2020

The role of topography, soil, and remotely sensed vegetation condition towards predicting crop yield

Trenton E. Franz
University of Nebraska-Lincoln, trenton.franz@unl.edu

Sayli Pokal
University of Nebraska-Lincoln, spokal@huskers.unl.edu

Justin P. Gibson
CropMetrics, North Bend, NE

Yuzhen Zhou
University of Nebraska - Lincoln, yuzhenzhou@unl.edu

Hamed Gholizadeh
University of Nebraska-Lincoln, hamed.gholizadeh@unl.edu

See next page for additional authors

The role of topography, soil, and remotely sensed vegetation condition towards predicting crop yield

Trenton E. Franz,[1] Sayli Pokal,[2] Justin P. Gibson,[1,3]
Yuzhen Zhou,[2] Hamed Gholizadeh,[1,4]
Fatima Amor Tenorio,[5] Daran Rudnick,[6]
Derek Heeren,[6] Matthew McCabe,[7]
Matteo Ziliani,[7] Zhenong Jin,[8]
Kaiyu Guan,[9] Ming Pan,[10] John Gates,[3]
and Brian Wardlow[1]

1 School of Natural Resources, University of Nebraska-Lincoln, Lincoln, NE 68583, USA
2 Department of Statistics, University of Nebraska-Lincoln, Lincoln, NE 68583, USA
3 CropMetrics, North Bend, NE 68649, USA
4 Center for Applications of Remote Sensing, Department of Geography, Oklahoma State University, Stillwater, OK 74078, USA
5 Department of Agronomy and Horticulture, University of Nebraska-Lincoln, Lincoln, NE 68583, USA
6 Biological Systems Engineering, University of Nebraska-Lincoln, Lincoln, NE 68583, USA
7 Environmental Science and Engineering, King Abdullah University of Science and Technology, Thuwal, 23955, Saudi Arabia
8 Department of Bioproducts and Biosystems Engineering, University of Minnesota - Twin Cities, St. Paul, MN 55108, USA
9 Department of Natural Resources and Environmental Sciences, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA
10 Department of Civil and Environmental Engineering, Princeton University, Princeton, NJ 08544, USA

Corresponding author — T. E. Franz, tfranz2@unl.edu

Published in *Field Crops Research* 252 (2020) 107788
DOI: 10.1016/j.fcr.2020.107788

Submitted 18 September 2019; revised 11 March 2020; accepted 23 March 2020.

Abstract

Foreknowledge of the spatiotemporal drivers of crop yield would provide a valuable source of information to optimize on-farm inputs and maximize profitability. In recent years, an abundance of spatial data providing information on soils, topography, and vegetation condition have become available from both proximal and remote sensing platforms. Given the wide range of data costs (between USD $0–50/ha), it is important to understand where often limited financial resources should be directed to optimize field production. Two key questions arise. First, will these data actually aid in better fine-resolution yield prediction to help optimize crop management and farm economics? Second, what level of priority should stakeholders commit to in order to obtain these data? Before fully addressing these questions a remaining challenge is the complex nature of spatiotemporal yield variation. Here, a methodological framework is presented to separate the spatial and temporal components of crop yield variation at the subfield level. The framework can also be used to quantify the benefits of different data types on the predicted crop yield as well to better understand the connection of that data to underlying mechanisms controlling yield. Here, fine-resolution (10 m) datasets were assembled for eight 64 ha field sites, spanning a range of climatic, topographic, and soil conditions across Nebraska. Using Empirical Orthogonal Function (EOF) analysis, we found the first axis of variation contained 60–85 % of the explained variance from any particular field, thus greatly reducing the dimensionality of the problem. Using Multiple Linear Regression (MLR) and Random Forest (RF) approaches, we quantified that location within the field had the largest relative importance for modeling crop yield patterns. Secondary factors included a combination of vegetation condition, soil water content, and topography. With respect to predicting spatiotemporal crop yield patterns, we found the RF approach (prediction RMSE of 0.2–0.4 Mg/ha for maize) was superior to MLR (0.3–0.8 Mg/ha). While not directly comparable to MLR and RF the EOF approach had relatively low error (0.5–1.7 Mg/ha) and is intriguing as it requires few calibration parameters (2–6 used here) and utilizes the climate-based aridity index, allowing for pragmatic long-term predictions of subfield crop yield.

Keywords: Maize and soybean, Yield, Spatiotemporal, Statistics, Remote sensing

1. Introduction

Understanding the spatiotemporal patterns of crop yield, along with our inability to accurately predict those patterns with a reasonable lead time, remain key limitations in making management decisions to optimize limited resources (e.g., water, energy, and fertilizer) while

maximizing on-farm profitability (Maestrini and Basso, 2018 and Gibson et al., 2019). In recent years, there has been a rapid rise in the types and scales of available remote sensing observations, with a number of new ground, unmanned and manned aircraft and satellite based platforms suitable for field-scale applications that fill this knowledge gap (Azzari et al., 2017; Bolton and Friedl, 2013; Mancini et al., 2013; McCabe et al., 2017a,b; Manfreda et al., 2018; Ziliani et al., 2018). Collectively, remote sensing observations from these data platforms provide a suite of variables that can describe topography, soils, vegetation condition, and qualitative crop health difference, all of which can be used as inputs to parameterize relatively simple (e.g. FAO56; Allen et al., 1998) and more complicated crop models (e.g. AquaCrop, Hybrid-Maize, DSSAT, APSIM) (Foster et al., 2017; Yang et al., 2013; Jones et al., 2003; Holzworth et al., 2014). Although the latter have been significantly improved within the last three decades (Jin et al., 2018), a major limitation of crop models remains their inability to be discretized spatially and provide information on spatial variations of actual within field condition (Kasampalis et al., 2018). It is expected that combining these new remote sensing and in-situ sensing technologies with crop models will lead to improved crop yield predictions. For example, several studies have combined statistical techniques (i.e. both linear and nonlinear approaches) with remote sensing to make yield predictions in the Midwest USA (Bolton and Friedl, 2013; Peng et al., 2018; Li et al., 2019), West Africa (Leroux et al., 2019; Gibon et al., 2018) and East Africa (Burke and Lobell, 2017) at field to regional scales. However, the cost for acquiring each data layer, as well as its spatial and temporal resolution and latency can be highly variable (McCabe et al., 2017a,b), so determining the relative cost-to-benefit ratio of these data for improving crop management is a key determinant in their utility.

While the range of sensing possibilities have expanded, the sensors that measure these geophysical, biophysical and biochemical properties utilize a range of wavelengths of the electromagnetic spectrum, making interpretation to useful agronomic information challenging (Maestrini and Basso, 2018; Haghverdi et al., 2015; Finkenbiner et al., 2019). For example, multispectral sensors onboard airborne and satellite platforms collect data in the visible and near infrared spectrum that can be used to describe various aspects of vegetation condition typically through the calculation of spectral-based vegetation indices

(e.g. normalized vegetation difference index, soil adjusted vegetation index, green chlorophyll content, pigment based indices, see Vina et al., 2011). However, there remains limited guidance on which specific index may work best for any particular case. Recent machine learning-related research have sought to explore this topic, indicating that unique combinations of many vegetation indices enhance prediction of key biophysical indicators (Shah et al., 2019) relative to using a single index. Ground based sensors are able to capture a wider range of the electromagnetic spectrum ($O{\sim}10^{3}$ m for broadband radio to $O{\sim}10^{-12}$ m for gamma rays) that can go beyond just sensing the vegetation canopy and penetrate deeper into the soil. These sensors can provide information about soil texture and soil water content (SWC) (see Robinson et al., 2008; Binley et al., 2015; Desilets et al., 2010; Finkenbiner et al., 2019) throughout the root zone and at spatial scales (tens of meters) that are more pragmatic for agricultural applications. However, the conversion and interpretation of the geophysical observation (e.g. bulk electrical conductivity towards predicting soil texture and neutron and gamma ray intensity towards predicting SWC) remains challenging and somewhat disconnected from agronomic decision making. In general, the scale difference between observations of state variables from remote and proximal sensing and the physically-based modeling parameters (e.g. prediction of saturated hydraulic conductivity, Binley et al., 1989) that control fluxes remains a challenge (Peters-Lidard et al., 2017).

For all these data sources (and available modeling approaches discussed above), several important unanswered questions remain. First, will these data actually aid in better fine-resolution yield prediction to help optimize crop management and farm economics? Second, what level of priority should the producer, farm manager, private consultant and/or state and federal agencies commit to in order to obtain these data (i.e., the value proposition)? As the answer to these questions requires information on economic costs (i.e. price of data, capacity to process data, cost to transform them into a decision making platform for producers, etc.), here, as a first step, we aim instead to quantify the benefits of the data on understanding and predicting subfield crop yield. To do this, we have compiled a unique fine-resolution (10 m) crop yield dataset from eight 64 ha study sites that span a climatic gradient across the state of Nebraska. At each site we have assembled data layers related to topography (freely available fine-resolution Light Detection and Ranging (LiDAR) system), soil texture

and soil water content (ground based hydrogeophysical mapping), and vegetation condition (freely-available Landsat satellite image archive). In order to separate the spatial and temporal components of crop yield we use the approach of Empirical Orthogonal Functions (EOF), which has been used in other scientific disciplines (e.g. Perry and Niemann, 2007) but limited use in agricultural research to our knowledge. The separation of space and time is a key advance of this work in better understanding crop yield patterns. Next we are able to explore the contribution of each covariate to understanding yield patterns by using common statistical approaches like Multivariate Linear Regression and Machine Learning (i.e. Random Forest). Importantly, we seek to develop a statistical framework that balances generality and parsimony for making fine-resolution predictions of crop yield.

2. Materials and methods

2.1. Study sites

A total of eight approximately 64 ha study sites were selected across a climatic and irrigation gradient within the state of Nebraska (**Fig. 1** and **Table 1**). Sites were identified based on the availability of historic crop yield maps and corresponding hydrogeophysical surveys that were compiled to generate a 10m resolution product that detailed soil, topographic, and vegetation condition. Seven of the study sites were irrigated by overhead sprinklers from center-pivots, with the remaining site being rainfed (S8). All sites primarily grow maize (*Zea mays* L.), with most of the eastern sites rotating in soybeans (*Glycine max*) in alternating years. Planting typically occurs in late April to early May depending on field and weather conditions. Irrigation generally starts in mid-June and continues through early September, depending on crop development and weather.

2.2. Data sources and processing

2.2.1. Climate

The general climate and monthly crop water use of Nebraska are detailed in Sharma and Irmak (2012a,b). Here, weather data for each study site was obtained from the nearest (< 20 km) Nebraska Mesonet

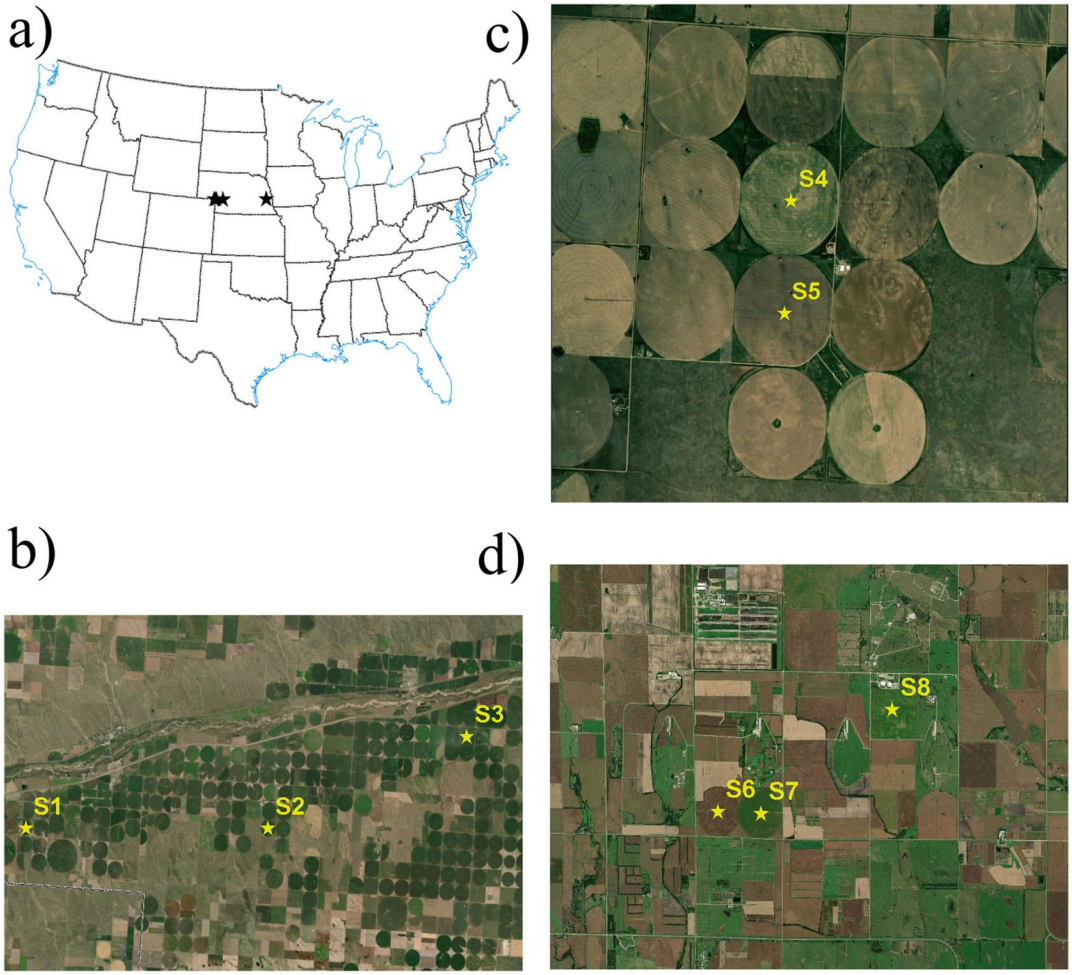

Fig. 1. a) Location of eight-64 ha study sites in the state of Nebraska, USA, in clusters in the b) west, c) central, and d) eastern part of the state. See Table 1 for a description of each site, Table 2 for a list of available datasets, and Supplementary Table S1 for 10m resolution QA/QC data.

station, formerly referred to as the Automated Data Weather Network (see https://mesonet.unl.edu/fordataaccessandQA/QCprocedures). **Table 1** provides the station name, years of record used for the analysis, and average growing season (May to September) precipitation (P) and reference evapotranspiration (ET_o) using a modified Penman approach (see https://hprcc.unl.edu/awdn.php#fordetails). With respect to quantifying year to year climate conditions we will use the growing season aridity index = P/ET_o. The aridity index (Budyko, 1974) is a simple yet powerful ratio that is capable of representing biogeographical distributions of vegetation across the globe (Kerkhoff et al., 2004).

Table 1 Summary data of each study site including: location, land use, recent climate data from Nebraska Mesonet, and net irrigation requirements for maize and soybean (Sharma and Irmak, 2012a,b).

Study Site	Latitude	Longitude	Dominant	Ownership Landuse	Mesonet Station (years used)	Precipitation Total May-Sept. (mm)	Potential ET May-Sept. (mm)	Net Irrigation Requirement for Maize (mm/yr)	Net Irrigation Requirement for Soybeans (mm/yr)
1	41.0043	−102.1065	Irrigated maize	Private Producer	Big Springs 8NE (2008–2017)	281.89	1143.45	525.00	425.00
2	41.0287	−101.9716	Irrigated maize	Research and Extension Farm	Brule 6SW (2008–2017)	299.74	1098.81	525.00	425.00
3	41.0728	−101.8498	Irrigated maize	Private Producer	Brule 6SW (2008–2017)	299.74	1098.81	525.00	425.00
4	41.0654	−101.1027	Irrigated maize and soybean rotation	Private Producer	Dickens 9 N (2008–2017)	324.27	1010.97	375.00	300.00
5	41.0583	−101.1027	Irrigated maize and soybean rotation	Private Producer	Dickens 9 N (2008–2017)	324.27	1010.97	375.00	300.00
6	41.1791	−96.4400	Irrigated maize	Research and Extension Farm	Ithaca 3E (1982–2017)	449.39	843.35	200.00	150.00
7	41.1791	−96.4400	Irrigated maize and soybean rotation	Research and Extension Farm	Ithaca 3E (1982–2017)	449.39	843.35	200.00	150.00
8	41.1791	−96.4400	Rainfed maize and soybean rotation	Research and Extension Farm	Ithaca 3E (1982–2017)	449.39	843.35	200.00	150.00

In addition, long-term forecasts (months instead of weeks) of aridity index are more pragmatic than daily weather given the current state of weather forecasting skill. For each year where crop yield information was available (see Supplementary Table S1 for each site year), the daily Mesonet data were downloaded and cumulated to seasonal totals for use in this research. The western sites (S1–3) are hot and dry, with a growing season P of approximately 300 mm, ET_o of around 1100 mm, daily temperature of 19 °C, and relative humidity of 50 %. The central sites (S4–5) are also hot and dry, with a growing season P of 325 mm, ET_o of 1000 mm, daily temperature of 19 °C, and relative humidity of 60 %. The eastern sites (S6–8) are hot and humid, with a growing season P of 450 mm, ET_o of 850 mm, daily temperature of 19 °C, and relative humidity of 70 %. The net irrigation requirement for maize and soybean generally follows the growing season P and ET_o, with net irrigation values in the east more than double those required in the west (i.e. reflecting the variability in rainfall along that gradient; see Table 1 for more details).

2.2.2. Topography

The topographic data were collected for each site using a fine-resolution 1m Digital Elevation Model (DEM) available for the state of Nebraska from LiDAR surveys provided by the United States Department of Agriculture Natural Resource Conservation Service (data processed on 9 July 2019). For each of the eight study sites, the DEM was clipped as a GeoTIFF file format to the field boundary and aggregated to the same 10m resolution grid as the other datasets using MATLAB R2018b (MathWorks, Natick, Massachusetts, USA) and a linear interpolation. Supplementary Table S1 provides the 10m dataset for each site and **Fig. 2** illustrates the distribution of relative elevation. The box and whisker plots illustrate the sites where large topographic relief exists (S2, 4, 5).

2.2.3. Soils

The soil texture and near surface soil water content (SWC) data were collected with a series of hydrogeophysical surveys (see Supplementary Table S1 for survey dates and all QA/QC data). Soil texture information is inferred from electromagnetic induction (EMI) surveys

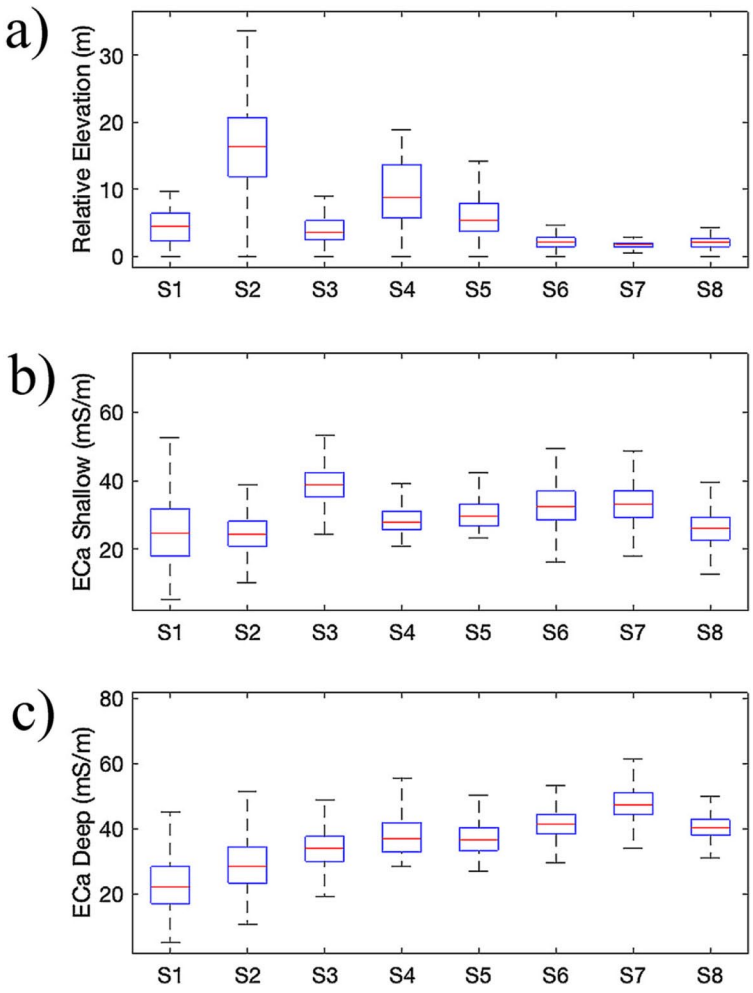

Fig. 2. Box and whisker plot of each study site's 10m resolution a) relative elevation, an b) Bulk electrical conductivity (ECa) shallow survey, and an c) ECa deep survey dataset to illustrate relative in-field variation between study sites. The red line is the median, the top and bottom of the boxes are the 25 and 75 % quantile, the top and bottom whiskers are the minimum and maximum.

measuring bulk electrical conductivity, EC_a (see for example Samouelian et al., 2005; Abdu et al., 2008), while near surface SWC is inferred from low energy neutron intensity surveys (Desilets et al., 2010; Zreda et al., 2012). The hydrogeophysical surveys were collected at each of the study sites using an all-terrain vehicle (ATV). Surveys were carried out between the spring of 2015 through 2018 when field access was available (see Finkenbiner et al., 2019 and Gibson and Franz, 2018 for additional site information of surveys). EC_a maps were collected

using a Dualem-21S EMI sensor (DUALEM, Milton, Canada). Four simultaneous depths of EC_a can be measured with the sensor given the dual-geometry receivers at separations of 1 and 2.1m from the transmitter. For this analysis, only one shallow (approx. 0–1 m) and one deep (approx. 0–3.2 m) survey sensor data were used. Measurements of EC_a were taken every second while being towed behind an ATV (on a plastic sled) traveling at speeds of around 8–15 km/h, with spacing every approx. 8 m, and surveys taking between 75–90 min. A Hemisphere GPS XF101 DGPS (Juniper Systems, Inc., Logan, UT) unit recorded the location of each measurement. Following basic QA/QC of the EC_a data (Franz et al., 2011), a spatial map of EC_a with 10m resolution was created from the more than 5000 observations using linear interpolation. We note that temporal differences in EC_a maps stem from changes in soil temperature, SWC, and soil solute concentration (Robinson et al., 2008). SWC has been shown to account for approximately 50 % of this variability (Brevik et al., 2006) between surveys. Removal of this confounding factor for isolating soil texture will be addressed in further detail in Sections 2.3.1 and 3.1. With respect to soil texture, low EC_a values generally indicate sandier soils, whereas higher EC_a values indicate higher silt and clay material. Fig. 2 illustrates the spatial distribution of a single bulk electrical conductivity survey collected from each site (presented here as ECa Shallow in mS/m), illustrating the degree of soil texture variation that exists within and between the study locations.

Near surface SWC data was collected using the same ATV setup for bulk conductivity, but deploying a mobile Cosmic-Ray Neutron Sensor (CRNS; Hydroinnova LLC, Albuquerque, NM; see Franz et al., 2015; Finkenbiner et al., 2019 for details of the instrument setup). The mobile CRNS used here records epithermal neutron intensity (from 0.25 to 1000 eV; Andreasen et al., 2017) integrated over one minute counting intervals. The change in epithermal neutron intensity is inversely correlated to the mass of hydrogen in the measurement volume (Zreda et al., 2012). The authors note that SWC changes are by far the largest driver of change in hydrogen mass (McJannet et al., 2014). Numerous validation studies across the globe (Bogena et al., 2013; Franz et al., 2012, 2016, Hawdon et al., 2014; Schron et al., 2018) have shown the CRNS to have area-average measurement accuracies of less than 0.03 cm^3/cm^{-3} root mean square errors (RMSE), when compared against a range of industry standard SWC point scale probes. The measurement

volume of the instrument is roughly a disk, with a 130–250m radius and a penetration depth of 0.15–0.40m (Kohli et al., 2015; Schron et al., 2017), depending on local conditions (e.g. elevation, water vapor, SWC etc.). For simplicity, a constant penetration depth of 0.3m was assumed for all surveys. In order to produce a 10m resolution SWC map (approx. 75–90 observations) we followed the same image sharpening procedure described in Gibson and Franz (2018). The image sharpening technique (i.e. drop-in-the-bucket inverse distance weighting) uses the series of overlapping coarse images to construct a finer resolution product and is routinely used in remote sensing applications with this type of data (see Chan et al., 2014 for details).

2.2.4. Vegetation condition

Vegetation condition information was inferred using a peak growing season, Green Chlorophyll Vegetation Index [GCVI = (NIR/Green) – 1] (Gitelson et al., 2003), calculated from multispectral, 30m Landsat imagery for the years 2000 to 2017 (see Supplementary Table S1 for site data by year). Given the approximately 16 day overpass repeat cycle of Landsat, and the obfuscating impacts of cloud cover and other non-ideal atmospheric conditions on obtaining clear-sky retrievals, the specific dates of the derived Landsat-based GCVI data may vary from mid-July to mid-August. Since the green and near-infrared spectral bands of Landsat image data used to calculate the GCVI were only available at 30m resolution, each 10m resolution pixel that forms the explanatory variables is assigned to a GCVI value using a nearest neighbor filter. We note that GCVI has been shown to outperform more commonly used indices like Normalized Difference Vegetation Index (NDVI, Rouse et al., 1973) for predicting maize yield and was thus selected here (see Burke and Lobell, 2017).

2.2.5. Crop yield

Dry crop yield information was provided by analyzing the combine harvester data. Typically, around 60,000 records were available for each 64 ha study site per year. With respect to yield data QA/QC, the yield monitoring equipment was calibrated at the beginning of the year for each grain type by weighing the grain collected from a known area using a certified scale and comparing it to the pressure

plate estimated grain total (personal communication with university farm managers and private producers, December 2019). Moisture content of the grain was tested at the elevator directly following harvest of the calibration area as well as on-site using a portable moisture meter at selected sites. These direct moisture content readings were compared with the combine moisture content readings. The calibrations were conducted across large representative areas to ensure that sufficient yield was collected for each calibration. Furthermore, calibrations were verified at the end of each field using the total scale ticket weights collected at the elevator. In order to aggregate these data points to a gridded 10m resolution, we removed those records (less than 0.1 %) outside of an expected range for these sites (0–24 Mg/ha for maize and 0–12 Mg/ ha for soybean) for the reported moisture-standardized, dry grain yield data. Next, given the inherently noisy data at the approximate 1m resolution, a smoothing filter technique was applied (i.e. an inverse distance weighted procedure with a search radius of 50 m) to produce a 10m product. Supplementary Table S1 contains the yearly yield data available for each field. Additional information about identification of yearly crop yield outliers will be discussed in Section 3.4.

2.3. Methodological framework

With the advance and widespread use of yield monitoring in commercial agricultural and fine-resolution remote sensing data (Burke and Lobell, 2017), it is now possible to obtain fine-resolution crop yield maps across the globe and for numerous crop types. However, connecting the mechanistic causes for yield variation with a capacity to make optimal input decisions remains challenging. One problem is the complex nature of spatiotemporal yield variation. In order to simplify this problem, we propose a set of statistical techniques to separate the spatial and temporal components of yield variation and investigate the environmental controls on each component. We note that the separation of variables is a common and powerful technique in solving partial differential equations (Haberman, 1998) and we take analogous steps here using the statistical technique of Empirical Orthogonal Function (EOF). In order to help guide the reader through

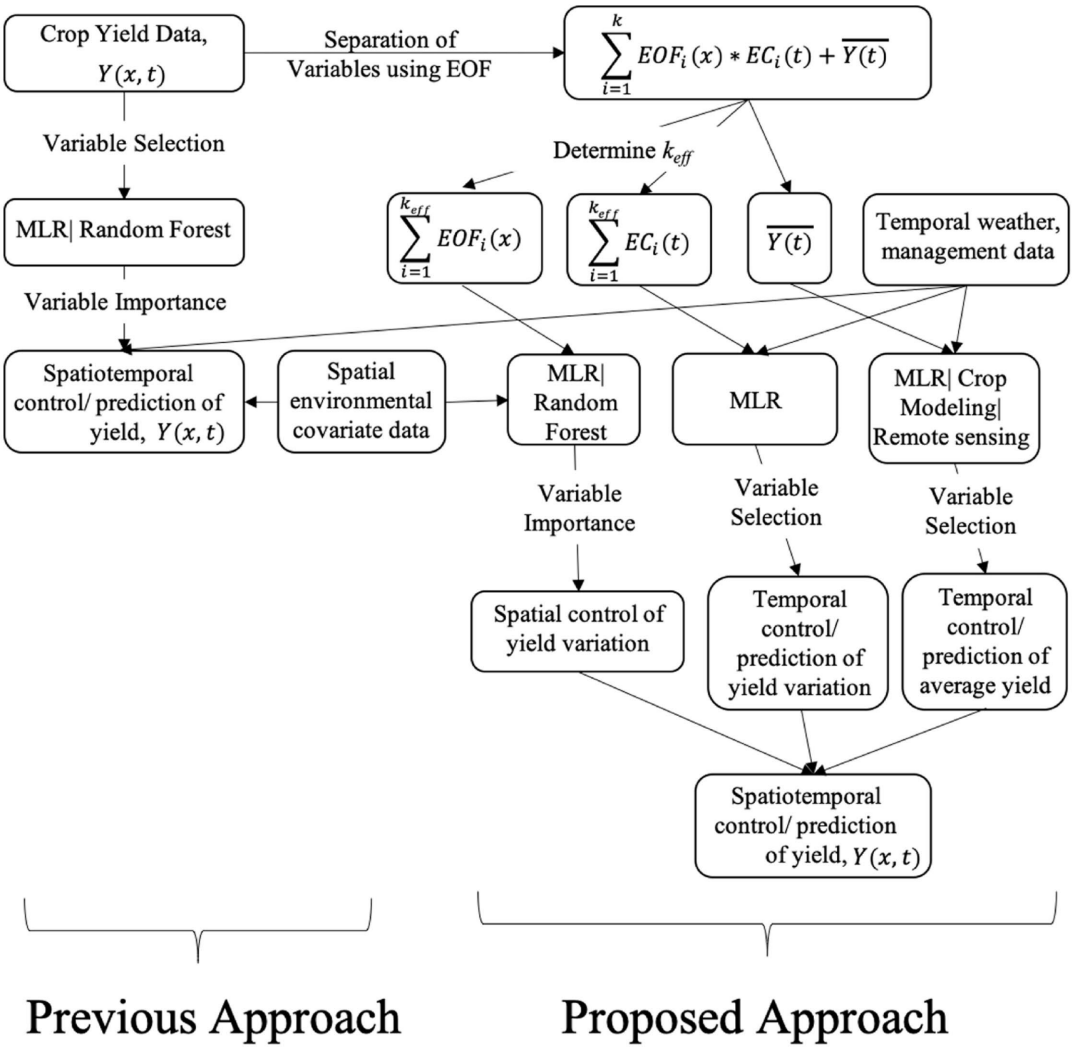

Fig. 3. Methodological framework that summarizes the EOF analysis that separates yield into spatial and temporal components. Subsequent steps illustrate different statistical analyses used to quantify the relative importance of spatial and temporal covariates as well as yield prediction. We note the left hand side has been used in previous work and the right hand side is the proposed approach here. Full documentation of R code steps and results are provided in Supplemental Table 4.

the various analysis used here, **Fig. 3** summarizes the methodological framework including the statistical techniques and data sources. The following sections will discuss each step in more detail.

2.3.1. Overview of empirical orthogonal function (EOF) analysis

With respect to the spatiotemporal variation of crop yield we are able to use the EOF technique to separate the spatial and temporal components as well as to reduce the dimensionality of the data (see Perry and Niemann, 2007; Korres et al., 2010). The crop yield data $\boldsymbol{Y}(x, t)$, where $\boldsymbol{x}$ is spatial location and $\boldsymbol{t}$ is time, can be decomposed in the following way:

$$\boldsymbol{Y}(x, t) = \sum_{\boldsymbol{i}=1}^{\boldsymbol{k}} \boldsymbol{EOF_i}(\boldsymbol{x})^{*}\boldsymbol{EC_i}(\boldsymbol{t}) + \bar{\boldsymbol{Y}}(t), \tag{1}$$

where $\boldsymbol{k}$ is the dimensionality of the input data (here the number of annual yield maps), ***EOF(x)*** is the set of time-invariant orthogonal spatial patterns, ***EC(t)*** is a set of time series expansion coefficients, and $\boldsymbol{Y}(t)$ is the field average yield. We note that EOF is nearly identical to Principal Component Analysis (PCA), save for the splitting of axes of variation into spatial and temporal coefficients instead of arbitrary linear combinations. Based on the calculated EOF, the original coordinate system is rotated into a new system aligned along perpendicular axes (similar to PCA). By retaining only significant EOF/EC pairs (based on a threshold of explained variance), EOF analysis can effectively reduce the dimensionality of the dataset (denoted by $\boldsymbol{k}_{\text{eff}}$) while preserving most of the variability present in the data.

Following Franz et al. (2017), for any time repeating dataset like crop yield, we have $\boldsymbol{n}$ locations (here a 10m resolution of approx. 7000 grid cells) and $\boldsymbol{k}$ observations (number of annual crop yield maps available), where the spatial anomalies of the crop yield observations can be computed as:

$$a_i(t) = s_i(t) - \frac{1}{n}\sum_{j=1}^{n} s_j(t) \tag{2}$$

where $\boldsymbol{a_i}(\boldsymbol{t})$ and $\boldsymbol{s_i}(\boldsymbol{t})$ are the crop yield observation spatial anomaly and yield observation at location $\boldsymbol{i}$ and time $\boldsymbol{t}$, respectively. A matrix of crop yield observation spatial anomalies, A (capital letters in bold denote matrices), can be constructed as:

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nk} \end{bmatrix} \tag{3}$$

Then, an empirical covariance matrix V can be calculated as:

$$\boldsymbol{V} = \frac{1}{k-1}\boldsymbol{A}^T\boldsymbol{A}, \tag{4}$$

where the superscript $\boldsymbol{T}$ indicates the matrix transpose. To perform EOF analysis, we find eigenvectors and eigenvalues for V, which satisfy the following equation:

$$\boldsymbol{V}\times\boldsymbol{E}=\boldsymbol{E}\times\boldsymbol{L} \tag{5}$$

where E contains eigenvectors (i.e. ECs) in columns:

$$\boldsymbol{E} = \begin{bmatrix} e_{11} & \cdots & e_{1k} \\ \vdots & \ddots & \vdots \\ e_{k1} & \cdots & e_{kk} \end{bmatrix} \tag{6}$$

and L contains eigenvalues along the diagonal:

$$\boldsymbol{L} = \begin{bmatrix} l_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & l_{kk} \end{bmatrix} \tag{7}$$

The above procedure rotates the original coordinate axes, with each axis indicating a sampling time, into a new set of orthogonal coordinate axes with each eigenvector representing a new axis. The eigenvalues explain the variance in the data along the direction of each corresponding new axis, and the portion of the explained variance ($\boldsymbol{EV_k}$) by the $\boldsymbol{i}_{\text{th}}$ new axis in the total variance can be computed as:

$$EV_i = \frac{l_{ii}}{\sum_{p=1}^{k} l_{pp}} \tag{8}$$

The eigenvectors are then arranged according to eigenvalues: the first axis explains the largest variance in the data, while each following axis explains the largest remaining variance and is orthogonal to other axes. The EOFs are then found by projecting A onto E

$$F = A \times E \tag{9}$$

where F contains each EOFs in columns. Based on the explained variance, only significant EOF/EC pairs are retained for the remaining

analysis (here defined as a threshold of greater than 10 % explained variance and denoted by $k_{eff;}$ see Peres-Neto et al. (2005) for a more complete discussion).

We note that the EOF analysis can be used on any time repeat spatial data, which will be explored in this work in order to analyze soil characteristics, vegetation condition, in addition to crop yield. Specifically we will perform EOF analysis on the EC_a, SWC, GCVI, and crop yield datasets for each available year for the eight study sites. Supplementary Table S1 contains the first axis EOF spatial coefficients (EOF_1) for each dataset and 10m grid, as well as the explained variance in the first axis (**Table 2**). Finkenbiner et al. (2019) and Gibson and Franz (2018) found that using EOF analysis on the EC_a dataset effectively isolated the soil hydraulic component of the signal, thus eliminating other confounding factors influencing EC_a values (such as temperature and SWC).

2.3.2. Relative importance of covariates in spatial and temporal patterns of crop yield

In order to understand the relative importance of the included covariates (topography, soil texture variation, SWC, vegetation condition) we utilized both linear and nonlinear statistical models. For all analyses we used the software R (R-3.6.1; R Core Team; www.r-project. org) and have included the results of each study site in the supplemental material. The supplemental tables include the input dataset, function used, and results of the analysis for full reproducibility and transparency of the work. We note that all input covariates were scaled by subtracting the mean and dividing by the standard deviation before running analyses. For the linear statistical model we used best subset Multivariate Linear Regression (MLR) to identify statistically-significant covariates using the lowest Bayesian Information Criterion to select the model. We also used the Variance Inflation Factor to check for and remove any mutlicollinearity (Chatterjee and Price, 1977). In order to calculate the model parameter relative importance for the selected model, we used the Lindeman, Merenda and Gold method (Gromping, 2006 and Soofi et al., 2000). For statistical evaluation, we split the data in half for training and testing. Using the test data, we computed various prediction

Table 2 Summary data of each study site including: elevation, number of soil water content maps, electromagnetic induction shallow and deep maps, GCVI data, and crop yield maps. Where available the percent of variance explained by the first axis using EOF is listed.

	Elevation	Soil Water		Soil Texture					Vegetation Condition				Crop Yield				
Study Site	Topography	Neutron Intensity Maps (# of)	Explained Variance in EOF1	ECa Shallow (0–1m) (# of)	Explained Variance in EOF1	ECa Deep (0–3m) (# of)	Explained Variance Maps EOF1	ECa Shallow to Deep ratio (# of)	Explained Variance in EOF1	Landsat GCVI Soybean (# of)	Explained Variance in EOF1	Landsat GCVI Corn (# of)	Explained Variance in EOF1	Soybean Yield Maps (# of)	Explained Variance in EOF1	Corn Yield Maps (# of)	Explained Variance in EOF1
1	NE LiDAR	4	74.8	3	94.7	3	95.7	3	83.5	NA	NA	3	95.7	NA	NA	2	83
2	NE LiDAR	4	59.5	1	NA	1	NA	1	NA	NA	NA	7	96.1	NA	NA	7	76.4
3	NE LiDAR	7	67.1	3	78.2	5	68.7	3	81.3	NA	NA	2	98	NA	NA	2	84.7
4	NE LiDAR	7	76.7	1	NA	1	NA	1	NA	NA	NA	3	98.5	NA	NA	3	79.7
5	NE LiDAR	10	69.5	1	NA	1	NA	1	NA	NA	NA	3	95.8	NA	NA	3	67.5
6	NE LiDAR	1	NA	1	NA	1	NA	1	NA	NA	NA	17	80.7	NA	NA	14	76.3
7	NE LiDAR	3	43.8	3	57.7	3	61.4	3	58.7	7	84.8	9	91.8	6	81.6	9	75.5
8	NE LiDAR	4	40	4	67.7	4	64.2	4	69.1	8	83.8	8	80.8	8	67.4	8	60.1

metrics including: coefficient of determination (R^2), bias, RMSE, and unbiased RMSE (ubRMSE):

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (P_i - \bar{O_i})^2}{\sum_{i=1}^{n} (O_i - \bar{O_i})^2} \tag{10}$$

$$Bias = \frac{1}{n}\sum_{i=1}^{n} (P_i - O_i) \tag{11}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (P_i - O_i)^2} \tag{12}$$

$$ubRMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (P_i + Bias - O_i)^2} \tag{13}$$

where $\boldsymbol{n}$ is the total number of crop yield data points in the test dataset, $\boldsymbol{i}$, $\boldsymbol{O_i}$, and $\boldsymbol{P_i}$, are the observed and predicted crop yield of the $\boldsymbol{i}$th data point, and the overbar denotes the average respectively.

For the nonlinear statistical model we used a Random Forest (RF) approach (Breiman, 2001). The RF is a non-parametric machine learning approach that uses many decision trees as base level classifiers, with each decision tree trained on a different sub-set of the input data, referred to as bootstrap sample. For each tree, the observations that are randomly included in the bootstrap sample are referred to as "in-bag" samples, the observations excluded are referred to as "out-of-bag" samples. The RF approach averages these multiple decision trees to provide a more robust prediction than would be available from any single decision tree. Machine learning approaches have proliferated in the literature in recent years (Belgiu and Dragut, 2016), with the application of many techniques for enhanced data analysis. One of the reasons for exploring the RF approach here is the history of technique in similar applications (Shah et al., 2019), where it has been shown to provide good accuracy, while avoiding the issues associated with overfitting. Here we used the software package R, with 1000 regression

trees, and a minimum leaf size of 5 (i.e. number of observations in the end node of a decision tree) on the training data (see Supplemental material for full functions and results). The out-of-bag samples are used to calculate the percentage increase in MSE when we remove a particular covariate from the model, this gives us information regarding the relative importance of each covariate. The test data set is used to calculate the prediction RMSE and prediction R^2.

3. Results

3.1. EOF analyses

EOF analyses (see Section 2.3.1) were performed on all site data for SWC, EC_a shallow, EC_a deep, ECa shallow to deep ratio, GCVI for maize and soybean, and crop yield for maize and soybean, where at least three datasets for each variable were available. Table 2 summarizes the explained variance of the first axis for each dataset (Supplementary Table S1 contains the EOF_1 spatial coefficients). Importantly we found that the first axis of variation for both maize and soybean yield dominates the explained variance (60–85 %) reducing the effective dimensionality of the problem to 1 ($\boldsymbol{k}_{\text{eff}} = 1$). This indicated that the spatial pattern of crop yield manifests repeatedly when the mean year-to-year changes are removed. The consequence of excluding second and higher order axes of variation will be discussed in Section 4 but note that Eq. (1) can handle higher order terms if necessary. In addition, we found that for the same field with different crops (only S7 and 8 had enough yield datasets) EOF_1 coefficients of crop yield did slightly change (Pearson correlation coefficients of ~0.8). This indicates that the crop type and plant physiology interact differently with the same input covariates in terms of crop yield patterns.

With respect to the other datasets, the vast majority of the explained variance (> 60 %) were also largely in the first axis of variation, allowing us to also reduce the dimensionality of the covariate data. This is beneficial as SWC, EC_a, and GCVI can vary greatly over time during and in-between growing seasons. As a result, only EOF_1 spatial covariates were used in subsequent statistical analyses, thus removing any confounding effects due to temporal changes (see Finkenbiner et al., 2019 for an example using SWC and Franz et al.,

2017 for using higher order EOFs). **Fig. 4** presents the 10m combine-harvest derived crop yield maps and **Fig. 5** the individual dataset and EOF_1 coefficients for each variable for S2 (irrigated maize in hot and dry climate, see Table S1 for all site data). Figs. 4 and 5 illustrate the predictor variables and response variables that will be used in the MLR and RF approaches in the next section. The purpose of the figures is to highlight the degree of spatial variability that exists within each dataset and the challenge of connecting any one covariate with the response variable of crop yield. Lastly, we selected at random 500 subsamples of each dataset ranging from 2 to k-1. We then computed the EOF_1 coefficients for the subsampled datasets averaged them and then compared them to the full dataset. We found that 3–5 samples for each dataset were sufficient to describe the scaled EOF_1 coefficients (i.e. scaling by min and max) within 5 % of the full dataset's EOF_1. This finding is consistent with Finkenbiner et al., 2019 who investigated only SWC data, where SWC is bounded by soil physical properties. We note that since crop yield range may change from year to year that the scaled EOF_1 coefficients show the consistent spatial pattern against the full dataset.

3.2. Relative importance of spatial covariates on crop yield pattern

In order to investigate the contribution of each input covariate to predicting crop yield EOF_1, we used the MLR and RF approaches described in Section 2.3.2. For the analysis, the response variable was EOF_1 coefficients for maize (or soybeans) crop yield, and the input covariates were grid cell location (polar coordinates of radius and angle relative to field center), relative elevation (set minimum grid cell field elevation to 0) and EOF_1 for SWC, EC_aS, EC_aD, EC_aSDR, and GCVI. The full results for each site are documented with R code and available in the Supplemental material. **Fig. 6**a summarizes the prediction RMSE for both MLR and RF for each study site and maize crop yield pattern (Supplementary Table S2 contains the MLR and RF results for both crops). The analysis illustrated a large prediction RMSE reduction (between 50–75 %) between the MLR and RF approaches for each field, thus indicating a nonlinear model is justified between input covariates and the response variables considered here. With respect to relative importance of variables Fig. 6b illustrates the results for the RF

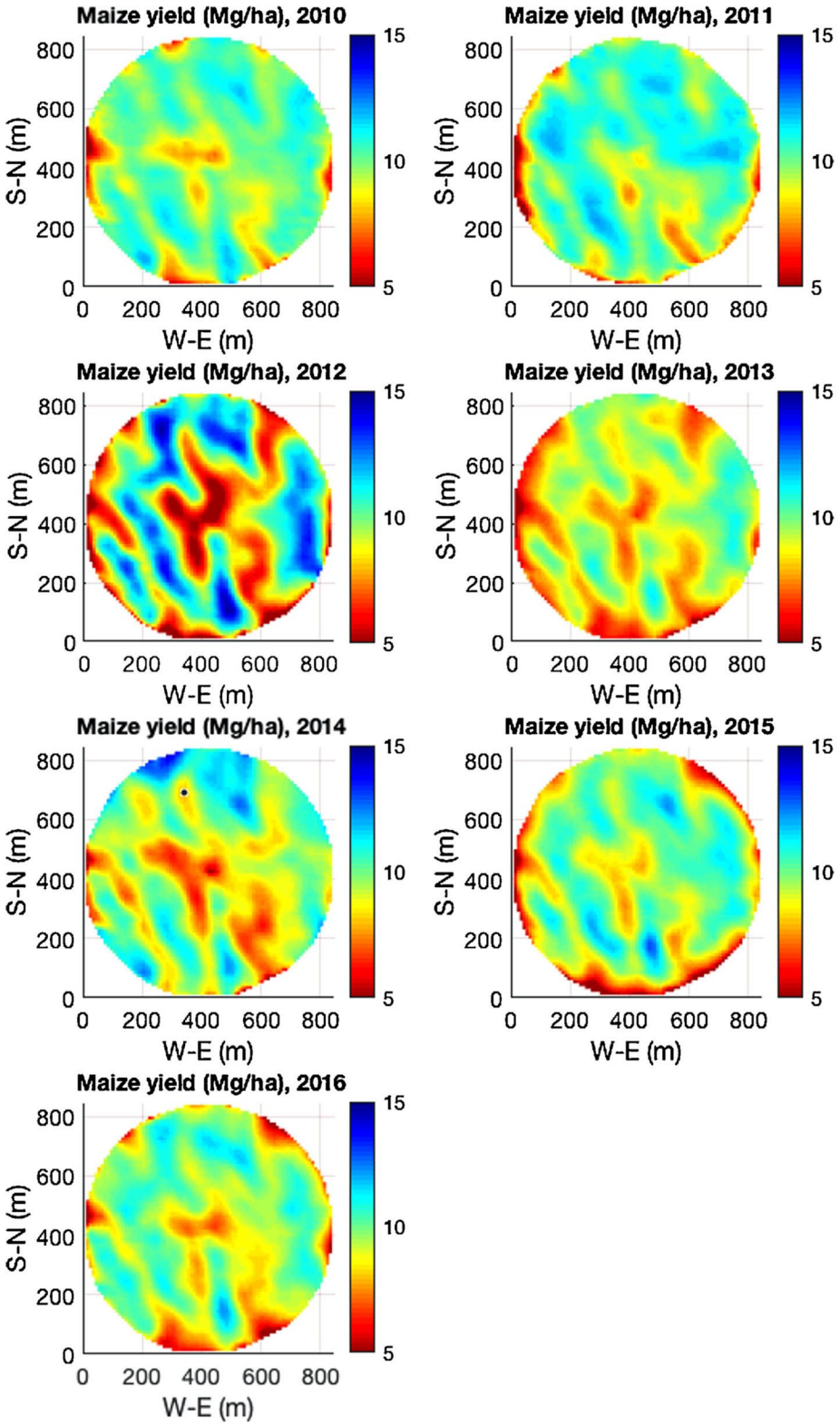

Fig. 4. 10m resolution QA/QC maize crop yield maps (Mg/ha) of S2 from 2010 to 2016.

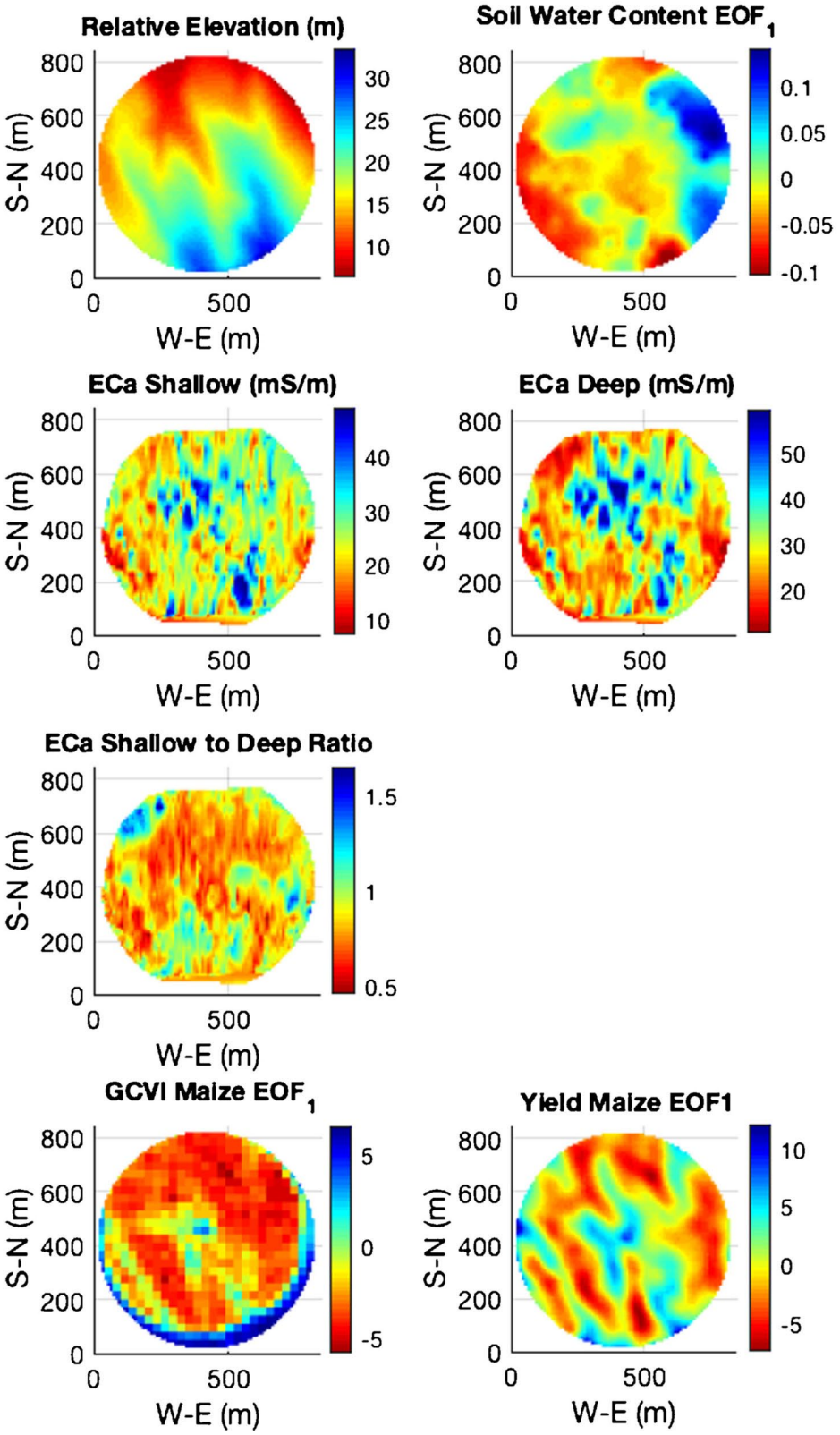

Fig. 5. 10m resolution maps of input covariates from S2 and EOF1 of maize crop yield.

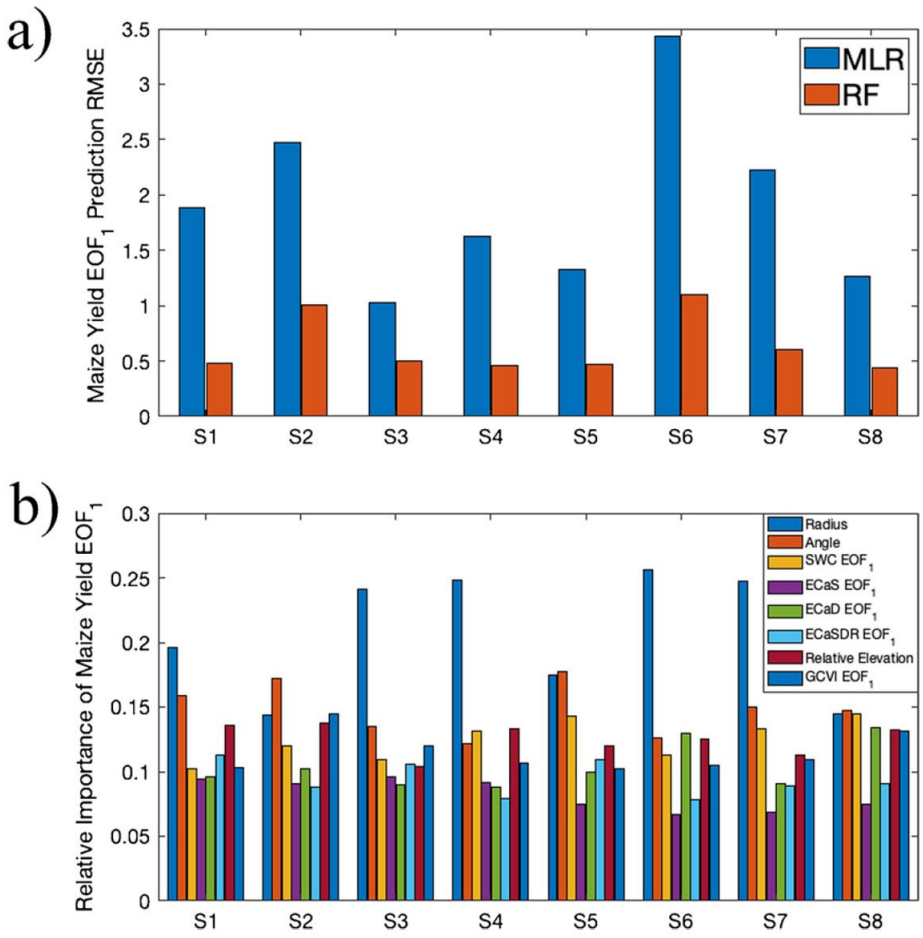

Fig. 6. a) Comparison of MLR versus RF analysis for predicting maize crop EOF1 using same input covariates. RF shows a marked decrease (50–75 %) in prediction RMSE. b) Relative importance of input covariates for RF analysis. Spatial location (radius, angle) tends to have largest relative importance followed by GCVI, SWC, and elevation depending on local site conditions (see Fig. 3 for distributions of elevation and soil texture).

analysis (MLR results are in Table S2). The results indicate that grid cell location (radius and angle) are overwhelming the primary contributors to explaining crop yield EOF_1. This is not overly surprising as grid cell location is correlated with many physical factors such as field boundaries, local depressions/high points, planting/irrigation/fertilization application pattern, wheel traffic, hydrologic barriers around the field boundary (i.e. roads, berms, fence lines) or internal hydrologic drainage patterns or structures (i.e. tile drains). The secondary factors that are important after location (combined effects of radius and angle) varied with relative field heterogeneity (see Fig. 3 for

topography and soil texture variation). For most field sites, the GCVI, SWC, and elevation had greater importance than the ECa data. In general, the interquartile range in Fig. 3 compares well with the relative importance magnitude in Fig. 6b, thus indicating *a priori* what may be the best secondary controlling factor(s) affecting crop yield pattern (topography vs. soils). However, the multiple contributing factors and collinearity between covariates makes it challenging to gain insight on how to best use the covariate data for decision making on crop yields. This will be discussed more in Section 4.

3.3. Relative importance of spatial covariates on temporal crop yield

Table 3 contains a summary of all statistical model results by year. In addition, the average field yield is reported, as well as a leave-one-out, cross validation RMSE for comparison of relative error. The leave-one-out, cross validation RMSE will serve as our benchmark (or null model) for comparison against the more complex statistical models presented. Using the same MLR and RF approach, we investigated the relative importance of each covariate on crop yield by year for each site. **Fig. 7**a illustrates for S2 the prediction RMSE by year using the MLR and RF approaches (Supplementary Table S3 contains results for all sites and all years and Table 3 contains a summary of results). The two sets of input covariates were selected to investigate the importance of in-season remote sensing GCVI data as well as using historical yield pattern (EOF_1 of crop yield) as an input covariate. For the first MLR and RF input covariate models (MLR1 and RF1), the response variable was crop yield by year, with the covariates as: grid cell location (radius and angle from field centers), relative elevation, the EOF_1 of SWC, ECaS, ECaD, ECaSDR, and GCVI for the same season as crop yield. For the second MLR and RF input covariate model (MLR2 and RF2), crop yield EOF_1 was added to the covariate set 1. With respect to relative importance

Fig. 7b and c illustrate again that location in the field had the largest importance in covariate set 1, followed by GCVI, SWC, and elevation. For set 2, crop yield EOF_1 had the largest importance followed by location, GCVI, SWC, and elevation. ECa continues to have the lowest relative importance. These results were consistent across all sites (see Supplementary Table S3 for data). Fig. 7a illustrates a large decrease

Table 3 Summary statistics of avg. field yield, leave one out cross validation, prediction RMSE for MLR and RF input covariate datasets 1 and 2, and bias, RMSE, and ubRMSE for EOF reconstruction, by study site and years included in the analysis. We note that three outlier years were identified and excluded from the statistical analyses (S6-2008, S6-2010, S7-2010).

Site Number	Crop	Year	Avg. Yield (Mg/ha)	RMSE Leave one out cross validation (Mg/ ha)	Pred. RMSE MLR set 1 (Mg/ha)	Pred. RMSE RF set 1 (Mg/ha)	Pred. RMSE MLR set 2 (Mg/ha)	Pred. RMSE RF set 2 (Mg/ ha)	EOF Bias (Mg/ha)	EOF RMSE (Mg/ha)	EOF ubRMSE (Mg/ha)
2	Maize	2010	9.79	0.838	0.733	0.315	0.400	0.184	0.391	0.698	0.578
2	Maize	2011	9.97	1.054	0.725	0.296	0.462	0.194	0.565	0.951	0.765
2	Maize	2012	9.34	1.528	1.787	0.703	0.404	0.247	−0.001	0.626	0.626
2	Maize	2013	8.77	1.025	0.752	0.298	0.353	0.161	−0.622	0.843	0.569
2	Maize	2014	9.42	1.137	0.954	0.357	0.594	0.210	0.000	0.964	0.964
2	Maize	2015	9.34	0.802	0.837	0.342	0.342	0.173	−0.007	0.716	0.716
2	Maize	2016	9.30	0.696	0.816	0.348	0.437	0.166	−0.069	0.592	0.588
4	Maize	2011	11.14	1.386	0.744	0.237	0.425	0.149	NaN	NaN	NaN
4	Maize	2012	12.58	1.401	1.185	0.384	0.365	0.172	NaN	NaN	NaN
4	Maize	2013	11.75	1.042	1.094	0.293	0.469	0.184	NaN	NaN	NaN
5	Maize	2011	10.59	3.437	0.854	0.225	0.677	0.198	NaN	NaN	NaN
5	Maize	2012	11.64	2.255	1.182	0.473	0.363	0.194	NaN	NaN	NaN
5	Maize	2015	16.09	5.093	0.607	0.211	0.492	0.165	NaN	NaN	NaN
6	Maize	2001	13.57	2.339	0.204	0.147	0.662	0.328	1.751	1.979	0.922
6	Maize	2002	12.48	1.506	0.824	0.305	0.673	0.275	0.441	1.182	1.097
6	Maize	2003	11.66	0.910	0.938	0.429	0.669	0.254	−0.170	0.846	0.828
6	Maize	2004	11.92	0.782	0.670	0.208	0.362	0.169	0.164	0.574	0.550
6	Maize	2005	11.16	0.877	0.654	0.278	0.428	0.188	−0.948	1.110	0.576
6	Maize	2006	10.63	1.451	0.403	0.227	0.378	0.168	−1.081	1.394	0.881
6	Maize	2007	11.59	0.701	0.859	0.306	0.497	0.213	0.285	0.805	0.753
6	Maize	2011	11.07	1.004	1.025	0.393	0.650	0.272	0.068	0.940	0.938
6	Maize	2012	12.31	1.242	0.726	0.273	0.516	0.191	0.106	0.843	0.836
6	Maize	2013	11.50	1.346	1.531	0.612	0.833	0.380	−0.128	1.303	1.297
6	Maize	2014	10.17	1.763	1.205	0.486	0.599	0.299	−0.694	0.983	0.696
6	Maize	2015	10.75	1.327	1.461	0.498	0.525	0.271	0.254	0.706	0.659
6	Maize	2016	11.24	1.318	1.561	0.562	0.728	0.273	0.127	0.991	0.982
6	Maize	2017	11.68	1.285	1.428	0.528	0.708	0.354	0.385	1.117	1.049
7	Maize	2001	13.28	1.110	0.666	0.226	0.442	0.163	0.494	0.910	0.764
7	Maize	2003	13.31	1.241	0.377	0.133	0.299	0.119	0.522	0.779	0.578
7	Maize	2005	12.38	0.752	0.528	0.187	0.321	0.155	−0.308	0.642	0.563
7	Maize	2007	12.53	0.860	0.644	0.206	0.402	0.169	−0.098	0.680	0.673
7	Maize	2011	11.65	1.429	0.805	0.255	0.503	0.200	−0.744	1.145	0.871
7	Maize	2012	12.55	1.247	0.946	0.246	0.635	0.189	0.027	1.154	1.153
7	Maize	2013	12.65	1.205	1.102	0.345	0.682	0.307	−0.102	1.427	1.424
7	Maize	2015	12.18	1.339	1.145	0.417	0.471	0.276	0.387	0.860	0.768
7	Maize	2017	13.04	1.644	1.248	0.414	0.504	0.252	0.409	1.775	1.728
7	Soybeans	2002	3.84	0.517	0.141	0.057	0.112	0.044	−0.074	0.246	0.235
7	Soybeans	2004	3.54	0.806	0.114	0.047	0.093	0.041	−0.375	0.421	0.192
7	Soybeans	2006	4.11	0.407	0.173	0.063	0.102	0.051	0.208	0.268	0.169
7	Soybeans	2008	4.17	0.412	0.288	0.100	0.213	0.079	0.252	0.492	0.423
7	Soybeans	2014	3.74	0.595	0.388	0.116	0.098	0.059	−0.147	0.230	0.176
8	Soybeans	2016	4.09	0.573	0.375	0.121	0.114	0.067	0.202	0.376	0.318
8	Maize	2001	8.57	1.589	0.732	0.257	0.636	0.209	−0.546	0.873	0.682
8	Maize	2003	7.44	2.769	0.625	0.236	0.527	0.179	−1.666	1.759	0.565
8	Maize	2005	8.96	1.081	0.426	0.139	0.382	0.114	0.405	0.624	0.475
8	Maize	2007	9.93	0.442	0.528	0.170	0.354	0.139	−0.160	0.411	0.379
8	Maize	2011	9.48	0.634	0.512	0.190	0.378	0.146	−1.128	1.227	0.484
8	Maize	2013	10.51	0.932	0.499	0.170	0.310	0.119	1.272	1.323	0.365
8	Maize	2015	11.49	2.035	0.693	0.228	0.407	0.171	0.082	0.519	0.512
8	Maize	2017	11.92	2.600	0.986	0.327	0.581	0.185	1.858	1.997	0.731
8	Soybeans	2002	3.14	0.619	0.732	0.257	0.636	0.209	0.077	0.315	0.306
8	Soybeans	2004	3.26	0.448	0.625	0.236	0.527	0.179	−0.481	0.507	0.159
8	Soybeans	2006	4.20	0.731	0.426	0.139	0.382	0.114	0.395	0.443	0.200
8	Soybeans	2008	4.03	0.618	0.528	0.170	0.354	0.139	−0.147	0.224	0.169
8	Soybeans	2010	4.07	0.675	0.512	0.190	0.378	0.146	0.028	0.122	0.118
8	Soybeans	2012	2.13	1.719	0.499	0.170	0.310	0.119	−0.004	0.295	0.295
8	Soybeans	2014	3.60	0.252	0.693	0.228	0.407	0.171	−0.372	0.661	0.546
8	Soybeans	2016	4.38	0.924	0.986	0.327	0.581	0.185	0.157	0.295	0.250

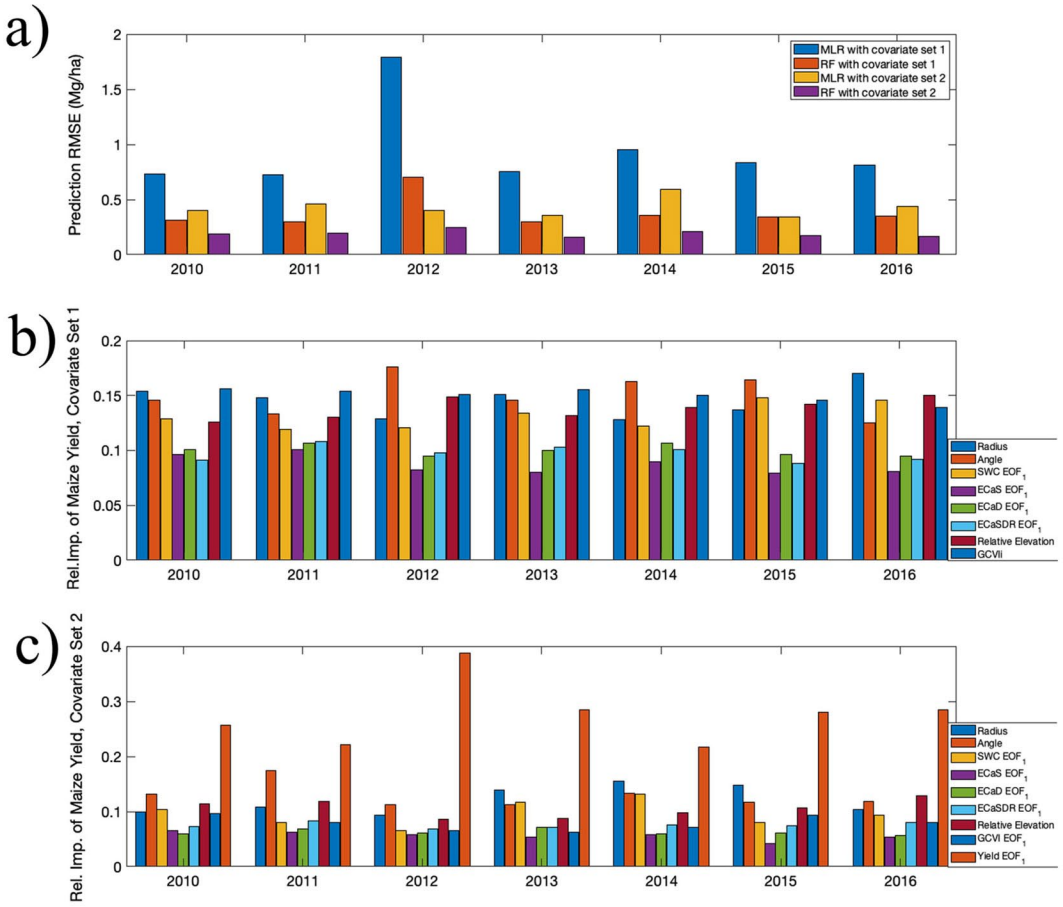

Fig. 7. a) Prediction RMSE of yearly crop yield using MLR and RF for two different sets of input covariates for S2. Covariate set 2 includes crop yield EOF1 as a input covariate. b and c) Relative importance results for RF analysis with covariate sets 1 and 2 (Supplementary Table S3 contains all sites and all years).

in prediction RMSE when using an RF vs MLR approach. However, MLR with data set 2 did perform considerably better compared against the leave-one-out cross validation benchmark (Table 3). Thus, a MLR approach using historic crop yield information appears to be satisfactory model. We do note that in covariate set 2, using the EOF_1 of crop yield as a covariate is somewhat circular. However, from cross validation we found that EOF_1 of crop yield requires only 3–5 years to estimate coefficients within 5 % of their values using the full dataset. Thus, operationally we argue that EOF_1 of crop yield could be established with historical data or remote sensing and used as an input covariate for prediction in future years. This prediction strategy will be discussed further in Section 4.

3.4. Spatiotemporal reconstruction and prediction of crop yield using EOF analysis

The benefit of the EOF-framework is that we can use this together with Eq. (1) to reconstruct and predict the yearly crop yield patterns. Table 2 illustrates that $\boldsymbol{EOF_1}$ of crop yield contains 60–85 % of the explained variance, thus greatly reducing the problem to a single axis, $\boldsymbol{k}_{\text{eff}} = 1$ and $Y(x, t) \sim \boldsymbol{EOF}_1(\boldsymbol{x}) * \boldsymbol{EC}_1(\boldsymbol{t}) + \boldsymbol{Y}(t)$. In order to perform the reconstruction and prediction, we need to first estimate functions that describe how the first expansion coefficient, $EC_1(t)$, and average field crop yield, $\bar{\boldsymbol{Y}}(t)$, change with time. For this research, we explore whether growing season aridity index is a good candidate to parameterize these relationships through time. Additional crop management information (e.g., planting density and applied nitrogen) and growing season weather conditions (e.g., min. and max temperature and incoming solar radiation) would certainly add further predictive value, particularly for better describing $\bar{\boldsymbol{Y}}(t)$.

Fig. 8 illustrates the relationship between growing season aridity index with EC_1 and $\bar{\boldsymbol{Y}}(t)$ for S2 (**Table 4** contains all available sites and fitting selection results for a zero to second order polynomial). It is evident from Fig. 8 and Table 4 that a second order polynomial fits the relationship between EC_1 and aridity index ($R^2_{adj} = 0.99$). However, no satisfactory relationship existed between $\bar{\boldsymbol{Y}}(t)$ and aridity index, so a zero order polynomial was selected (i.e. a constant value of $\overline{Y} = 9.974$ Mg/ha). Here, additional management information may be needed to capture the small year-to-year variations in $\bar{\boldsymbol{Y}}(t)$ and thus eliminate any systematic bias in the EOF reconstructed/predicted crop yield. **Table 4** summarizes the selected models for EC_1 and $\bar{\boldsymbol{Y}}(t)$ for each site. Given the small sample sizes (6–14), only second order polynomial models were considered and R^2_{adj} was used for model selection. In addition, when plotting the relationships, three outlier years were evident visually (S6-2008, S6-2010, S7-2010) and were subsequently removed from the fitting and model selection process. It was not immediately clear if the outliers were due to machinery breakdowns, pest pressure, hail damage or other factors. As in all such studies, local qualitative information from the producer is extremely valuable in identifying and eliminating outliers. The selected models in Table 4 varied between zero and second order polynomials depending on

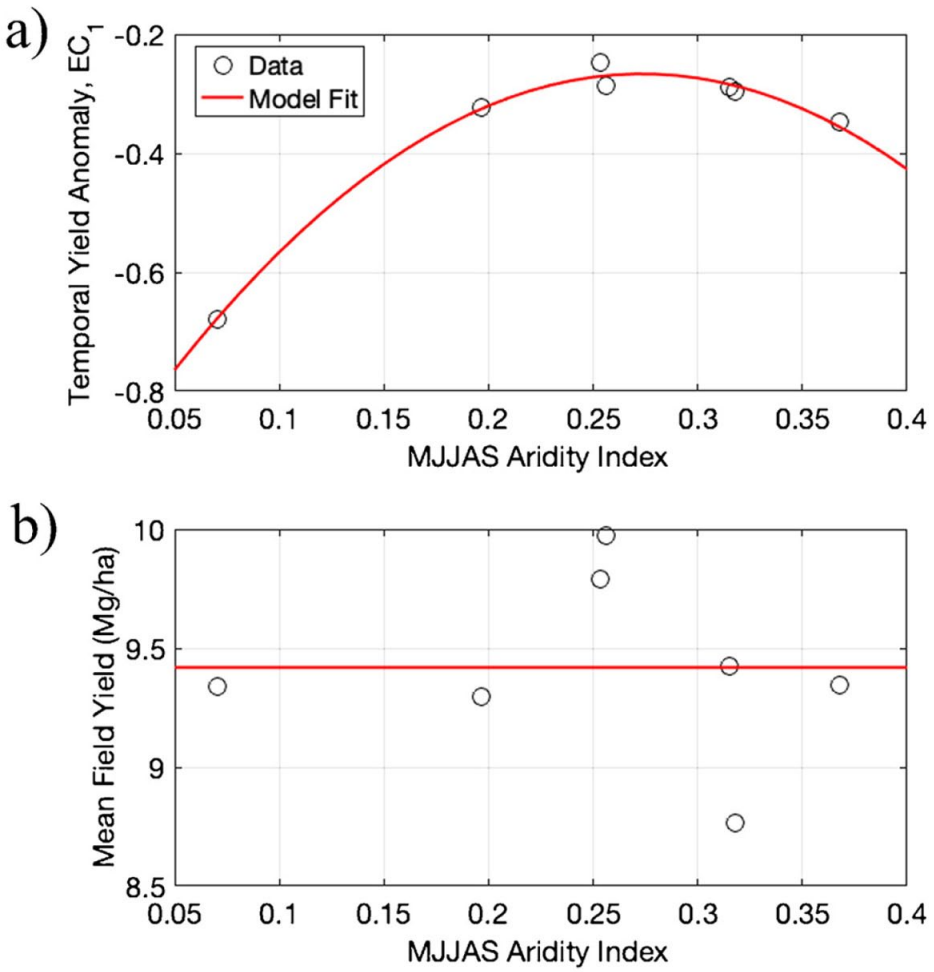

Fig. 8. EOF reconstruction for S2. The a) temporal anomaly (EC1) is fitted with a second order polynomial using growing season aridity index. b) The mean field yield is not a function of aridity index so a constant is assumed. See Table 4 for model selection and fit for all sites where enough crop yield data is available (> 5 years).

Table 4 Summary of fitting analysis between temporal anomaly (EC1) versus aridity index and average field yield (**Y**) vs. aridity index for EOF reconstruction of site data with a sufficient number of years. See Table 3 for specific years included for each site.

				Temporal Anomaly, EC1 (Aridity Index)					Field Average Yield, Y (Aridity Index)				
Site Name	Site Number	Crop	Number of Years	Functional form	a	b	c	adjR²	Functional form	a	b	c	adjR²
BWL2	2	Maize	7	y=a*x^2+b*x+c	−1.01	5.45	−9.98	0.99	y=a	9.97			NA
CSP1	6	Maize	14	y=a*x+b	0.31	0.08		0.34	y=a*x+b	−2.45	12.90		0.29
CSP2	7	Maize	9	y=a*x+b	0.31	0.14		0.10	y=a*x+b	13.28	−1.13	0.11	
CSP2	7	Soybeans	6	y=a*x+b	0.95	−0.19		0.52	y=a	3.92			NA
CSP3	8	Maize	7	y=a*x+b	−0.10	−0.28		0.09	y=a*x+b	4.63	7.18		0.29
CSP3	8	Soybeans	7	y=a*x^2+b*x+c	4.70	−4.10	0.99	0.67	y=a*x^2+b*x+c	−11.58	15.11	−0.69	0.83

each field. Further improvements to model selection will be discussed in Section 4. Returning to S2, we used the derived functions for EC_1 and $\boldsymbol{Y}$ (from Fig. 8 and Table 4) and used Eq. (1) with $\boldsymbol{k}_{\text{eff}}$ =1 to reconstruct the 10m resolution crop yield patterns for 2010–2016 (see **Fig. 9** and Table 3). The reconstructed RMSE ranged from 0.6 to 1 Mg/ha, which is comparable to the MLR models presented in Fig. 7a. However, and most notably, only three derived parameters and aridity index are needed to describe EC_1 and $\boldsymbol{Y}$ to estimate the reconstruction and subsequent future predictions of crop yield. Table 3 and **Fig. 10** present a statistical summary for all sites and years using the EOF reconstruction and fitted relationships from Table 4. We found that in most years and most sites, RMSE is between 0.5–1.7 Mg/ha for maize, and 0.2–0.6 Mg/ha for soybean, indicating a relative error of less than 10 % of the mean and a RMSE reduction of 10–40 % compared against the leave-one-out cross validation prediction. However, approximately 10 % of site years contained a significant bias and reduced ubRMSE vs RMSE. Here an improved statistical or crop model describing $\boldsymbol{Y}^{-}(t)$ is needed, which will be discussed further in Section 4.

4. Discussion

Perhaps the key result from the EOF analysis is that the crop yield pattern manifests itself year after year following the removal of the mean for the eight sites across NE. Moreover, the first axis of variation was able to capture 60–85 % of the explained variance, thus greatly reducing the complexity and dimensionality of the problem. That is when $\boldsymbol{k}_{\text{eff}}$ =1, Eq. (1) reduces to $\boldsymbol{Y}(x, t) \sim \boldsymbol{EOF}_1(\boldsymbol{x}) * \boldsymbol{EC}_1(\boldsymbol{t}) + \boldsymbol{Y}^{-}(t)$, effectively allowing us to separate the spatial and temporal components of yield and form a better understanding of the underlying physical mechanisms and associated data. We do note that certain fields may require consideration of additional axes of variation (particularly with explained variances>10 %). The EOF framework is able to handle additional axes in the reconstruction, but fitting $EC_i\,(t)$ may be more problematic (see Franz et al., 2017 for an example). Here we found the MLR and RF analysis, undertaken using a range of covariates describing topography, SWC, soil texture and vegetation condition over the growing season, were also challenging to interpret (Maestrini and

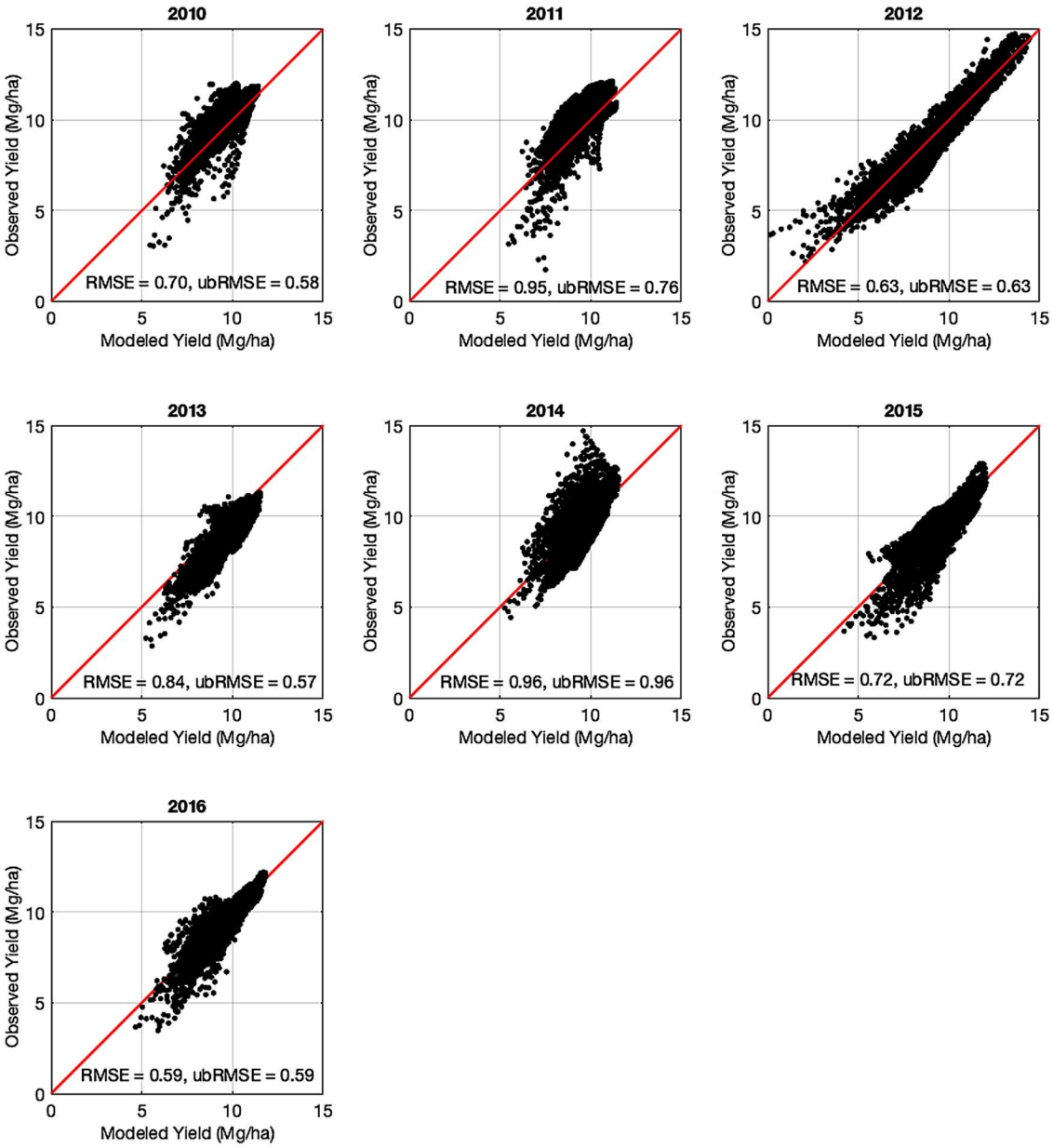

Fig. 9. One to one comparison of 10m resolution observed crop yield and modeled yield using EOF technique (Eq. 1 with keff=1) including the EOF1 spatial anomalies and functions found between aridity index and summarized in Fig. 8 and Table 4 for S2.

Basso, 2018). First, it is clearly demonstrated by the reduction in prediction RMSE between the MLR and RF approaches, that the relationship between crop yield and the selected covariates is nonlinear. Moreover, the relative importance analysis revealed that location within the field was the largest primary factor for all eight fields. As mentioned

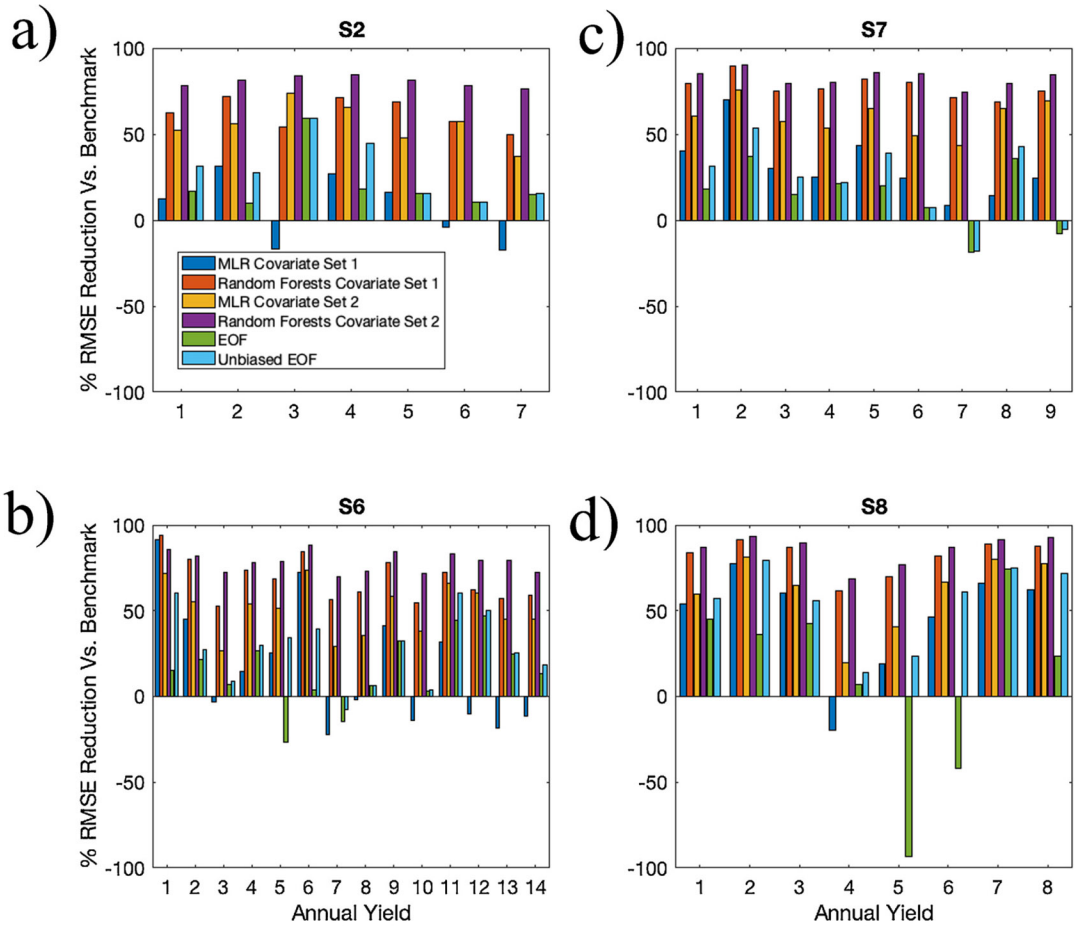

Fig. 10. Evaluation of various statistical model performance vs. a benchmark for maize at S2, 6, 7, and 8 where sufficient annual crop yield data exists. Here the benchmark is set as the leave-one-out cross validation for each location and year. Negative values indicate an inferior model compared to the benchmark. See Table 3 for complete results of all sites.

before, location within the field is correlated with several known and unknown factors including: field boundaries, planting/irrigation/ fertilization application pattern, wheel traffic, hydrologic barriers around the field boundary, and/or internal hydrologic drainage patterns or structures. This spatial autocorrelation also manifests in a geostatistical variogram analysis and needs to be considered in any spatial regression analysis (see Supplemental Table 4 for results). Secondary factors varied between site GCVI, SWC, and elevation, roughly following the interquartile range of those covariates illustrated in Fig. 2, whereas soil texture (EC_a) seemed to have a tertiary influence. We

note that GCVI was derived from Landsat data at 30m and reprocessed to 10 m. Given that GCVI was a strong secondary factor we expect it to have an even greater importance as the data resolution becomes finer (Burke and Lobell, 2017) from newer satellites, aircraft or unmanned aerial vehicles (UAVs) (Houborg and McCabe, 2018a).

A key question that was posed at the outset of this analysis was: what level of investment should the producer invest in order to obtain these datasets? Some datasets are freely available (i.e. topography, Landsat archive), whereas others cost a few USD per ha (i.e. imagery from aircraft or private satellites). In contrast, some spatial data sets require an investment of a few to several tens of USD per ha (i.e. hydrogeophysical). With respect to the benefit of predicting crop yield, we found that incorporating historic crop yield maps are superior to any other covariate, as illustrated by the difference in the two different MLR and RF input covariate sets in Fig. 7. In addition, fine-resolution elevation data (freely available) and GCVI seem to be a good secondary covariate to include in a statistical model of crop yield. A key benchmark for any statistical or crop model prediction will be to outperform the leave-one-out cross validation using crop yield data (Fig. 10). This benchmark and test can be used as an estimate of economic return for any new dataset (crop yield price multiplied by prediction RMSE reduction vs. the benchmark) and can be used as guidance for determining price point.

Of course, near real-time imagery from aircraft and satellites can be used to diagnose other useful problems, like clogged sprinklers and disease onset and outbreak. Two potentially important aspects that have not been explored here are the spatial and temporal resolution of available remotely sensed data, and how these might impact predictions. For instance, the Landsat data employed here provide only a single peak value of GCVI during the season, whereas there might be information content in more frequent observations or during critical growth periods related to crop yield (i.e., flowering stage for soybeans and silking stage for maize). The availability of visible and near-infrared data from Sentinel-2 satellites offer a 5-day repeat time, while recently developed CubeSats such as PlanetScope (Houborg and McCabe, 2018a) offer the capacity for daily retrieval, with variables such as NDVI and LAI being routinely retrieved (McCabe et al., 2017a,b). Perhaps equally important is the spatial resolution of the available products, with these new satellite systems

offering native spatial resolutions of high as 3 m, providing two orders of magnitude more spatial detail than a comparable 30m Landsat pixel (Houborg and McCabe, 2018a,b). It is also worth noting that only a single metric of crop condition (i.e. the Green Chlorophyll Vegetation Index) was explored here, whereas there are numerous vegetation indices available from satellite and airborne platforms that may provide much greater insight into crop yield (see for example Shah et al., 2019 and Burke and Lobell, 2017).

With respect to hydrogeophysical mapping, Finkenbiner et al. (2019) and Gibson and Franz (2018), clearly showed hidden soil texture boundaries beyond the freely-available soil datasets such as SSURGO (Soil Survey Staff, 2016). In addition, irrigation depths and watering patterns may be updated based on the information related to soil hydraulic properties (i.e. wilting point and field capacity). Interestingly, the relative contribution of the remote sensing and hydrogeophysical data layers seemed to have less impact in predicting crop yield when compared to historical crop yield benchmark quantified by the leave one- cross validation analysis. We note from personal experience that this benchmark is often used implicitly by producers and crop consultants for agronomic decisions and represents the historic "local knowledge". Supplanting and augmenting this information from remote and proximal sensing datasets will be challenging and is likely a large contributing factor for the slower adoption of alternative yield metrics. However, with increasing farm size and distance of producers/managers to those farms, effective use of emerging datasets and techniques will be crucial for maximizing on-farm profitability.

The ability of the EOF framework to make predictions about crop yield using growing season weather information (aridity index) was also explored. It was found that the aridity index provided satisfactory models of $EC_1(t)$ and $\boldsymbol{Y}(t)$ in order to perform EOF reconstruction for crop yield with RMSE between 0.5–1.7 Mg/ha for maize, and 0.2–0.6 Mg/ha for soybean (comparable to MLR approach). Another important aspect for prediction using EOF vs other statistical models is our inability to make accurate long-term daily weather forecasts (months vs. weeks). We argue that aridity index is a more pragmatic prediction for long-term forecasts compared to daily weather given the current state of forecasting. With only 2–6 calibrated parameters needed to produce a 10m resolution crop yield prediction

(from approx. 7000 total cells) this approach represents a balance of model specificity and parsimony. Moreover, the EOF framework spatial representation can be used at multiple scales and or different geometries that are advantageous for matching constraints imposed by farming operations. However, given the inherent noise in crop yield data, we would suggest a certain minimum spatial resolution (here 10 m) and smoothing filter of the dataset in addition to the standard QA/QC discussed in Section 2.2.5. A key remaining challenge (particularly outside of countries with large commercial production) will be gaining access to historical yield maps. In regions with many small holder farms, remote sensing estimates of crop yield will likely need to be used as a surrogate (Burke and Lobell, 2017). A final challenge with this framework is better describing the $EC_1(t)$ and $\boldsymbol{Y}(t)$ relationships in order to reduce systematic bias in the crop yield predictions. From our EOF analysis we found about 10 % (6 of the 52 field years in Table 3) had large bias and three other field years (S6-2008, S6-2010, S7-2010) were clear outliers. Here, additional crop management information (i.e. planting density, applied nitrogen, pump failure, hail damage, replanting etc.) and/or growing season weather conditions (i.e. min. and max temperature, incoming solar radiation etc.) should be incorporated with a more robust statistical (Lobell et al., 2013, 2014) or crop model (e.g. Allen et al., 1998; Foster et al., 2017; Yang et al., 2013; Jones et al., 2003). Since the model needs only to describe $EC_1(t)$ and $\bar{\boldsymbol{Y}}(t)$ at the field scale a variety of simplified to more complex models (i.e. FAO56, AquaCrop, Hybrid-Maize, DSSAT, APSIM) may be selected. For example van Bussel et al. (2015) and Grassini et al. (2015) explore the minimum data needed to assess crop yield potential and gaps, and found that 3–5 years of crop yield data was needed to adequately describe connections with weather in an irrigated setting, or 5–8 years for a rainfed setting. However, since input values used in crop growth models will inevitably contain uncertainty (i.e., due to random and systematic measurement errors and spatial and temporal variation observed in many of these inputs), calibration of these models is critical to lead to better simulation of environmental response and further reduce bias and identify outliers.

5. Conclusions

In this work, we presented a statistical methodology to separate the spatial and temporal components of crop yield variation. Using a unique dataset of soils, topography and crop condition, we were able to quantify the relative importance of those datasets on understanding and predicting subfield crop yield, thus better quantifying the data utility. With respect to crop yield prediction, we found that historical yield maps are by far the best predictor, followed by crop condition (GCVI), SWC, and elevation data, depending on site. Soil texture (ECa) generally had the lowest importance but is useful in other agronomic decisions, such as determining irrigation depth. While the required crop yield datasets are available from harvest machinery, challenges of data access, format and privacy remain significant hurdles for use in the public sector, which may have to rely on remote sensing estimates of crop yield. We note that the presented methodology could be used by private companies and crop advisers with non-disclosure agreements given access to the yield data. In order to conduct a full economic analysis of the datasets, future work should quantify both the benefits and costs of acquiring, processing, and delivering the results to a producer. A key benchmark for evaluating the benefits of any new dataset, statistical approach, or crop model prediction, would be to outperform the leave-one-out cross validation using historical crop yield data. This benchmark can be used to quantify economic benefit and help determine the appropriate price point for acquiring new datasets.

CRediT authorship contribution statement

Trenton E. Franz: Conceptualization, Formal analysis, Writing - original draft, Funding acquisition, Supervision, Resources. **Sayli Pokal:** Software, Formal analysis, Writing - original draft, Writing - review & editing. **Justin P. Gibson:** Conceptualization, Writing - review & editing, Formal analysis. **Yuzhen Zhou:** Conceptualization, Software, Formal analysis, Writing - original draft, Writing - review & editing. **Hamed Gholizadeh:** Writing - original draft, Writing - review & editing. **Fatima Amor Tenorio:** Writing - original draft, Writing - review & editing. **Daran Rudnick:** Funding acquisition, Resources, Writing - original draft, Writing - review & editing. **Derek Heeren:** Funding acquisition, Resources, Writing - original draft, Writing - review & editing. **Matthew McCabe:** Writing - original draft, Supervision, Resources.

Matteo Ziliani: Writing - review & editing. **Zhenong Jin:** Writing - review & editing. **Kaiyu Guan:** Conceptualization, Funding acquisition, Writing - review & editing, Supervision, Resources. **Ming Pan:** Conceptualization, Funding acquisition, Writing - review & editing. **John Gates:** Conceptualization, Writing - review & editing. **Brian Wardlow:** Writing - review & editing, Supervision, Resources.

Declaration of Competing Interest — The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments — T.E.F. acknowledges the financial support of the USDA National Institute of Food and Agriculture, Hatch project #1009760 and project # 2019-67021-29312, as well as the Joint FAO/IAEA Programme of Nuclear Techniques in Food and Agriculture CRP D1.50.17. D.R. acknowledges the financial support of the USDA National Institute of Food and Agriculture, Hatch project #1015698. We would also like to thank Nathan Thorson of the Eastern Nebraska Research and Extension Center, the West Central Research and Extension Center, Paulman Farms, and Jacob Fritton of The Nature Conservancy for providing crop yield information, access to study sites, and liaison with private land owners. We also would like to thank Les Howard for providing the processed USGS DEM data, Yaping Cai for providing processed Landsat data, and Catherine Finkenbiner, William Avery and Matthew Russell for collecting hydrogeophysical surveys. M.F.M and M.Z. were supported by the King Abdullah University of Science and Technology.

References

Abdu, H., Robinson, D.A., Seyfried, M., Jones, S.B., 2008. Geophysical imaging of watershed subsurface patterns and prediction of soil texture and water holding capacity. Water Resour. Res. 44 https://doi.org/10.1029/2008wr007043. Wood18.

Allen, R.G., Pereira, L.S., Raes, D., Smith, M., 1998. Crop evapotranspiration. Guidelines for Computing Crop Water Requirements. FAO Irrigation and Drainage Paper 56. Food and Agriculture Organization of the United Nations, Rome, Italy.

Andreasen, M., Jensen, K.H., Desilets, D., Franz, T.E., Zreda, M., Bogena, H.R., Looms, M.C., 2017. Status and perspectives on the cosmic-ray neutron method for soil moisture estimation and other environmental science applications. Vadose Zone J. 16 (8), 11. https://doi.org/10.2136/vzj2017.04.0086.

Azzari, G., Jain, M., Lobell, D.B., 2017. Towards fine resolution global maps of crop yields: testing multiple methods and satellites in three countries. Remote Sens. Environ. 202, 129–141. https://doi.org/10.1016/j.rse.2017.04.014.

Belgiu, M., Dragut, L., 2016. Random forest in remote sensing: a review of applications and future directions. ISPRS-J. Photogramm. Remote Sens. 114, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011.

Binley, A., Beven, K., Elgy, J., 1989. A physically based model of heterogeneous hillslopes. 2. Effective hydraulic conductivities. Water Resour. Res. 25 (6), 1227–1233.

Binley, A., Hubbard, S.S., Huisman, J.A., Revil, A., Robinson, D.A., Singha, K., Slater, L.D., 2015. The emergence of hydrogeophysics for improved understanding of subsurface processes over multiple scales. Water Resour. Res. 51 (6), 3837–3866. https://doi.org/10.1002/2015wr017016.

Bogena, H.R., Huisman, J.A., Baatz, R., Franssen, H.J.H., Vereecken, H., 2013. Accuracy of the cosmic-ray soil water content probe in humid forest ecosystems: the worst case scenario. Water Resour. Res. 49 (9), 5778–5791. https://doi.org/10.1002/wrcr.20463.

Bolton, D.K., Friedl, M.A., 2013. Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. Agric. For. Meteorol. 173, 74–84. https://doi.org/10.1016/j.agrformet.2013.01.007.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32. https://doi.org/10.1023/a:1010933404324.

Brevik, E.C., Fenton, T.E., Lazari, A., 2006. Soil electrical conductivity as a function of soil water content and implications for soil mapping. Precis. Agric. 7 (6), 393–404. https://doi.org/10.1007/s11119-006-9021-x.

Budyko, M.I., 1974. Climate and Life. Academic Press, New York and London.

Burke, M., Lobell, D.B., 2017. Satellite-based assessment of yield variation and its determinants in smallholder African systems. Proc. Natl. Acad. Sci. U. S. A. 114 (9), 2189–2194. https://doi.org/10.1073/pnas.1616919114.

Chan, S., Njoku, E.G., Colliander, A., 2014. Soil Moisture Active Passive (SMAP), Algorithm Theoretical Basis Document, Level 1C Radiometer Data Product, Revision A. 20 pp. Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA.

Chatterjee, S., Price, B., 1977. Regression Analysis by Example. John Wiley & Sons, New York, NY.

Desilets, D., Zreda, M., Ferre, T.P.A., 2010. Nature's neutron probe: land surface hydrology at an elusive scale with cosmic rays. Water Resour. Res. 46. https://doi.org/10.1029/2009wr008726.

Finkenbiner, C.E., Franz, T.E., Gibson, J., Heeren, D.M., Luck, J., 2019. Integration of hydrogeophysical datasets and empirical orthogonal functions for improved irrigation water management. Precis. Agric. 20 (1), 78–100. https://doi.org/10.1007/s11119-018-9582-5.

Foster, T., Brozovic, N., Butler, A.P., Neale, C.M.U., Raes, D., Steduto, P., Fereres, E., Hsiao, T.C., 2017. AquaCrop-OS: an open source version of FAO's crop water productivity model. Agric. Water Manage. 181, 18–22. https://doi.org/10.1016/j.agwat.2016.11.015.

Franz, T.E., King, E.G., Caylor, K.K., Robinson, D.A., 2011. Coupling vegetation organization patterns to soil resource heterogeneity in a central Kenyan dryland using geophysical imagery. Water Resour. Res. 47 https://doi.org/10.1029/2010wr010127.W07531

Franz, T.E., Zreda, M., Rosolem, R., Ferre, P.A., 2012. Field validation of cosmic-ray soil moisture sensor using a distributed sensor network. Vadose Zone J. 11, 4. https://doi.org/10.2136/vzj2012.0046.

Franz, T.E., Wang, T., Avery, W., Finkenbiner, C., Brocca, L., 2015. Combined analysis of soil moisture measurements from roving and fixed cosmic ray neutron probes for multiscale real-time monitoring. Geophys. Res. Lett. 42, 3389–3396. https://doi.org/10.1002/2015GL063963.

Franz, T.E., Wahbi, A., Vreugdenhil, M., Weltin, G., Heng, L., Oismueller, M., Straub, P., Dercon, G., Desilets, D., 2016. Using cosmic-ray neutron probes to monitor landscape scale soil water content in mixed land use agricultural systems. Appl. Environ. Soil Sci. 2016. https://doi.org/10.1155/2016/4323742.

Franz, T.E., Loecke, T.D., Burgin, A.J., Zhou, Y.Z., Le, T., Moscicki, D., 2017. Spatiotemporal predictions of soil properties and states in variably saturated landscapes. J. Geophys. Res.-Biogeosci. 122 (7), 1576–1596. https://doi.org/10.1002/2017jg003837.

Gibon, F., Pellarin, T., Roman-Cascon, C., Alhassane, A., Traore, S., Kerr, Y., Lo Seen, D., Baron, C., 2018. Millet yield estimates in the Sahel using satellite derived soil moisture time series. Agric. For. Meteorol. 262, 100–109. https://doi.org/10.1016/j.agrformet.2018.07.001

Gibson, J., Franz, T.E., 2018. Spatial prediction of near surface soil water retention functions using hydrogeophysics and empirical orthogonal functions. J. Hydrol. 561, 372–383. https://doi.org/10.1016/j.jhydrol.2018.03.046.

Gibson, K.E.B., Gibson, J.P., Grassini, P., 2019. Benchmarking irrigation water use in producer fields in the US central Great Plains. Environ. Res. Lett. 14 (5), 8. https://doi.org/10.1088/1748-9326/ab17eb.

Gitelson, A.A., Gritz, Y., Merzlyak, M.N., 2003. Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves. J. Plant Physiol. 160 (3), 271–282. https://doi.org/10.1078/0176-1617-00887.

Grassini, P., van Bussel, L.G.J., Van Wart, J., Wolf, J., Claessens, L., Yang, H.S., Boogaard, H., de Groot, H., van Ittersum, M.K., Cassman, K.G., 2015. How good is good enough? Data requirements for reliable crop yield simulations and yield-gap analysis. Field Crop. Res. 177, 49–63. https://doi.org/10.1016/j.fcr.2015.03.004.

Gromping, U., 2006. Relative importance for linear regression in R: the package relaimpo. J. Stat. Softw. 17 (1).

Haberman, R., 1998. Elementary Applied Partial Differential Equations with Fourier Series and Boundary Value Problems, 3rd ed. Prentice Hall, Upper Saddle River.

Haghverdi, A., Leib, B.G., Washington-Allen, R.A., Ayers, P.D., Buschermohle, M.J., 2015. Perspectives on delineating management zones for variable rate irrigation. Comput. Electron. Agric. 117, 154–167. https://doi.org/10.1016/j.compag.2015.06.019.

Hawdon, A., McJannet, D., Wallace, J., 2014. Calibration and correction procedures for cosmic-ray neutron soil moisture probes located across Australia. Water Resour. Res. 50 (6), 5029–5043. https://doi.org/10.1002/2013wr015138.

Holzworth, D.P., Huth, N.I., Devoil, P.G., Zurcher, E.J., Herrmann, N.I., McLean, G., Chenu, K., van Oosterom, E.J., Snow, V., Murphy, C., Moore, A.D., Brown, H., Whish, J.P.M., Verrall, S., Fainges, J., Bell, L.W., Peake, A.S., Poulton, P.L., Hochman, Z., Thorburn, P.J., Gaydon, D.S., Dalgliesh, N.P., Rodriguez, D., Cox, H., Chapman, S., Doherty, A., Teixeira, E., Sharp, J., Cichota, R., Vogeler, I., Li, F.Y., Wang, E.L., Hammer, G.L., Robertson, M.J., Dimes, J.P., Whitbread, A.M., Hunt, J., van Rees, H., McClelland, T., Carberry, P.S., Hargreaves, J.N.G., MacLeod, N., McDonald, C., Harsdorf, J., Wedgwood, S., Keating, B.A., 2014. APSIM - evolution towards a new generation of agricultural systems simulation. Environ. Model. Softw. 62, 327–350. https://doi.org/10.1016/j.envsoft.2014.07.009.

Houborg, R., McCabe, M.F., 2018a. A Cubesat enabled Spatio-Temporal Enhancement Method (CESTEM) utilizing planet, Landsat and MODIS data. Remote Sens. Environ. 209, 211–226. https://doi.org/10.1016/j.rse.2018.02.067.

Houborg, R., McCabe, M.F., 2018b. Daily retrieval of NDVI and LAI at 3 m resolution via the fusion of CubeSat, Landsat, and MODIS data. Remote Sens. 10 (6), 23. https://doi.org/10.3390/rs10060890.

Jin, X.L., Kumar, L., Li, Z.H., Feng, H.K., Xu, X.G., Yang, G.J., Wang, J.H., 2018. A review of data assimilation of remote sensing and crop models. Eur. J. Agron. 92, 141–152. https://doi.org/10.1016/j.eja.2017.11.002.

Jones, J.W., Hoogenboom, G., Porter, C.H., Boote, K.J., Batchelor, W.D., Hunt, L.A., Wilkens, P.W., Singh, U., Gijsman, A.J., Ritchie, J.T., 2003. The DSSAT cropping system model. Eur. J. Agron. 18 (3–4), 235–265. https://doi.org/10.1016/s1161-0301(02)00107-7.

Kasampalis, D.A., Alexandridis, T.K., Deva, C., Challinor, A., Moshou, D., Zalidis, G., 2018. Contribution of remote sensing on crop models: a review. J. Imaging 4 (4), 19. https://doi.org/10.3390/jimaging4040052.

Kerkhoff, A.J., Martens, S.N., Milne, B.T., 2004. An ecological evaluation of Eagleson's optimality hypotheses. Funct. Ecol. 18 (3), 404–413. Kohli, M., Schron, M., Zreda, M., Schmidt, U., Dietrich, P., Zacharias, S., 2015. Footprint characteristics revised for field-scale soil moisture monitoring with cosmic-ray neutrons. Water Resour. Res. 51 (7), 5772–5790. https://doi.org/10.1002/2015wr017169.

Korres, W., Koyama, C.N., Fiener, P., Schneider, K., 2010. Analysis of surface soil moisture patterns in agricultural landscapes using Empirical Orthogonal Functions. Hydrol. Earth Syst. Sci. 14 (5), 751–764. https://doi.org/10.5194/hess-14-751-2010.

Leroux, L., Castets, M., Baron, C., Escorihuela, M.J., Begue, A., Lo Seen, D., 2019. Maize yield estimation in West Africa from crop process-induced combinations of multidomain remote sensing indices. Eur. J. Agron. 108, 11–26. https://doi.org/10.1016/j.eja.2019.04.007.

Li, Y., Guan, K.Y., Yu, A., Peng, B., Zhao, L., Li, B., Peng, J., 2019. Toward building a transparent statistical model for improving crop yield prediction: modeling rainfed corn in the U.S. Field Crop. Res. 234, 55–65. https://doi.org/10.1016/j.fcr.2019.02.005.

Lobell, D.B., Hammer, G.L., McLean, G., Messina, C., Roberts, M.J., Schlenker, W., 2013. The critical role of extreme heat for maize production in the United States. Nat. Clim. Change 3 (5), 497–501. https://doi.org/10.1038/nclimate1832.

Lobell, D.B., Roberts, M.J., Schlenker, W., Braun, N., Little, B.B., Rejesus, R.M., Hammer, G.L., 2014. Greater sensitivity to drought accompanies maize yield increase in the US Midwest. Science 344 (6183), 516–519. https://doi.org/10.1126/science.1251423.

Maestrini, B., Basso, B., 2018. Drivers of within-field spatial and temporal variability of crop yield across the US Midwest. Sci. Rep. 8. https://doi.org/10.1038/s41598-018-32779-3.

Mancini, F., Dubbini, M., Gattelli, M., Stecchi, F., Fabbri, S., Gabbianelli, G., 2013. Using Unmanned Aerial Vehicles (UAV) for high-resolution reconstruction of topography: the structure from motion approach on coastal environments. Remote Sens. 5 (12), 6880–6898. https://doi.org/10.3390/rs5126880.

Manfreda, S., McCabe, M.E., Miller, P.E., Lucas, R., Madrigal, V.P., Mallinis, G., Dor, E., Helman, D., Estes, L., Ciraolo, G., Mullerova, J., Tauro, F., de Lima, M.I., del Lima, J., Maltese, A., Frances, F., Caylor, K., Kohv, M., Perks, M., Ruiz-Perez, G., Su, Z., Vico, G., Toth, B., 2018. On the use of unmanned aerial systems for environmental monitoring. Remote Sens. 10 (4), 28. https://doi.org/10.3390/rs10040641.

McCabe, M.F., Rodell, M., Alsdorf, D.E., Miralles, D.G., Uijlenhoet, R., Wagner, W., Lucieer, A., Houborg, R., Verhoest, N.E.C., Franz, T.E., Shi, J.C., Gao, H.L., Wood, E.F., 2017a. The future of earth observation in hydrology. Hydrol. Earth Syst. Sci. 21 (7), 3879–3914. https://doi.org/10.5194/hess-21-3879-2017.

McCabe, M.F., Aragon, B., Houborg, R., Mascaro, J., 2017b. CubeSats in hydrology: ultrahigh- resolution insights into vegetation dynamics and terrestrial evaporation. Water Resour. Res. 53 (12), 10017–10024. https://doi.org/10.1002/2017wr022240.

McJannet, D., Franz, T.E., Hawdon, A., Boadle, D., Baker, B., Almeida, A., Silberstein, R., Lambert, T., Desilets, D., 2014. Field testing of the universal calibration function for determination of soil moisture with cosmic-ray neutrons. Water Resour. Res. 50 (6), 5235–5248. https://doi.org/10.1002/2014wr015513.

Peng, B., Guan, K.Y., Pan, M., Li, Y., 2018. Benefits of seasonal climate prediction and satellite data for forecasting US maize yield. Geophys. Res. Lett. 45 (18), 9662–9671. https://doi.org/10.1029/2018gl079291.

Peres-Neto, P.R., Jackson, D.A., Somers, K.M., 2005. How many principal components? Stopping rules for determining the number of non-trivial axes revisited. Comput. Stat. Data Anal. 49 (4), 974–997. https://doi.org/10.1016/j.csda.2004.06.015.

Perry, M.A., Niemann, J.D., 2007. Analysis and estimation of soil moisture at the catchment scale using EOFs. J. Hydrol. 334 (3–4), 388–404. https://doi.org/10.1016/j.jhydrol.2006.10.014.

Peters-Lidard, C.D., Clark, M., Samaniego, L., Verhoest, N.E.C., van Emmerik, T., Uijlenhoet, R., Achieng, K., Franz, T.E., Woods, R., 2017. Scaling, similarity, and the fourth paradigm for hydrology. Hydrol. Earth Syst. Sci. 21 (7), 3701–3713. https://doi.org/10.5194/hess-21-3701-2017.

Robinson, D.A., Binley, A., Crook, N., Day-Lewis, F.D., Ferre, T.P.A., Grauch, V.J.S., Knight, R., Knoll, M., Lakshmi, V., Miller, R., Nyquist, J., Pellerin, L., Singha, K., Slater, L., 2008. Advancing process-based watershed hydrological research using near-surface geophysics: a vision for, and review of, electrical and magnetic geophysical methods. Hydrol. Process. 22 (18), 3604–3635.

Rouse, J.W., Haas, R.H., Schell, J.A., Deering, D.W., 1973. Monitoring vegetation systems in the Great Plains with ERTS. In: Third ERTS Symposium. NASA SP-351, (Washington D.C.). pp. 309–317.

Samouelian, A., Cousin, I., Tabbagh, A., Bruand, A., Richard, G., 2005. Electrical resistivity survey in soil science: a review. Soil Tillage Res. 83 (2), 173–193.

Schron, M., Kohli, M., Scheiffele, L., Iwema, J., Bogena, H.R., Lv, L., Martini, E., Baroni, G., Rosolem, R., Weimar, J., Mai, J., Cuntz, M., Rebmann, C., Oswald, S.E., Dietrich, P., Schmidt, U., Zacharias, S., 2017. Improving calibration and validation of cosmicray neutron sensors in the light of spatial sensitivity. Hydrol. Earth Syst. Sci. 21 (10), 5009–5030. https://doi.org/10.5194/hess-21-5009-2017.

Schron, M., Rosolem, R., Kohli, M., Piussi, L., Schroter, I., Iwema, J., Kogler, S., Oswald, S.E., Wollschlager, U., Samaniego, L., Dietrich, P., Zacharias, S., 2018. Cosmic-ray neutron rover surveys of field soil moisture and the influence of roads. Water Resour. Res. 54 (9), 6441–6459. https://doi.org/10.1029/2017wr021719.

Shah, S.H., Angel, Y., Houborg, R., Ali, S., McCabe, M.F., 2019. A random forest machine learning approach for the retrieval of leaf chlorophyll content in wheat. Remote Sens. 11 (8), 26. https://doi.org/10.3390/rs11080920.

Sharma, V., Irmak, S., 2012a. Mapping spatially interpolated precipitation, reference evapotranspiration, actual crop evapotranspiration, and net irrigation requirements in Nebraska: part I. Precipitation and reference evapotranspiration. Trans. ASABE 55 (3), 907–921.

Sharma, V., Irmak, S., 2012b. Mapping spatially interpolated precipitation, reference evapotranspiration, actual crop evapotranspiration, and net irrigation requirements in Nebraska: part II. Actual crop evapotranspiration and net irrigation requirements. Trans. ASABE 55 (3), 923–936.

Soil Survey Staff, 2016. Soil Taxonomy: a Basic System of Soil Classification for Making and Interpreting Soil Surveys. U.S. Department of Agriculture Handbook 436, 2nd ed. [online] Available from:. 2nd ed. Natural Resources Conservation Service. http://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/nrcs142p2_051232.pdf.

Soofi, E.S., Retzer, J.J., Yasai-Ardekani, M., 2000. A framework for measuring the importance of variables with applications to management research and decision models. Decis. Sci. 31 (3), 595–625. https://doi.org/10.1111/j.1540-5915.2000.tb00936.x.

van Bussel, L.G.J., Grassini, P., Van Wart, J., Wolf, J., Claessens, L., Yang, H.S., Boogaard, H., de Groot, H., Saito, K., Cassman, K.G., van Ittersum, M.K., 2015. From field to atlas: upscaling of location-specific yield gap estimates. Field Crop. Res. 177, 98–108. https://doi.org/10.1016/j.fcr.2015.03.005.

Vina, A., Gitelson, A.A., Nguy-Robertson, A.L., Peng, Y., 2011. Comparison of different vegetation indices for the remote assessment of green leaf area index of crops. Remote Sens. Environ. 115 (12), 3468–3478. https://doi.org/10.1016/j.rse.2011.08. 010.

Yang, H.S., Dobermann, A., Cassman, K.G., Walters, D.T., Grassini, P., 2013. Hybrid- Maize (v.2013.4). A Simulation Model for Corn Growth and Yield. Nebraska Coop. Extension, Univ. Nebraska-Lincoln, Lincoln, NE.

Ziliani, M.G., Parkes, S.D., Hoteit, I., McCabe, M.F., 2018. Intra-season crop height variability at commercial farm scales using a fixed-wing UAV. Remote Sens. 10 (12), 25. https://doi.org/10.3390/rs10122007.

Zreda, M., Shuttleworth, W.J., Xeng, X., Zweck, C., Desilets, D., Franz, T.E., Rosolem, R., 2012. COSMOS: the COsmic-ray soil moisture observing system. Hydrol. Earth Syst. Sci. 16, 4079–4099. https://doi.org/10.5194/hess-16-1-2012.