 km, and the precipitation is 4 km (6.25, 100, and 1600 ha, respectively). Temporal revisit times are also high with all of these products as raw input data for each are collected daily. On any given date cloud cover impacts the quality of imagery collected from satellites and thus MODIS composite products which pick the "best of" pixels over an eight day period are distributed to end users helping simplify data management and processing. The precipitation dataset used is also produced and distributed daily but not impacted by cloud cover due to the penetrating ability of radar. Composited precipitation type products are also produced but they are accumulation periods of a week or month and not exactly in alignment with MODIS datasets. Thirdly, each of the datasets has a reasonable history depth. Terra data collection goes back to the year 2000 and Aqua 2002. These start years were soon after the launch of each satellite. The NWS precipitation product archive began in 2005. All products have a finite lifespan but it is believed that each of these dataset will continue to be at least available into the near future. For MODIS, while it is aging, at the time of writing there is no sense that the sensor is about to fail or be shut off, and even if one platform does fail the other can be considered a backup. Furthermore, for MODIS the VIIRS sensor is considered a replacement. More VIIRS sensors are planned for the operational Joint Polar Satellite System so whatever is learned about yield forecasting from MODIS will likely be directly applicable for years to come. And finally, and perhaps most importantly, all four of these datasets are provided free of charge and accessible immediately via the Internet. They each have little lag time from collection to distribution. The last point is particularly critical if trying to forecast yields in an operational and time sensitive environment.

The respective Terra and Aqua satellite from which the MODIS data were chosen was done for purposeful reasons. Terra was used for NDVI because its approximately 10:30 a.m. overpass time is thought to provide the most optimal time for observing surface reflectance versus Aqua which is at 1:30 p.m. and more likely on average to have encountered afternoon clouds. Conversely, Aqua was chosen for LST observations instead of Terra because the daytime overpass of 1:30 p.m. aligns closer in time with the maximum daytime temperature. Likewise, the Aqua nighttime temperature product was also chosen over Terra's because it would be more representative of the low temperature point of the diurnal cycle. Both daytime high and nighttime low temperatures are often talked about in terms of crop yield impacts, albeit usually anecdotally, so it was desirable to include a proxy for both. The 8-day composited products were chosen as a compromise over single day observations or even longer compositing time windows. Furthermore, the 8-day compositing period is seen as usually sufficient time in which to have obtained at least one clear sky observation, in the Corn Belt at least, for inclusion into the mosaic.

All of the MODIS MOD09Q1 and MYD11A2 data were obtained via File Transfer Protocol (FTP) over the Internet from the Land Process Distributed Active Archive Center (LP DAAC). The MODIS data is global in nature and distributed by standardized five degree by five degree "tiles." Six tiles were needed to cover the entire Corn Belt region of interest and consisted of "horizontal/vertical" referenced tiles 9/4, 9/5, 10/4, 10/5, 11/4 and 11/5. Thirty-four time periods of the 8-day data

were pulled for the years 2006 through 2012. The start time of the first composite for each year was February 10th, or day-of-year (DOY) 41. The start date of the final annual composite acquired was November 1st which equals DOY 305. The entire data series time window was seen as more than comprehensive enough not only to cover the Corn Belt crop growing season, which is typically from May to October, but also allowed for spring pre-season information to be investigated. The MODIS raw files were obtained natively in the Hierarchical Data Format. The "Collect 5" version was used.

The NOAA/NWS data were also obtained via Internet FTP but from the NWS Advanced Hydrologic Prediction Service website. Because the rainfall data were to be aggregated to the same 8-day schedule as MODIS, all the daily data from February 10th through November 8th were gathered. The same 2006–2011 years were spanned. The NWS information was natively in Environment Systems Research Institute's (ESRI) Shapefile vector format in a standardized and regularized point grid of approximately 4 km in spacing. The Shapefile data were provided by the NWS in "compressed".zip format.

The dependent variable for which the remote sensed dataset would ultimately be compared came from annual county-level corn and soybean yields as published by the USDA NASS. The data were downloaded from the Internet in tabular form from the USDA/NASS Quick Stats website. An annual average yield for each county, for both corn and soybeans, across the Corn Belt, years 2006–2011, was ultimately gathered. There are roughly 1000 counties in the region so with seven years of data a large amount of data points resulted. A handful of the counties did not contain data in certain or all years because they failed to meet NASS publication standards of having a sufficient number of farmers or large enough response rate. This typically means there was very little or none of that crop produced in that particular county and any estimate would have been unreliable.

Finally, land cover "masks" depicting crop areas were obtained. NASS produces a robust annual land cover classification over the US called the Cropland Data Layer (CDL). The first US wide coverage was developed in 2009 (Johnson & Mueller, 2010) and has continued annually since (Boryan, Yang, Mueller, & Craig, 2011). Year 2008 was completed retrospectively for the nation and for the Corn Belt area CDL classifications exist in the years 2006 and 2007 as well. CDL accuracies for corn and soybean fields are particularly strong and thus can act as an excellent proxy for ground validation data. All of the CDL information was obtained over the 12 state region from 2006 to present. The 2006 start year of the CDLs also drove the starting point for inclusion of the MODIS and precipitation datasets even though they both exist farther back in time. The CDLs were ultimately used to isolate only the relevant areas of corn and soybeans from the MODIS and precipitation data.

2.3. Data preparation

A consistent grid-based reference frame was first established in which all of the data products used would harmoniously nest. The MODIS-based "250 meter" sinusoidal equal area map projection was chosen. Selecting this coordinate system minimized the overall amount of data processing needed since the majority of the data used was from MODIS. It also avoided any resampling caused degradation of the MODIS data since it would not have to undergo a map reprojection. Of note, while the MODIS data are labeled as 250 m in resolution they are more precisely 231.66 and indeed that spacing is what was used. The generalized 250 m, and multiples of it, nomenclature will be continued here.

With the reference grid established, a 250 m resolution raster-based mask file for each state (again, Arkansas, Illinois, Indiana, Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, and Wisconsin) was created to define which pixels were within a particular state's border. All boundary definitions were derived from ESRI's *Data and Maps* Detailed Counties Shapefile (dtl_cnty.shp) by first converting the vector file into the sinusoidal projection and then determining

which 250 m reference grid cell belonged to what state. A majority area rule was used to assign pixels that were split by two or more states. Each final state mask was ultimately aligned spatially with the MODIS reference grid and stored in 1-bit Earth Resources Data Analysis System (ERDAS) Imagine format with pixels coded zero defined as residing outside the state border and those coded one being inside.

Similarly and in conjunction, 250 m pixel identifications grids were developed by state to define which county each pixel best fit. Grid cells were assigned to their Federal Information Processing Standard (FIPS) codes which typically are ordered alphabetically by county name starting at a value of one and increasing by odd numbers only. The ESRI County Shapefile was again used as input. The output however was in 16-bit format to account for FIPS values that could range above 255.

Corn and soybean masks were next developed, by state and year, from the available CDL 2006–2011 annual information. These masks were developed to provide a means to isolate corn or soybean areas from the MODIS and precipitation data (Kastens et al., 2005). A four step process was used to tailor the information from each 30 or 56 m CDL into something appropriate at the 250 m level. The first step was to reproject the raw CDL grids from their native Albers equal-area conic projection to that of the MODIS sinusoidal. The output grid cell size was kept the same as the input (thus, either 30 or 56 m). Next, the number of crop of interest pixels from the CDL that were contained within each 250 m pixel grid cell was calculated. The total possible was also established. Depending on the CDL pixel size, and the vertical and horizontal alignment of it versus the 250 m grid, the maximum counts possible could vary between 16 and 49. Next, the total corn (or soybean) counts in each MODIS grid cell were divided by the maximum possible to give an areal proportion. Finally, that proportion information was compared to a threshold of 0.9 to establish whether the 250 m cell contained enough of the crop of interest. This 90 percent threshold value was subjectively determined but was felt to be a good compromise between the possibility of excluding too many pixels, because of noise or minor cover type mixing, yet conservative enough to be robust. The final corn and soybean masks for each state were a 1-bit formatted file where a value of one equaled at least 90% of the MODIS-scaled pixel to be containing corn (or soybeans) as a cover type while a value of zero was less than 90%. In the end, each of the state-level mask grids had minimally thousands of pixels declared as corn or soybeans and most counties within them hundreds of pixels. This was deemed more than sufficient to build a high quality sample for intersection against the NDVI, LST and precipitation datasets.

With the state, county and yearly corn and soybean masks all in place, next the remotely sensed data was managed. For the MODIS surface reflectance data (MOD09Q1) each of the 6 Hierarchical Data Format (HDF) tiles comprising the Corn Belt per 8-day time step were successively mosaicked into a single HDF file. Next, the red and near-infrared channels (HDF bands 0 and 1, respectively) were extracted and stored in ERDAS Imagine format. Finally, the NDVI was calculated for each entire image mosaic. All data management was kept in the original 16-bit scaled by 10,000 format and thus the output NDVI was also scaled. There were ultimately 34 NDVI mosaic images per year created and thus a grand total of 204 over the six years from 2006 through 2011.

Next all of the corresponding MODIS LST datasets were ingested. The processing used was very similar as to the surface reflectance data in that each date of six tiles were first read and mosaicked. However, after combining the tiles, it was HDF layers 0 and 4 that were imported into ERDAS Imagine format. Those particular bands corresponded to the daytime and nighttime information, respectively. No further processing was then done and thus temperatures were left in their native, scaled by 50 K, 16-bit and 1 km spatial resolution, format. In total there were 204 images for both the daytime and nighttime LST information.

Through subjective observation of the MODIS composite imagery it was noticed that there were often times when the 8-day window was

not long enough to assure noise free pixels. The errors tended to be more pronounced in the early or later periods of the growing season and in more northerly locations. It was also concluded that for both NDVI and LST when noise occurred it usually erred in the negative direction. These findings were consistent with what would result when periods of prolonged clouds or haze or even standing snow or water existed. Many methods for minimizing noise within remotely sensed time series data have been proposed and used to smooth the data. They range from simple to sophisticated and include analysis methods like least squares (Jönsson & Eklundh, 2004), harmonics (Moody & Johnson, 2001) or wavelets (Galford et al., 2008). Choosing a best method is not always straightforward (Hird & McDermid, 2008) and a disadvantage of them all is they usually perform best when a true periodic length of data is available. For many applications this may not be an issue. But, if trying to analyze and forecast crop yields in the middle of the season, the full time series of the current year will not be available, or in itself will need to be modeled. Furthermore, potential outlier or unusual data needs to be scrutinized carefully for any smoothing methods for fear of imputing additional error into the time series.

Thus, here only a simple and mild error checking of the data was employed to help minimize the effects of the typically lower error values. To do so, each pixel within the time series was compared to the pixel one composite date prior and one date after. If that middle value was lower than both, it was deemed to be bad. It was then corrected by taking the simple average of that prior and subsequent pixel. The only caveat is the first and last images in each year's time series were not able to be error checked since they are at the ends of the time series. But, this is why 34 time steps were downloaded per year when ultimately only the central 32 would be used for direct analysis. This simple error checking method was sufficient to improve the data in the vast majority of cases yet not so extreme as to be altering perfectly good pixels.

Next, the precipitation information was ingested to harmonize with the MODIS time series data. The native format of the rainfall data was markedly different from the MODIS since it was stored as vector points. So, an independent series of steps was taken to prepare the data. First, the data were grouped by 8-day intervals in the same time windows as for each of the 8-day MODIS composites and added together to derive an accumulated 8-day precipitation at each vector point. These 8-day accumulated precipitation files were each then projected to the MODIS sinusoidal projection. Finally, the values were interpolated from vector points to raster-based cells via an inverse distance weighting methodology. Ultimately, a four km grid (more closely 3706 m) with floating point values of rainfall measured in inches was a result. An output value of zero meant no rainfall during the 8-day period. The precipitation output grid was established so that the 1 km LST and 250 m NDVI data nested perfectly to the four km precipitation pixels. No temporal error checking or smoothing of the precipitation data was undertaken.

2.4. Data integration

With the 8-day composited NDVI, daytime and nighttime LST, and accumulated precipitation information in place alongside the reference masks, a final step was performed to determine the time series of county-level averages for both corn and soybeans. In summary, the 32 8-day periods, spanning from mid-February to late October, for each variable were "stacked" into one file and then intersected, by county, state, and crop mask for each of the years. Global mean statistics were then calculated for the pertinent pixels within each county for all of the 32 time periods. The result was a 32 value long vector containing county-level averages for NDVI, daytime and nighttime LST, and precipitation over only the corn and soybean areas. So in total, for each county and for each year 128 variables were derived.

After all the vectors were derived, they were combined and managed into a table in which each row in the database pertained to a particular year and county. A state identifier, county identifier, and year were affixed. Most important, the NASS published county-level average yield values (reported in the US customary bushels per acre for units) to be used as the independent variable were also added to each record. A few counties did not have any crop pixels to analyze so those were ultimately dropped. Counties that were shown to have crop pixels but no corresponding NASS yield were also dropped. Also, Arkansas and North Dakota counties were not included in the corn dataset since they were not part of the NASS Objective Yield survey region. Wisconsin was excluded from the soybeans list for the same reason. The resulting corn and soybean tables had 4824 and 4693 records, respectively.

3. Results

Having the corn and soybean databases built, simple exploration to determine what dependencies existed between each variable by date and the crop yields was undertaken. The Pearson product-moment correlation coefficient (r) was used for this and the results for the entire combined Corn Belt region over all of the years are shown in Fig. 2a for corn and Fig. 2b for soybeans. Both corn and soybeans showed similar relationships and patterns. For corn the strongest correlations came positively from NDVI peaking during the first week of August (r = 0.73). A simple linear regression of NDVI versus yield at that date gave a $R^2 = 0.54$. An exponential regression was tighter with an $R^2 = 0.63$. There was a continuous and smooth correlation relationship going forward or backward from that time but it progressively weakened. Inversely, daytime LST showed a good negative correlation to corn yields. It was also optimized during early August (r = −0.58) and steadily weakened in both time directions. A linear regression at the most negative correlation point yielded an $R^2 = 0.33$ while again an exponential regression was better with $R^2 = 0.53$. Soybean NDVI and daytime LST peak correlation responses were a week later and slightly less (r = 0.70, $R^2 = 0.49$ and r = −0.51, $R^2 = 0.26$). The scatterplots during the peak correlation time for NDVI and daytime LST, corn and soybeans, are shown in Fig. 3a, b, c and d.

Another noticeable, albeit weak feature, was the inverse relationship of NDVI to yields during spring. There was also some suggestion of positive spring daytime LST relating to better yields. The nighttime LST however showed very little correlation to crop yield regardless of the time of the season or crop but there was possibly some minor suggestion of being consistently negative throughout. Precipitation also showed little to no relationship with corn or soybean yield and with equal distribution along the zero line throughout the 32 period time window.

The corn and soybean datasets were further analyzed for NDVI and daytime LST correlation coefficient response by dividing them into subgroups. The first stratification was done by state. In summary, each state had a similar response to the average and varied mainly by degree. Thus, all state's corn peaked at or around early August with most having r greater than 0.70. Missouri, Indiana and Ohio were somewhat weaker with peak r = 0.53, 0.46 and 0.39, respectively. Likewise, the corn daytime LST relationships by state were all consistently inverse to yield. Those peak magnitudes were somewhat more widely spread ranging though and varied from −0.29 (Kansas) to −0.83 (South Dakota). For soybeans the state responses were also all similar in shape to the overall average. NDVI in Indiana had the lowest correlation but was still 0.46. Minnesota was the highest with r = 0.83 and all states peaked at or around the middle of August. Daytime LST for soybeans for each states was also consistent and inverted for each state and ranged more weakly from −0.22 (Nebraska) to −0.64 (Illinois and Missouri).

A second stratification methodology was performed but by year instead. Like by state, the results showed for both crops that each stratum followed a similar summer pattern to the combined average,
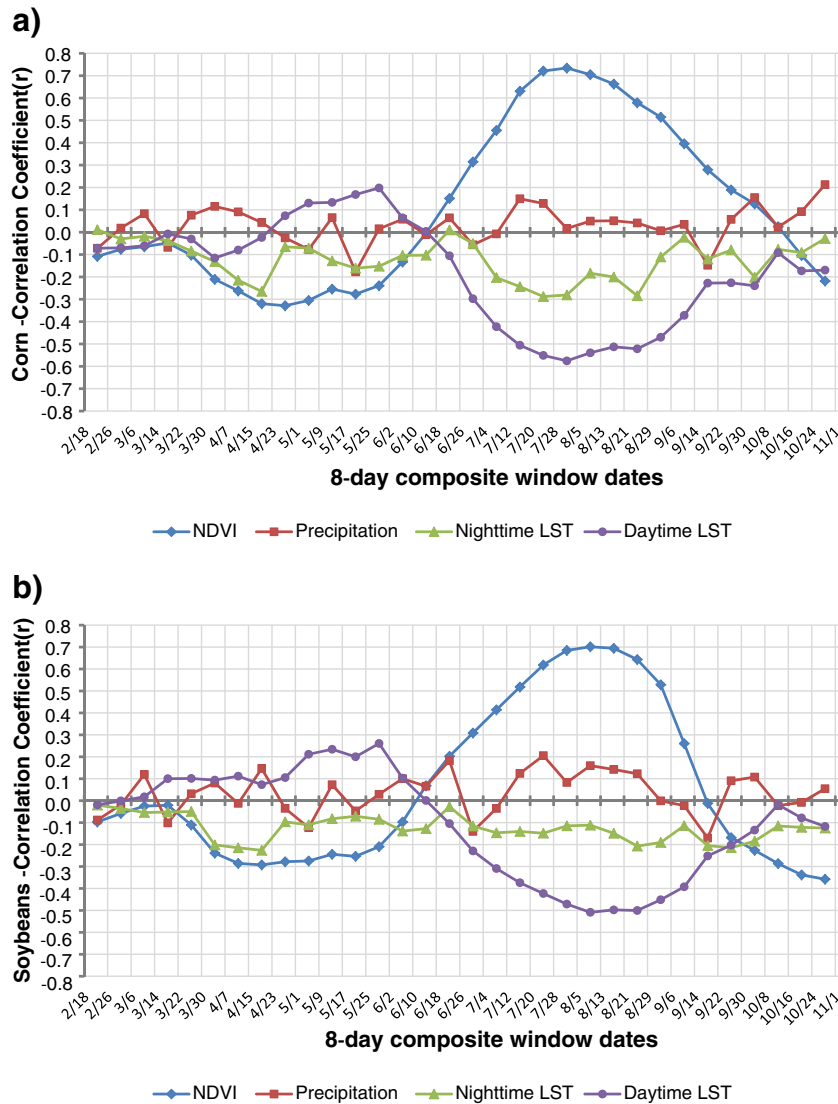
a)

b)

Fig. 2. Corn (a. upper) and soybeans (b. lower) remotely sensed variable correlation by date.

regardless of year. For corn the NDVI to yield correlation still peaked mid-summer for all strata on or just around the July 28–August 4th composite. Year 2011 saw the highest peak correlation with r = 0.81 and 2008 with the lowest equal to 0.64. Also closely consistent was the inverted response of daytime LST. However, in years 2009 and 2010 daytime LST peaked an 8-day period earlier during the July 20–27th and in 2008 it was a period later than the average. For soybeans the peaks all occurred in August and most centered with the overall average. For NDVI the peaks were fairly consistent and ranged from 0.67 for 2008 and 0.77 for 2010. For LST the relationships were also clustered in August but weaker with 2009 notably lower than the rest with only −0.24. Overall, the strength and weakness years for soybeans were the same as that for corn.

4. Discussion

4.1. Yield relationships to remote sensing variables

The positive mid-summer NDVI to yield relationship for both corn and soybeans was not surprising and expected given past research by others. The results here simply reinforce the fact and fine tune the knowledge that early August NDVI is the very best relationship for

understanding corn yields from MODIS NDVI and about a week later is the optimal time for understanding soybean yields. From the scatterplot there is suggestion that the values are saturating and the best regression fit is non-linear. The mild inverse relationship to NDVI shown around the late April period is perhaps more interesting and new knowledge.

More of an unknown going into this work was the relationship of the LST variables since most agronomy type research involves simple air temperature instead of land surface temperatures. Furthermore, the limited LST research has involved the morning overpass of the MODIS sensor from the Terra satellite instead of the afternoon overpass of Aqua which confounds the question. The early afternoon daytime LST was clearly inversely related mid-summer in this study for corn and just slightly less for soybeans. This is consistent with agronomy work that suggests that as temperatures get overly warm plant productivity begins to suffer, particularly for corn. It was expected that there would also be found a relationship between the nighttime LST and crop yields given, albeit anecdotal, information that plants prefer a cooling off period during the evening. However, nothing of the sort was proven here. With either the daytime or nighttime temperature case it needs to be noted that the relationships shown here would not necessarily reflect what might be found if using more traditional air temperatures instead. Also, since the LST information is natively at 1 km resolution
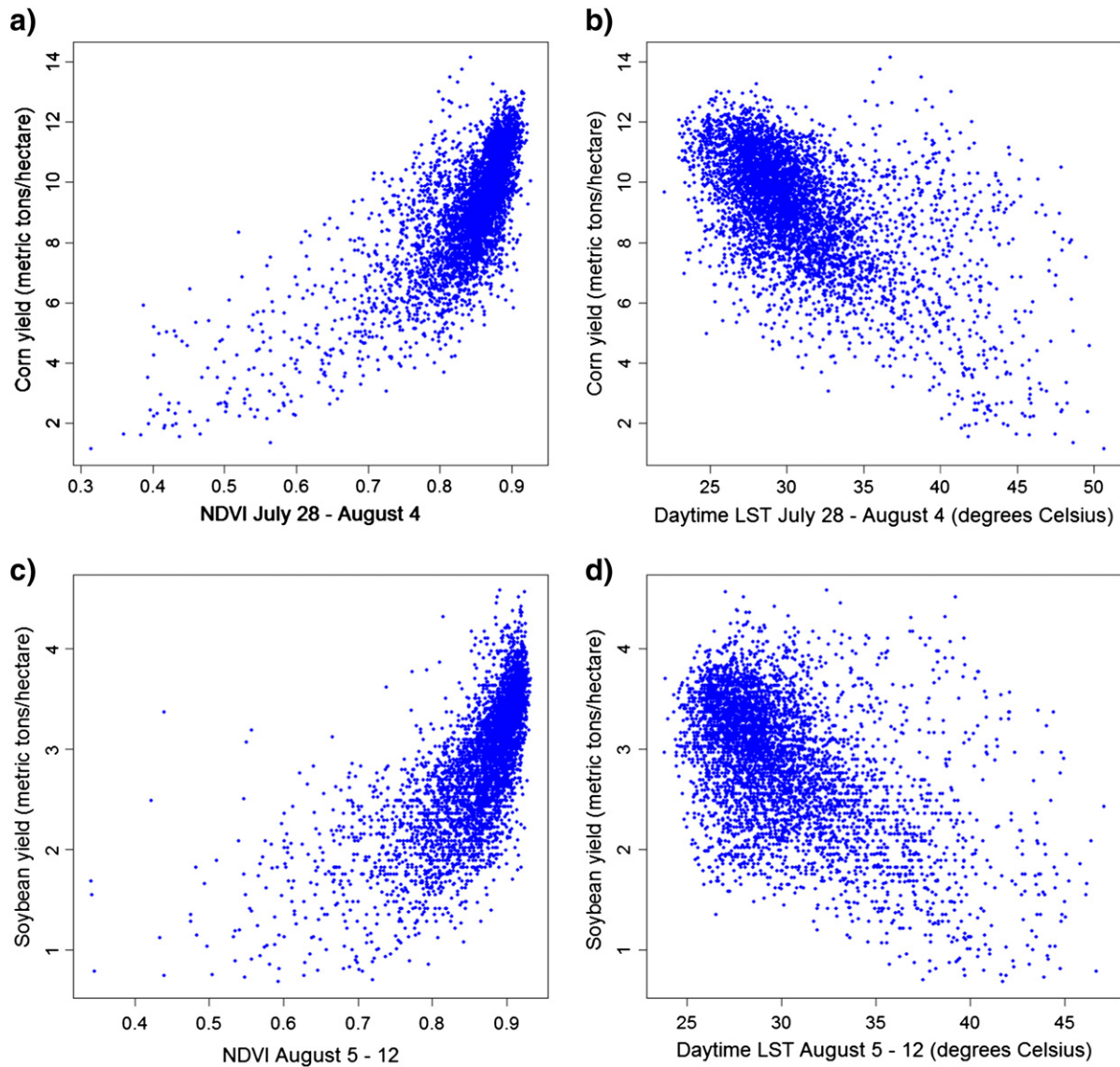
Fig. 3. Relationship between NDVI and daytime LST and crop yield at the point of peak of correlation. NDVI corn (a. upper left), daytime LST corn (b. upper right), NDVI soybeans (c. lower left), and daytime LST soybeans (d. lower right).

there is a potential that finer resolution data would reduce pixel mixing of the land cover types and thus further improve the relationships.

The results of the lack of any sort of relationship between precipitation and yields were more surprising and perplexing because it is conventional wisdom that crops need rainfall to thrive. The reasons that no relationships were found, regardless of time of year, could be many. First, there are certainly areas, particularly in Nebraska and Kansas, that are heavily reliant on irrigation and thus regardless whether rain occurs or not the corn and soybean crops can still have adequate moisture. Second, the NWS rainfall grids are effectively a model of the rainfall amounts derived from radar reflectivity data and may not in fact be overly accurate even though they appear reasonable. Third, it could just be that rainfall during the season is not as helpful to yield production as it might seem. Perhaps soil moisture is built up over a longer time period and at sufficient depths that the plants can still reach and utilize it even during times of minimal rainfall.

Finally, precipitation comes in different severities, types, and durations and likely some are more beneficial than others. For example, a slow but steady rainfall over a long period like days or weeks may have better impact on plant health and yields than that from a quick

but intense storm lasting only hours or minutes long. Those strongest of storms could also result in hail which if large enough outright could destroy the crop. The accumulated 8-day period was likely not time specific enough to capture the intensity attribute. On the flipside it could be speculated that longer accumulation periods are needed to develop a relationship to yields. This is something that was actually tested. The precipitation data were explored more intensively by recursively combining the original 32 values by multiples of 2 to further accumulate the data. In other words, a variety of 16-day, 32-day, 64-day and 128-day accumulated grids were formed. These types of snapshots provided something more akin to more commonly used monthly or seasonal totals. However, regardless of what periodic window was tried, no better relationships were found.

While the mid-season NDVI and daytime LST relationships are obviously strongest, the suggestion of a pre-season indicator of crop yield via the inverse NDVI in late April is particularly compelling. It is reinforced, even though weak, because it occurs over many dates which reduces the likelihood that it is just noise. Why this inverse pre-season relationship exists is not immediately intuitive but an explanation could be that low NDVI tends to occur in damp soils with little

vegetation present and thus those are the areas which will lean toward having better yields as the season progresses. Likewise, there may also be a very slight suggestion that warm spring land surface temperatures are helpful to crop yields. This could possibly be explained by reason knowing that warm spring temperatures provide earlier and better planting conditions which should improve the yield outcomes. These leading season indicators are tantalizing and could be seen as extremely important predictors for crop yields that potentially could be used, at least minimally, to nudge yield forecasts up or down from a trend line.

Finally, the stratifying of databases by state and by year was done to understand if there were dominant dependencies based on geography within the Corn Belt or certain years. In all cases the individual state or year levels followed the overall trend of being peaked for NDVI in early August, and inversed for daytime LST, for corn. Soybeans peaked a period or two later than but otherwise each state and year followed the over trends. These results suggest that the NDVI and daytime LST relationships to crop yield are similar throughout the widespread Corn Belt area in addition to over each of the six years studied. This ultimately bolsters the idea that a singular model across time and space integrating all of the county-level information is appropriate.

4.2. Forecast application of the remote sensing variables

Armed with a better understanding of what variables are correlated with crop yields, the next objective was to put them into practice for forecasting. Many types of modeling options exist but in this context it was decided to use a tool called Rulequest Cubist and allow it to "data mine" for the best ways relate the historical county-level input variables to crop yields. Cubist is a machine learning tool which autonomously derives best fitting piecewise linear models used to predict a continuous outcome variable (Quinlan, 1993). It has the ability to analyze the input data for "nearest neighbor" relationships and can run iteratively multiple times to form ensemble or "committee" models. Furthermore, it is clever in that is has a built in multi-fold cross-validation ability to self-test the models it has created. Finally, Cubist also provided output so a user can assess the rules set for importance of which dependent variables are used most. Decision tree type models have been used for remote sensing of agriculture applications but have tended to be toward crop type mapping (Chang, Hansen, Pittman, Carroll, & DiMiceli, 2007; Friedl & Brodley, 1997) versus crop yield mapping (Lobell et al., 2005).

Only NDVI and daytime LST were used going forward for modeling efforts since both nighttime LST and precipitation provided little correlation to yield regardless of the time of season or crop type. However, the entire time series of NDVI and daytime LST were used even though it was acknowledged that some dates, particularly early and late in the season, had likely no value. It was fully anticipated that with more information than from just the single optimal dates, as summarized earlier, the modeling efforts would improve due to the accumulation of information over the season, but it was not obvious exactly which dates to include to maximize the information gain. So, a liberal stance was taken to include all which was bolstered by the knowledge that decision tree analysis has the inherent ability to ignore, or prune, input data that is not useful, or might even be harmful, compared to other modeling methodologies.

By omitting nighttime LST and precipitation, the data tables where effectively halved in terms of the number of data fields before letting Cubist assess them. To reiterate, at this point each record contained variables for state, county, year, along with the seasonal 32 NDVI values and 32 daytime LST values. Corn and soybeans were run separately. Since Cubist has a few input parameters for tweaking its models, many trial-and-error runs of the software were performed in an attempt to determine its most effective and parsimonious use. Eventually, the method settled upon was letting Cubist use "instances" while limiting it to a single rule set with a "committee" model of five iterations. A 10-fold cross-validation was done to assess the model but this option ignored when the final rule-set model was built.

The county-level yield model results using all of the data gave a self-reported cross-fold validation R^2 of 0.93 for both corn and soybeans. The average absolute error at the county-level was self-reported 0.50 metric tons per hectare for corn and 0.17 for soybeans. Interestingly, allowing Cubist to derive a more sophisticated model with multiple rules instead of a single one showed little performance gain. Thus the simplest was used to make the results more interpretable and reduce the potential for model over-fitting. Using Cubist's instances option instead of rules mode did provide significantly better results. The iterative committee models also improved the results noticeably. The derived rules for both the soybean and corn models tended to rely most heavily on the mid-summer variables for both NDVI and daytime LST. This is reassuring and expected given that they were known to have the strongest correlations with yields. Some time periods and/or variable types were not used at all, particularly those early or late in the season. Finally, some variables were indeed used, such as during the green-up and senescence times, but had a small coefficient in the resulting model so are viewed more as tweaking the overall results rather than driving it.

The initial Cubist runs incorporated the whole years' worth of data. However, the corn and soybean models were also tested to understand the performance when withholding the data available towards year end. This would mimic the scenario when trying to forecast crop yields mid-season. So, R^2 assessments of the models were calculated by systematically removing end of year data. Surprisingly, the models still derived an $R^2 = 0.93$ when only provided with information up through the middle of August. Subsequent reduction of the end of year data finally started to impact the models but the decrease in the R^2 was not overly dramatic. For example, in early July the cross-validated R^2 for corn and soybeans 0.90 translating to absolute error averages about 20% greater than with having the full season's information. This ability to derive reasonable mid-season yields is heartening and suggests that even without a full years' worth of data there is yet enough information in the time series data to forecast accurately.

4.3. Validation of forecasting

Cubist's self-assessment tools are useful for understanding the performance of its models but it is not a true out-of-sample validation. So, to fully test the implications of the remote sensed corn and soybean forecasting rules, they were run against the MODIS data collected for 2012 and results compared to the NASS 2012 published county-level yields. This assessment was undertaken by first downloading the 2012 NDVI data from Terra and daytime LST data from Aqua and managed in the exact same manner as for the 2006–2011 datasets. Likewise, the 2012 CDL data were also gathered and processed to create the corn and soybean masks and then were used to extract the county-level NDVI and daytime LST averages for 2012.

Next, an executable, called Cubistsam.exe provided with the Cubist software, was run to easily implement the previously derived, full-season rules against the new predictor data. The final result was an estimated yield for each county based solely on the MODIS data. Of note, several 2012 images early and late in the season were very noisy beyond what could be error checked and thus were excluded as the model inputs. They were the first eight NDVI and last three NDVI periods and the first ten and last three daytime LSTs. Effectively what was left was the data spanning the middle of May through early October. Also, all modeling work here was completed prior to the official release of those numbers so as to avoid retrospect altering of the model to, perhaps, tune results.

Fig. 4a shows the relationship between the modeled yields for corn and what NASS ultimately published. Each point represents a county in the Corn Belt region. Fig. 4b shows the same for soybeans. For corn the relationship between the county-level predicted and published years gave a good $R^2 = 0.77$ and a root-mean-square error (RMSE) of 1.26 metric tons per hectare. The linear regression relationship was $y = 0.96x + 0.27$. Because the slope was close to one there not much

a)

y = 0.9623x + 0.2691
R² = 0.7737
RMSE = 1.2619

b)

y = 0.8098x + 0.4757
R² = 0.7128
RMSE = 0.4222
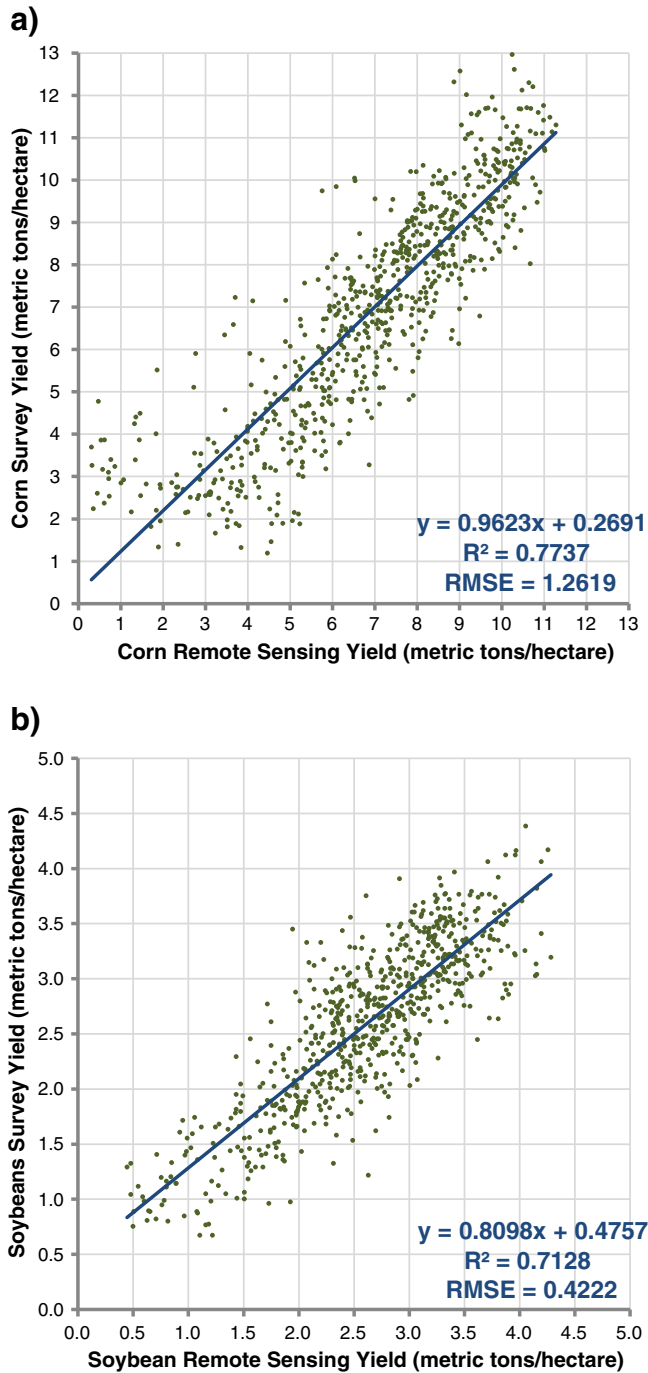
Fig. 4. 2012 corn (a. upper) and soybeans (b. lower) county-level remote sensing forecasted yields versus NASS published.

suggestion of bias to the estimates dependent on the magnitude of yield. However, there is some visual evidence by way of increased scatter that the model is less accurate on the low yielding counties. The soybean results were also good but not as strong as for corn. The soybean regression gave an $R^2 = 0.71$, RMSE = 0.42 metric tons per hectare, and a linear model equation of $y = 0.81x + 0.48$. The slope was not as close one as hoped and showed that remote sensing tended to underestimate on the low end and overestimate on the high. Because the R^2 is lower than it was for corn the soybean model is not as precise overall.

It was expected for both crops that the modeled results would have a tighter relationship to the final NASS yield given the self-reported R^2

performance of Cubist at 0.93. Three explanations to the decrease are hypothesized. The first, and perhaps the most obvious, is that Cubist is simply overestimating how powerful it is. The program is ultimately proprietary and complicated and it is not externally obvious how it is constructing its results. However, even if it is indeed overstating its usefulness it is still believed valid to use comparative results to understand the utility of different models' parameters and input variables. Secondly, it is recognized that the time series of input data variables which are input into Cubist are not purely independent. For example, NDVI from successive time period is usually similar and one would be a good predictor for a previous or successor value because it usually does not vary much from one 8-day period to the next. And in addition to the temporal autocorrelation there are likely spatial autocorrelation effects too. This overall lack of date independence is usually considered a violation of purely statistical type modeling and may be coming into play here as well.

Finally, and perhaps most importantly, 2012 turned out to be an outlier year in terms of US weather in the Corn Belt where the bulk of crop production exists. The area underwent what was considered the biggest agricultural drought since at least 1988 due to unusually prolonged heat and dryness which occurred directly in the middle of the summer. Final yields for corn were far under trend expectations across the region and most depressed in non-irrigated southern and western parts of the Corn Belt region. On average they were down about 16% from 2011 and 25% from the record high set in 2009 — this when usual year to year yield variability tends to be less than 10%. Ultimately, corn yields were the lowest since 1995 which is remarkable considering the improved seed varieties and planting and management techniques that have progressed since. Soybean yields were also negatively impacted in 2012 although to a lesser extent since they are more heat tolerant to begin with and late season rains helped spare the crop. They were down over 5% from 2011 and 10% from the record high set in 2009. Still 2012 was the lowest soybean yield since 2003. So, both the corn and soybean models here having been derived from relatively good to excellent 2006–2011 data had to extrapolate to a certain extent its results for 2012. It is believed that any modeling efforts which purely relied on contemporary datasets like this would have struggled to properly predict 2012 values given that they were unusually low on average.

A second, albeit subjective, validation was also performed to understand the utility of the remote sensing estimates. It was done by creating 2012 yield maps at the 250 m MODIS resolution. Fig. 5a shows the studied corn states and Fig. 5b the soybeans. They were created in ERDAS Imagine by calculating for each pixel a yield from the final decision tree model equation against the time series of the 2012 MODIS data. The original 30 m resolution 2012 CDLs were also integrated to better mask cartographically the pertinent field areas for each crop. The map results were particularly interesting because they begin to show the spatial heterogeneity that occurs even across a county. At the finest scale and in the most homogeneous of pixels one can attempt to estimate a field-level yield. Unfortunately, no field-level in situ information is known to be widely available in order to fully understand the accuracy of the maps. However, the patterns look reasonable and certainly highlight the known lower yielding areas, particularly those toward the south and west. The maps also reinforce the irrigated areas common throughout Nebraska which ultimately maintained good yields in 2012.

A final assessment of the practicality of the model was to recursively perform hindcasting for each of the years 2006–2011. In other words 2006–2010 and 2012 were used to predict for 2011, then 2006–2009 and 2011–2012 to predict for 2010, then 2006–2008 and 2010–2012 to predict for 2009, and so on for each of the individual years back to 2006. This is hypothetical because one would never have future data to predict for the current but it does allow for more out-of-sample validation trials to help increase the understanding of the model performance. It also provides evidence of the model accuracy beyond trying to predict for the anomalous 2012 drought year.

Results showing RMSE (metric tons per hectare), R^2, model regression line slope and intercept (again, metric tons per hectare) for each of the scenarios are shown in Table 1. For corn the best R^2 was for the year 2012 but RMSE was tied for worst in 2012, along with 2007. Overall, no years for corn were shown to be a real under or over outlier in terms model performance. A notable consistency though was the model slope always being less than one. This suggests that the model is always prone to underestimating on the low end and/or overestimating on the high. For soybeans the overall results and implications were similar with slopes also always less than one. RMSE and R^2 were also fairly stable across analysis years for soybeans although 2008 only had an R^2 of 0.47. Its RMSE of 0.41 was relatively average though so in general not of much concern. In terms of an average coefficient of determination the corn models outperformed the soybean models with $R^2 = 0.70$ versus 0.65. Average model slope was closer to one for corn as well (0.83 versus 0.77). In summary, the hindcasting results reinforce that the modeling efforts are strong for both crops but with corn being somewhat more robust.

4.4. Modeling for 2013

With the 2012 data now in hand, it was deemed constructive to test the overall models by including it into the pool of 2006–2011 data. It was believed that having the preponderance of low 2012 data points should only help for future forecasting efforts. Indeed once added, correlations to yields for the 2006–2012 increased the NDVI and daytime LST relationships further. For corn the NDVI peak correlation r was still found to be the first week of August and increased to 0.78 from 0.73. The inverted peak daytime LST relationship increased to an r of -0.60 from -0.58. Similarly the updated soybean relationship also increased with r = 0.73, from 0.70, for middle August NDVI and -0.56 from -0.51 for the same date daytime LST. Additionally, the modest pre-season inverse NDVI relationship strengthened for both crops going from -0.33 to -0.39 for corn and -0.29 to -0.34 for soybeans. But, any early season positive daytime LST to yield relationship became even weaker and thus it is truly not believed to be useful as a yield predictor for either crop. Finally, updating of the Cubist rules to include the 2012 data showed the self-reported full-season R^2 to increase to 0.95 for corn and 0.94 for soybeans. It was 0.93 for both previously. Since 2012 was an unusually low year for yields, the inclusion of that data has now resulted in even stronger models for the future.

5. Conclusions

In summary, this work was undertaken to substantiate the use of four timely and spatially detailed remotely sensed datasets for inclusion into county-level corn and soybean yield forecasting models for use across the US Corn Belt. Those datasets tested were NDVI, daytime LST and nighttime LST as collected from the space-borne MODIS sensors and precipitation as derived from the ground-based NWS Nexrad weather monitoring system. All datasets were assessed as 8-day composited products chosen as a compromise between data processing simplicity and retention of sufficient temporal detail. A large and robust cross-sectional panel of county-level data from NASS spanning the years 2006 through 2011 was used to verify the explanatory power throughout the spring and summer of each input dataset. Detailed land cover information helped isolate within the datasets the corn and soybean areas.

NDVI was found to be strongly and positively correlated with both corn and soybean yields throughout the summer and most dramatically in early August. Also, and unexpectedly, there was a suggestion that early season NDVI is negatively correlated to crop yields, albeit weakly. Daytime LST strongly mirrored NDVI and was found to be negatively correlated with yields for both crops throughout the summer season and peaking in early to mid-August. Nighttime LST and precipitation
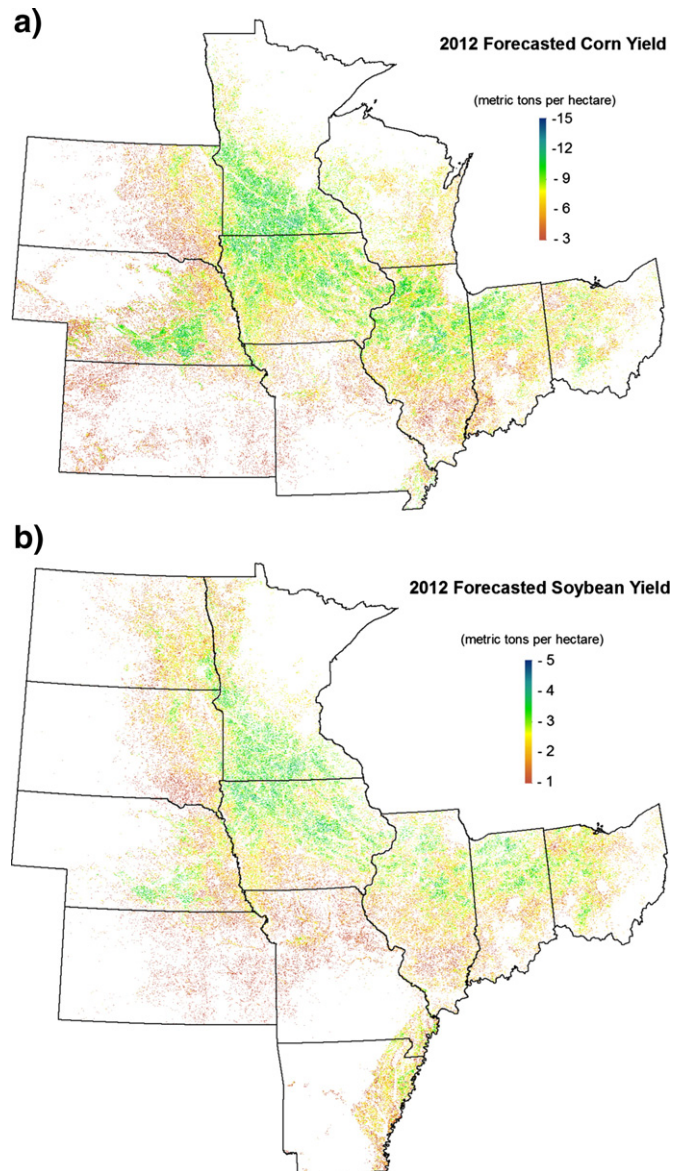
Fig. 5. Corn (a. upper) and soybean (b. lower) remotely sensed pixel-level forecast yield map.

were also investigated but showed to have no correlation to corn or soybean yields regardless of the seasonality. Not finding a precipitation to yield relationship was particularly surprising.

Using the full time series of NDVI and daytime LST variables for inclusion into a decision rule-based predictive model gave very good results ($R^2 = 0.93$) for both corn and soybeans as ascertained through the self-reported, cross-validation diagnostics of Cubist software. Dates of the strongest correlation in August were shown to most heavily used by the models as expected. Removal of data from mid-August and later degraded the model little suggesting that yield forecasting results in early August are nearly as robust as would be after the season has completed. Furthermore, forecasts even earlier in the season are possible with model precision dropping steadily into the spring to where the error is about double. This may still be useful in certain applications.

Out-of-sample validation of the model with a full season's worth of 2012 NDVI and daytime LST remote sensing input data against the 2012 published county estimates was calculated and showed corn performing good with an $R^2 = 0.77$ and soybeans, somewhat less so, with an $R^2 = 0.71$. These correlations were lower than was suggested by model cross validation but 2012 turned out to be an anomalous

Table 1
County-level corn (a. upper) and soybeans (b. lower) model result statistics for forecasting year 2012 and hindcasting of years 2006–2011.

Year	RMSE	R^2	m (slope)	b (intercept)	Type
Corn					
2006	1.08	0.71	0.89	0.98	Hindcast
2007	1.26	0.67	0.84	1.95	Hindcast
2008	0.99	0.66	0.76	2.06	Hindcast
2009	0.96	0.66	0.87	1.44	Hindcast
2010	1.07	0.66	0.74	2.57	Hindcast
2011	1.09	0.74	0.74	2.18	Hindcast
2012	1.26	0.77	0.96	0.27	Forecast
Soybeans					
2006	0.47	0.66	0.75	0.90	Hindcast
2007	0.38	0.69	0.81	0.62	Hindcast
2008	0.41	0.47	0.65	1.03	Hindcast
2009	0.38	0.59	0.73	0.71	Hindcast
2010	0.41	0.70	0.81	0.34	Hindcast
2011	0.41	0.71	0.81	0.52	Hindcast
2012	0.42	0.71	0.81	0.48	Forecast

yield year in which to forecast for due to a large and deep drought which lowered yields, some dramatically, across most of the study region. Thus, much of the model error may have been a result of trying to predict what would eventually become outlier events. Hindcasting the individual years 2006–2011 showed similar results to that of 2012. All the results taken together show the modeling efforts to be robust for both crops but with corn performing somewhat better on average.

While the focus on this paper has been the forecasting of county-level corn and soybean yields, the ability to derive broader estimates at the state- and national-level was also a strong motivation for this work. Those results were not shown but were indeed constructed by weighting each county by typical harvested area statistic to come up with a state and then regional estimate. Those estimates ultimately corresponded with even greater accuracy against the official 2012 NASS estimates. Finally and perhaps most importantly, was the strong desire to derive timely and objective early and mid-season yield estimates at these state- and national-levels. The Cubist modeling methodology employed appears flexible and robust even when a full season's worth of data is not yet available for input.

Moving forward, Collect 6 version of the MODIS data archive is currently under construction and should provide the final and most revised history for the dataset. It is said to be modestly improved in terms of data precision and stability over time (Wang et al., 2012) and thus hoped to fine tune the crop yield modeling further when released. In the end, after both MODIS sensors are finally retired, this type of work will undoubtedly transition to VIIRS. Exploration of that dataset will need to be undertaken when a reasonable history of data has been collected. The good LST relationships found from Aqua MODIS should be bolstered by the fact that the VIIRS instruments will also be in afternoon orbits. Analogous NDVI work utilizing the Aqua sensor instead of only Terra should also be tested before VIIRS is deemed operational to better understand if it shows markedly different results. Conversely, Terra LST could also be analyzed, for completeness, but since there is no future morning LST sensor there is little incentive to test.

Finally, this work focused only on corn and soybeans as grown in the Corn Belt region of the US and it was predicated on having detailed crop specific land cover information in which to best isolate the relevant areas from the input remote sensing datasets. In most settings this information does not exist with the same spatial detail or timeliness. However, anecdotal testing here suggests that using more generalized masks for cropland cover (Johnson, 2013) can still be effective for modeling corn and soybean yields although the model precision does decrease. Also, and regardless of the detail of land cover information available, it is not directly evident whether the results found here can be assumed to work or extrapolated to other crops or areas. But given the clear

seasonal relationship of both corn and soybeans to NDVI and daytime LST it is believed that, minimally, functionally related crops in similar style growing regimes should indeed respond similarly.

Note

Since the initial submission of this manuscript, soybean yield values were adjusted by NASS for the 2011 crop season. The average county-level change was up fractionally by 0.02 metric tons per hectare, or less than 1%. The analysis was rerun using the updated 2011 numbers but found to be insignificant on the results so the initial 2006–2011 dataset was kept as the basis for this work. NASS corn yield estimates did not change.

References

Allen, R., Hanuschak, G., & Craig, M. (2002). *Limited use of remotely sensed data for crop condition monitoring and crop yield forecasting in NASS.* (Retrieved June 12, 2013, from http://www.nass.usda.gov/research/avhrr/remoteuse.pdf)

Baruth, B., Royer, A., Klisch, A., & Genovese, G. (2008). The use of remote sensing within the MARS crop yield monitoring system of the European commission. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 37*, 935–939.

Basnyat, P., McConkey, B., Lafond, G. P., Moulin, A., & Pelcat, Y. (2004). Optimal time for remote sensing to relate to crop grain yield on the Canadian prairies. *Canadian Journal of Plant Science, 84*, 97–103.

Becker-Reshef, I., Vermote, E., Lindeman, M., & Justice, C. (2010). A generalized regression-based model for forecasting winter wheat yields in Kansas and Ukraine using MODIS data. *Remote Sensing of Environment, 114*, 1312–1323.

Benedetti, R., & Rossini, P. (1993). On the use of NDVI profiles as a tool for agricultural statistics: The case study of wheat yield estimate and forecast in Emilia Romagna. *Remote Sensing of Environment, 45*, 311–326.

Bolton, D. K., & Friedl, M.A. (2013). Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. *Agricultural and Forest Meteorology, 173*, 74–84.

Boryan, C., Yang, Z., Mueller, R., & Craig, M. (2011). Monitoring US agriculture: The US Department of Agriculture, National Agricultural Statistics Service, Cropland Data Layer program. *Geocarto International, 26,* 341–358.

Chang, J., Hansen, M. C., Pittman, K., Carroll, M., & DiMiceli, C. (2007). Corn and soybean mapping in the United States using MODIS time-series data. *Agronomy Journal, 99,* 1654–1664.

Dixon, B.L., Hollinger, S. E., Garcia, P., & Tirupattur, V. (1994). Estimating corn yield response models to predict impacts of climate change. *Journal of Agricultural and Resource Economics, 19,* 58–68.

Doraiswamy, P. C., Hatfield, J. L., Jackson, T. J., Akhmedov, B., Prueger, J., & Stern, A. (2004). Crop condition and yield simulations using Landsat and MODIS. *Remote Sensing of Environment, 92,* 548–559.

Doraiswamy, P. C., Moulin, S., Cook, P. W., & Stern, A. (2003). Crop yield assessment from remote sensing. *Photogrammetric Engineering and Remote Sensing, 69,* 665–674.

Doraiswamy, P. C., Sinclair, T. R., Hollinger, S., Akhmedov, B., Stern, A., & Prueger, J. (2005). Application of MODIS derived parameters for regional crop yield assessment. *Remote Sensing of Environment, 97,* 192–202.

Feddema, J. J., & Egbert, S. L. (2005). MODIS land surface temperature composite data and their relationships with climatic water budget factors in the central Great Plains. *International Journal of Remote Sensing, 26,* 1127–1144.

Fensholt, R., & Sandholt, I. (2005). Evaluation of MODIS and NOAA AVHRR vegetation indices with in situ measurements in a semi-arid environment. *International Journal of Remote Sensing, 26,* 2561–2594.

Ferencz, Cs, Bognár, P., Lichtenberger, J., Hamar, D., Tarcsai, Gy, Timár, G., et al. (2004). Crop yield estimation by satellite remote sensing. *International Journal of Remote Sensing, 25,* 4113–4149.

Friedl, M.A., & Brodley, C. E. (1997). Decision tree classification of land cover from remotely sensed data. *Remote Sensing of Environment, 61,* 399–409.

Funk, C., & Budde, M. E. (2009). Phenologically tuned MODIS NDVI-based production anomaly estimates for Zimbabwe. *Remote Sensing of Environment, 113,* 115–125.

Galford, G. L., Mustard, J. F., Melillo, J., Gendrin, A., Cerri, C. C., & Cerri, E. P. (2008). Wavelet analysis of MODIS time series to detect expansion and intensification of row-crop agriculture in Brazil. *Remote Sensing of Environment, 112,* 576–587.

Gallego, J., Carfagna, E., & Baruth, B. (2010). *Accuracy, objectivity and efficiency of remote sensing for agricultural statistics.* Agricultural Survey Methods. West Sussex, United Kingdom: John Wiley & Sons, Ltd, 193–211.

Gitelson, A. A., Peng, Y., Masek, J. G., Rundquist, D. C., Verma, S., Suyker, A., et al. (2012). Remote estimation of crop gross primary production with Landsat data. *Remote Sensing of Environment, 121,* 404–414.

Goetz, S. J. (1997). Multi-sensor analysis of NDVI, surface temperature and biophysical variables at a mixed grassland site. *International Journal of Remote Sensing, 18,* 71–94.

Guindin-Garcia, N., Gitelson, A. A., Arkebauer, T. J., 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.

Hatfield, J. L. (1983). Remote sensing estimators of potential and actual crop yield. *Remote Sensing of Environment, 13,* 301–311.

Hays, M. J., & Decker, W. L. (1996). Using NOAA AVHRR data to estimate maize production in the United States Corn Belt. *International Journal of Remote Sensing, 17,* 3189–3200.

Hird, J. N., & McDermid, G. J. (2008). Noise reduction of NDVI time series: An empirical comparison of selected techniques. *Remote Sensing of Environment, 113,* 248–258.

Huete, A., Didan, K., Miura, T., Rodriguez, E. P., Gao, X., & Ferreira, L. G. (2002). Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment, 83,* 195–213.

Johnson, D.M. (2013). A 2010 map estimate of annually tilled cropland within the conterminous United States. *Agricultural Systems, 114,* 95–105.

Johnson, D.M., & Mueller, R. (2010). The 2009 cropland data layer. *Photogrammetric Engineering and Remote Sensing, 76,* 1202–1205.

Jönsson, P., & Eklundh, L. (2004). TIMESAT — A program for analyzing time-series of satellite sensor data. *Computers & Geosciences, 30,* 833–845.

Justice, C. O., Townshend, J. R. G., Vermote, E. F., Masuoka, E., Wolfe, R. E., Saleous, N., et al. (2002). An overview of MODIS Land data processing and product status. *Remote Sensing of Environment, 83,* 3–15.

Kastens, J. H., Kastens, T. L., Kastens, D. L. A., Price, K. P., Martinko, E. A., & Lee, R. -Y. (2005). Image masking for crop yield forecasting using AVHRR NDVI time series imagery. *Remote Sensing of Environment, 99,* 341–356.

Kauffman, R. K., & Snell, S. (1997). A biophysical model of corn yield: Integrating climatic and social determinants. *American Journal of Agricultural Economics, 79,* 178–190.

Lee, B. -H., Phil Kenkel, P., & Brorsen, B. W. (2013). Pre-harvest forecasting of county wheat yield and wheat quality using weather information. *Agricultural and Forest Meteorology, 168,* 26–35.

Liu, J., Pattey, E., Miller, J. R., McNairn, H., Smith, A., & Hu, B. (2010). Estimating crop stresses, aboveground dry biomass and yield of corn using multi-temporal optical data combined with a radiation use efficiency model. *Remote Sensing of Environment, 114,* 1167–1177.

Lobell, D. B., Ortiz-Monasterio, J. I., Asner, G. P., Naylor, R. L., & Falcon, W. P. (2005). Combining field surveys, remote sensing, and regression trees to understand yield variations in an irrigated wheat landscape. *Agronomy Journal, 97,* 241–249.

Maselli, F., & Rembold, F. (2001). Analysis of GAC NDVI data for cropland identification and yield forecasting in Mediterranean African countries. *Photogrammetric Engineering and Remote Sensing, 67,* 593–602.

Mildrexler, D. J., Zhao, M., & Running, S. W. (2011). A global comparison between station air temperatures and MODIS land surface temperatures reveals the cooling role of forests. *Journal of Geophysical Research, 116,* G03025.

Mkhabela, M. S., Bullock, P., Raj, S., Wang, S., & Yang, Y. (2011). Crop yield forecasting on the Canadian prairies using MODIS NDVI data. *Agricultural and Forest Meteorology, 151,* 385–393.

Mkhabela, M. S., Mkhabela, M. S., & Mashinini, N. N. (2005). Early maize yield forecasting in the four agro-ecological regions of Swaziland using NDVI data derived from NOAA's AVHRR. *Agricultural and Forest Meteorology, 129,* 1–9.

Moody, A., & Johnson, D.M. (2001). Land-surface phenologies from AVHRR using the discrete Fourier transform. *Remote Sensing of Environment, 75,* 305–323.

Patel, N. R., Parida, B. R., Venus, V., Saha, S. K., & Dadhwal, V. K. (2012). Analysis of agricultural drought using vegetation temperature condition index (VTCI) from Terra/MODIS satellite data. *Environmental Monitoring and Assessment, 184,* 7153–7163.

Prasad, A. K., Chai, L., Singh, R. P., & Kafatos, M. (2006). Crop yield estimation model for Iowa using remote sensing and surface parameters. *International Journal of Applied Earth Observation and Geoinformation, 8,* 26–33.

Prihodko, L., & Goward, S. N. (1997). Estimation of air temperature from remotely sensed surface observations. *Remote Sensing of Environment, 60,* 335–346.

Quinlan, J. R. (1993). Combining instance-based and model-based learning. *Proceedings of the Tenth International Conference on Machine Learning* (pp. 236–243). San Mateo, California: Morgan Kaufmann Publishers.

Reeves, M. C., Zhao, M., & Running, S. W. (2005). Usefulness of limits of MODIS GPP for estimating wheat yield. *International Journal of Remote Sensing, 26,* 1403–1421.

Reynolds, C. A., Yitayew, M., Slack, D. C., Hutchinson, C. F., Huete, A., & Petersen, M. S. (2000). Estimating crop yields and production by integrating the FAO crop specific water balance model with real-time satellite data and ground-based ancillary data. *International Journal of Remote Sensing, 21,* 3487–3508.

Rojas, O. (2007). Operational maize yield model development and validation based on remote sensing and agro-meteorological data in Kenya. *International Journal of Remote Sensing, 28,* 3775–3793.

Sakamoto, T., Gitelson, A. A., & Arkebauer, T. J. (2013). MODIS-based corn grain yield estimation model incorporating crop phenology information. *Remote Sensing of Environment, 131,* 215–231.

Salazar, L., Kogan, F., & Roytman, L. (2007). Use of remote sensing data for estimation of winter wheat yield in the United States. *International Journal of Remote Sensing, 28,* 3795–3811.

Schlenker, W., & Roberts, M. J. (2009). Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. *Proceedings of the National Academy of Sciences of the United States of America, 106,* 15594–15598.

Seo, D. -J. (1998). Real-time estimation of rainfall fields using radar rainfall and rain gauge data. *Journal of Hydrology, 208,* 37–52.

Seo, D. -J., Breidenbach, J., & Johnson, E. (1999). Real-time estimation of mean field bias in radar rainfall data. *Journal of Hydrology, 223,* 131–147.

Shanahan, J. F., Schepers, J. S., Francis, D.D., Varvel, G. E., Wilhelm, W. W., Tringe, J. M., et al. (2001). Use of remote-sensing imagery to estimate corn grain yield. *Agronomy Journal, 93,* 583–589.

Smith, W. J. (1914). The effect of weather upon the yield of corn. *Monthly Weather Review, 42,* 78–93.

Tannura, M.A., Irwin, S. H., & Good, D. L. (2008). Weather, technology, and corn and soybean yields in the U.S. Corn Belt. *Marketing and outlook research report 2008-01* Department of Agricultural and Consumer Economics, University of Illinois at Urbana-Champaign (Retrieved June 12, 2013, from http://www.farmdoc.illinois. edu/marketing/morr/morr_08-01/morr_08-01.pdf)

Tucker, C. J. (1979). Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment, 8,* 127–150.

United States Department of Agriculture (2010). Field crop usual planting and harvesting dates. *Agricultural Handbook Number 628* (Retrieved June 12, 2013, from http:// usda01.library.cornell.edu/usda/current/planting/planting-10-29-2010.pdf)

United States Department of Agriculture (2012). The yield forecasting program of NASS. *SMB staff report number SMB 12-01* (Retrieved June 12, 2013, from http:// www.nass.usda.gov/Publications/Methodology_and_Data_Quality/Advanced_Topics/ Yield%20Forecasting%20Program%20of%20NASS.pdf)

Vermote, E. F., Kotchenova, S. Y., & Ray, J. P. (2011). *MODIS surface reflectance user's guide.* (Retrieved June 12, 2013, from http://modis-sr.ltdri.org/products/ MOD09_UserGuide_v1_3.pdf)

Viña, A., Gitelson, A. A., Rundquist, D. C., Keydan, G., Leavitt, B., & Schepers, J. (2004). Monitoring maize (*Zea mays* L.) phenology with remote sensing. *Agronomy Journal, 96,* 1139–1147.

Wall, L., Larocque, D., & Léger, P. -M. (2008). The early explanatory power of NDVI in crop yield modelling. *International Journal of Remote Sensing, 29,* 2211–2225.

Wallace, H. A. (1920). Mathematical inquiry into the effect of weather on corn yield in the eight Corn Belt states. *Monthly Weather Review, 48,* 439–446.

Wan, Z. (2007). *Collection-5 MODIS land surface temperature products users' guide.* (Retrieved June 12, 2013, from http://www.icess.ucsb.edu/modis/LstUsrGuide/ MODIS_LST_products_Users_guide_C5.pdf)

Wan, Z. (2008). New refinements and validation of the MODIS Land-Surface Temperature/Emissivity products. *Remote Sensing of Environment, 112,* 59–74.

Wan, Z., Wang, P., & Li, X. (2004). Using MODIS land surface temperature and normalized difference vegetation index products for monitoring drought in the southern Great Plains, USA. *International Journal of Remote Sensing, 25,* 61–72.

Wang, D., Morton, D., Masek, J., Wu, A., Nagol, J., Xiong, X., et al. (2012). Impact of sensor degradation on the MODIS NDVI time series. *Remote Sensing of Environment, 119,* 55–61.