Computers and Electronics in Agriculture 208 (2023) 107768

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Preseason crop type prediction using crop sequence boundaries

Jonathon Abernethy [a,*], Peter Beeson [b], Claire Boryan [a], Kevin Hunt [a], Luca Sartore [a,c]

[a] USDA NASS, 1400 Independence Ave SW, Washington, DC 20250, USA
[b] Global Conservation Institute, 156 Thankhohay Poe, Santa Fe, NM 87506, USA
[c] NISS, 1750 K Street NW Suite 1100, Washington, DC 20006, USA

ARTICLE INFO

Keywords:
Crop type prediction
Machine learning
Cropland data layer
Big Data
Data compression

ABSTRACT

Preseason crop-type prediction has emerged as a valuable tool for agricultural use. A reliable algorithm for early crop-type prediction has many applications, including crop mapping, planted acreage prediction, crop yield prediction, disaster response, area sample design, crop survey imputation, and more. The primary source of data for preseason crop prediction in the United States is the United States Department of Agriculture National Agricultural Statistics Service's Cropland Data Layer, which is an annual crop specific land cover data set produced using satellite imagery and administrative data. Historical crop rotations taken from the Cropland Data Layer can be used by machine learning models to predict the future crop type in any given land area. The dataset obtained from the Cropland Data Layer is large, containing hundreds of millions of pixels per state. Current approaches for predictive modeling have utilized sampling, resulting in more scalable machine learning. In this paper, the authors propose an alternative method that uses all available Cropland Data Layer data in a rapid and memory-efficient manner. The proposed method relies on a novel technique for identifying groups of pixels with homogenous cropping history. These pixel groups are summarized as polygons representing field boundaries, referred to as crop sequence boundaries. Use of these new polygons for modeling significantly reduces the computational burden of incorporating all the Cropland Data Layer data and eliminates any increased uncertainty brought on by sampling. This novel polygon-based approach competes well with existing methods in scalability and accuracy, achieving the highest overall accuracy in 23 out of 24 tests performed.

1. Introduction

The United States Department of Agriculture's (USDA's) National Agricultural Statistics Service (NASS) is mandated to provide timely and accurate statistics in service to the domestic agriculture economy as well as to inform producers and industries that support and benefit from it. Various products are provided to meet this role. One of these products is the Cropland Data Layer (CDL) (Boryan et al., 2011). The CDL program uses inputs including satellite imagery, farmer-reported data, and other ancillary data to produce a crop-specific, raster-formatted, geo-referenced, crop type map at 30 m resolution (900 square meters) for the continental United States (Boryan et al., 2011).

Accurate crop land cover maps like the CDL are crucial for decision making in Agriculture. Reliable crop maps can aid in planted acreage estimation and prediction (Lark et al., 2017; Zhang et al., 2019a), small area estimation (Wang et al., 2018), area frame stratification (Boryan and Yang, 2017), disaster monitoring (Boryan et al., 2018), yield prediction (Johnson, 2014) and many other applications (Yaramasu et al., 2020; Zhang et al., 2019a). The CDL for a given year is released to the public around February of the following year. Consequently, the CDLs are not publicly available during the current growing season.

The absence of CDL information during the current planting season has inspired a wide variety of research to produce CDL like analogues. These efforts can be divided into pre-season and in-season crop mapping. In-season mapping typically refers to the prediction of crop maps sometime after planting but before harvest. The in-season case is a relatively mature field, which has been covered extensively, see, for example, Johnson and Mueller (2021); Orynbaikyzy et al. (2019); Rauf et al., (2022); You and Dong (2020); Zhang et al. (2021a), and the references therein.

Pre-season crop mapping, on the other hand, assumes that crop predictions are needed before planting occurs. This means that unlike the in-season case, satellite imagery and other in-season data are not available to train models. Typically, the only available data that can be used for prediction are the historical crop rotations from previous CDLs and possibly time invariant data for the area of interest. This leads to the

* Corresponding author.
E-mail address: jake.abernethy@usda.gov (J. Abernethy).

https://doi.org/10.1016/j.compag.2023.107768
Received 7 December 2022; Received in revised form 6 March 2023; Accepted 8 March 2023
Available online 20 March 2023

following question: can a model be created to perform pre-season crop mapping using rotations derived from previous years' CDLs?

While still a relatively new field, pre-season rotation modeling using the CDL has seen an abundance of recent work. In particular, Zhang et al. (2019a), use rotations from historical CDLs and train an artificial neural network to predict the current season crop type for a variety of agricultural districts in the US corn belt. As a sampling strategy Zhang et al. (2019a) splits states by agricultural district and then collect training rotations by taking a one-tenth subset of single pixel temporal windows from each district. Each pixel comes with a nine-year temporal window, where the crop types in the first eight years are used to predict the ninth year. This method of crop type prediction was also used for trusted pixel selection (Zhang et al., 2021a) and agro-geoinformation discovery (Zhang et al., 2021b).

Another deep learning approach to pre-season crop prediction is taken in Yaramasu et al. (2020), where a spatiotemporal encoder and decoder setup is used for the prediction model. The authors demonstrate that their model can take spatial information into account and is superior to a Markovian model in the state of Nebraska. Unlike the approach of Zhang et al. (2019a), where the unit of analysis is a single pixel, Yaramasu et al. (2020) use randomly selected 512 by 512 square regions of pixels for training. Using the squares may help the model better learn spatial features (Yaramasu et al., 2020).

Apart from neural networks, other machine learning approaches, including random forests (Johnson and Mueller, 2021; Yao and Di, 2021), extreme gradient boosting (Yao and Di, 2021), and naïve Bayes (Yao and Di, 2021) have been used to solve the pre-season crop mapping problem. Like the approach of Zhang et al. (2019a), subsets of single pixel time series are used as units of analysis (Johnson and Mueller, 2021). In particular, the area of interest is split into counties and a 0.25 percent sample of pixels is taken (Johnson and Mueller, 2021). Finally, all single pixel time series are used in Yao and Di (2021), where the area of interest is a single county with about 1.2 million pixels.

Each example of pre-season crop type prediction discussed so far is at the pixel level. A potential drawback with pixel-level modeling is that the pixel is typically not the spatial unit where the land cover change occurs (Ballestores and Qiu, 2012; Irwin and Geoghegan, 2001; National Council, 2013; Sohl et al., 2017). The true spatial unit of change is referred to as a parcel, and in agricultural applications corresponds to the farm field (Sohl et al., 2017). Therefore, a natural approach to pre-season crop type prediction is to group pixels into field polygons representing areas of common management and then train machine learning models to make predictions at the field polygon scale. The field polygon-based approach to prediction is the one taken in this paper.

Previous examples of field-level crop type prediction leveraging historical crop rotations are based on Markov chains and are typically applied outside of the United States (Aurbacher and Dabbert, 2011; Osman et al., 2015; Xiao et al., 2014). These models are fit on relatively small datasets using fields taken from administrative or survey data, the analogues of which are not available to the public in the US. Furthermore, it has been demonstrated that other, more recent, machine learning approaches can have superior performance when compared to Markov chains (Yaramasu et al., 2020).

Field-based landcover predictions have also been used as a component in the Forecasting Scenarios of Land-use Change (FORE-SCE) model (Dornbierer et al., 2021; Sohl et al., 2019, 2017). These models typically use linear logistic regressions to assign landcover probabilities and do not use historical crop rotations as predictor variables (Dornbierer et al., 2021; Sohl et al., 2019, 2017). In many cases, the fields used in these models are not publicly available, so old boundaries from 2008 are used (Dornbierer et al., 2021; Sohl et al., 2019, 2017). Furthermore, these fields boundaries can contain multiple crops (Sohl et al., 2017). As an alternative to administrative field boundaries, synthetic boundaries have also been used, but the fields derived from these boundaries may also contain multiple crop types (Yan and Roy, 2014).

In summary, current models leveraging machine learning and historical crop rotations for pre-season crop type prediction are pixel based. As an alternative to pixel-based analysis, the field can be used as a potentially closer analogue to the true decision-making unit for predicting crop type. Current field-based approaches either use outdated field boundaries, boundaries that can include multiple crop types, or have not been scaled to large land areas. These approaches have also not leveraged machine learning and historical CDL crop rotations in a comparable way to the pixel-based work (Johnson and Mueller, 2021; Yao and Di, 2021; Yaramasu et al., 2020; Zhang et al., 2019a).

This research adds a novel approach to pre-season crop type prediction by combining a machine learning model with synthetically derived field polygons. The fields are obtained from polygons based on consistent cropping sequences identified in the set of historical CDLs through a modified version of the algorithm found in Beeson et al. (2020). The field polygons are referred to as crop sequence boundaries (CSBs). The CSBs are designed so that every pixel within the boundary has the same crop rotation history. The CSBs are used in place of the CDL pixels as the primary unit of modeling to implement a novel field-level CSB-based machine learning model.

Crop sequence boundary-based models have several benefits. The first is scalability. The common rotation history within each CSB polygon allows for an efficient compression of the pixel-based CDL stack to a set of polygons with attributes. The number of CSB polygons is much smaller than the number of CDL pixels, which allows for efficient modeling of large land areas including multiple states or even the continental United States. The second benefit is timeliness. Because the boundaries are synthetic and CDL based, the boundary lag is only one year, as opposed to relying on potentially old administrative boundaries. The third is prediction accuracy, as predictions from the CSB-based model are found to have consistently higher accuracies with respect to USDA Farm Service Agency (FSA) in situ ground reference common land unit data than two alternative pixel-based models.

The paper is organized as follows: in Section 2 the study area is presented, the CDL and FSA data are introduced, and the novel CSB-based predictive modeling framework is described. The metrics used for evaluating model performance are also described in Section 2. In Section 3, experiments comparing the new field polygon-based approach with selected existing methods from the literature are conducted and the results are presented. In Section 4 the results are further discussed, and the benefits and limitations of field-level pre-season modeling are examined. Finally, Section 5 includes the conclusion and a description of future work.

2. Materials and methods

2.1. Study area and crop types

To make a reliable comparison between models, it is important to reduce outside sources of error as much as possible. In this case, outside error comes from the CDLs, which are not perfectly accurate (Boryan et al., 2011). Error rates can be found on the Cropland Data Layer metadata page (USDA-NASS, 2022). Minimizing pixel-level CDL error is important as the CDLs are used for the predictor variables, response variable and model selection (cross validation). For this reason, a study area with reliable CDL pixels is desirable.

The study area was selected by identifying states with a CDL principal crop accuracy of at least 89.5%, for all years 2008–2016. Crop types with both user and producer accuracy at least 89.5% for all years 2008–2016 are selected within these states, provided the crop types are identified in at least 1000 CDL pixels each year. Note that these criteria can be checked using the CDL metadata (USDA-NASS, 2022). There are eight states that constitute the study area: Illinois, Indiana, Iowa, Louisiana, Minnesota, Missouri, Nebraska, and Ohio (Table 1). Within each state, the crop types that satisfy the accuracy criteria are used. Crops grown vary by state, but include corn, soybeans, rice, sugarcane, spring wheat, cotton, and sugar beets. Land cover types that do not meet the

Table 1
States and crop types in study area with percent coverage.

State	Type 1	Type 2	Type 3	Type 4	Type 5
Illinois	Corn	Soybean	Other		
	49%	46%	5%		
Indiana	Corn	Soybean	Other		
	45%	49%	6%		
Iowa	Corn	Soybean	Other		
	55%	40%	5%		
Louisiana	Corn	Rice	Sugarcane	Other	
	15%	15%	12%	58%	
Minnesota	Corn	Soybean	Spring Wheat	Sugar beet	Other
	41%	40%	7%	2%	10%
Missouri	Corn	Soybean	Cotton	Rice	Other
	26%	43%	26%	2%	3%
Nebraska	Corn	Soybean	Other		
	49%	28%	23%		
Ohio	Corn	Soybean	Other		
	35%	50%	15%		

accuracy criteria are coded as "other".

The coverage of the non-other crop types in the study area is typically high for each state, usually at least 75% (Table 1). The only exception is Louisiana, where soybeans (a major crop) were not included as they technically failed the CDL user-accuracy threshold in 2009. Note that the percentages in Table 1 are calculated based on the total acreage of all field crops including potatoes planted in the state.

2.2. Cropland data Layer

NASS produces the Cropland Data Layer, which is crop-specific, raster-formatted, geo-referenced, land cover data set at 900 square meter resolution (Boryan et al., 2011). The CDL program inputs include medium resolution satellite imagery acquired throughout the summer growing season; Farm Service Agency Common Land Unit farmer-reported data as ground reference information, and other ancillary data such as the United States Geological Survey's (USGS's) National Land Cover Data set (Yang et al., 2018). A decision tree-supervised classification method is used to generate the state-level crop specific classifications.

The first state-level CDL (North Dakota) was produced in 1997. Cropland Data Layers for the conterminous U.S. have been produced annually since 2008. Total crop identification accuracies for the CDLs are well documented in the literature and generally vary from 85% to 95% for the major agricultural commodities. For detailed information on the CDL, see Boryan et al. (2011). The CDLs are available to the public for download and on-line analysis on the CropScape web application (Han et al., 2012; Zhang et al., 2019b). The eight states used in this study are mostly in corn and soybean dominated regions of the USA (Fig. 1).

For each year, each CDL pixel has a label, which identifies the land cover type for that year. For example, a value of 1 refers to corn, a value of 5 refers to soybeans, and a value of 176 refers to pasture. The land cover grids provided by the CDL are used as inputs to create fields with homogenous crop type.

2.3. Common land unit and 578 administrative data

USDA Farm Service Agency (FSA) Common Land Unit (CLU) data are a standardized GIS layer of U.S. crop fields supporting commodity and conservation programs. The FSA CLU data are updated every growing season when farmers report crop type and acreage for their fields to FSA county offices. The farmer reports are included in FSA 578 administrative databases which are then linked to the FSA CLU polygons. Data for each program participant typically arrive later in the season (July and August). The FSA CLU and 578 data are administratively confidential and not available for public dissemination (Heald, 2002; USDA-FSA, 2017). These data only include program participants, so the coverage of all planted acres is not complete. Planted acres for crops considered in the study area (Section 2.1) are well covered, however, with the lowest coverages ranging from 95% to almost 100%.

The FSA CLU data are used as ground reference evaluation data for crop categories in this study. In particular, the FSA data are used to compare the predictive accuracy of the proposed approach to alternatives from the existing literature. Note that the FSA data are only used to evaluate model performance, not to train models. This is because, as

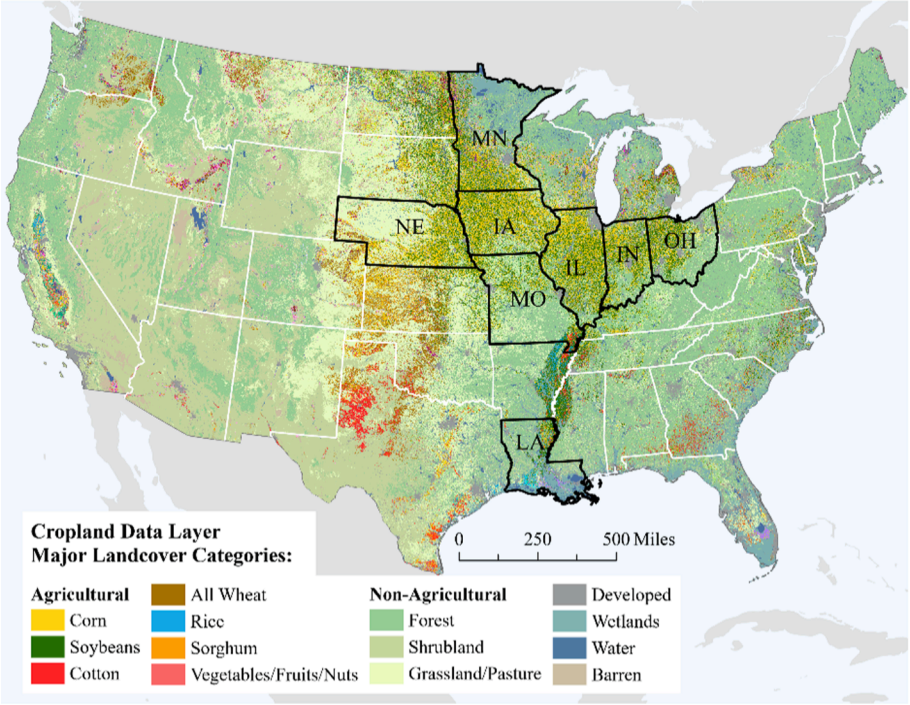


Fig. 1. The NASS Cropland Data Layer product with highlighted (black state boundaries) study area.

mentioned, these data are not publicly available and do not provide full coverage of all planted acres. The ground reference status, however, guarantees that differences in model performance are not simply artifacts of CDL error, which makes the FSA a better tool for comparison of model predictions than the CDL.

2.4. A field polygon centered approach to crop type prediction

In this section, a new procedure for crop type prediction using units of consistent cropping sequences derived from the CDL is proposed. The first step is representative field creation, where homogeneous cropping polygons are created using historical CDLs as input. The second step is the creation of a tabular dataset containing attributes for each polygon. The final step includes training a machine learning model on this tabular dataset. The trained model can predict the future crop type inside each field polygon.

This approach is based on three major assumptions. The first is that polygons of consistent cropping sequence can be derived to use as inputs, so that the unique rotation history within each polygon can be used as input to train a machine learning model. One procedure to accomplish this is reviewed in section 2.4.1 and 2.4.2. The second assumption is that the polygon boundaries do not change much during the year for which the prediction is needed. For example, if a field is split in two during the prediction year, assuming a single crop type within the boundary may cause prediction errors. While this assumption is difficult to verify directly, the competitive prediction accuracy established in the experiments when compared to purely pixel-based approaches in Section 3 suggest error due to field boundary change is minimal. Finally, the third assumption is that farmers rarely break their rotations. Any model that uses historical crop rotations to predict future planting choices relies on the assumption that rotations are stable, so that past cropping activity can predict future decisions. Violations of this assumption include rotation breakage due to outside factors such as unusual economic incentives or weather activity. Accounting for these is out of the scope of this study.

2.4.1. Field definition

The first step is to derive field polygons from the CDL. To do this, properties of an ideal field must be defined. For the pre-season prediction application, the properties are:

1. A contiguous non-multi-segmented polygon
2. A common cultivated landcover type within the boundary each year
3. The polygons exist over a fixed time window
4. Complete coverage of the cultivated area of interest

The first and second requirements are to correct for potential CDL noise. For example, if a single soybean pixel exists in a much larger homogenous collection of corn pixels, the soybean pixel is assumed to be erroneous and is labeled as part of the larger cornfield.

The second and third requirements exist so that the field polygons can be converted to a tabular dataset where the rows are the polygon identifiers (ID), and the columns are the unique crop type for each year. This is ensured when all pixels share a common crop type for each year (see Section 2.4.2 and 2.4.3). Note that the crop type can vary between years, but not between pixels within the field boundary during a fixed year. Finally, for the fourth requirement, the proposed fields should have complete coverage of the area of interest, so that predictions can be made anywhere desired.

As an example, an acceptable field over three years would have all CDL pixels within the boundary following a $C-S-C$ rotation, where C refers to corn and S refers to soybeans. Note that the crop type may vary between years one, two, and three, but is constant within year. Other acceptable examples would be $C-C-C$ and $C-S-W$ (where W refers to "wheat"). An example of an unacceptable field boundary would be a rotation like $C-C \cup S-C$. The $\cup$ notation denotes "both" (i.e., two distinct crop types). This field boundary is unacceptable because there is at least one year where the within field crop type is not constant. Other unacceptable examples would be $C \cup S-C \cup S-C \cup S$ and $C \cup S \cup W-C-C$.

As another example, consider the FSA CLUs described in Section 2.3. The CLUs are one type of 'field' that could potentially be used as a land unit for predictive modeling. However, the FSA CLUs violate the ideal field requirements because a single CLU can contain multiple crop types. The CLUs also do not have a fixed time of existence, as their creation and destruction depend on changes in reporting to FSA. One CLU could be ten years old while another is defined during the current year. Any predictive model utilizing the CLUs would have to account for the crop heterogeneity and variable time intervals. Finally, as mentioned in Section 2.3, the coverage of the CLUs, while high in many areas, is not complete. This means that there are areas of land where a CLU-based model would not be able to make predictions. These issues, as well as the fact that the FSA CLUs are administratively confidential and not available to the public, suggest that a new set of crop field polygons with desirable properties for predictive modeling would be useful.

2.4.2. Crop sequence Boundaries: Field polygons derived from the CDL

Given the requirements defined above, an algorithm to generate the field boundary polygons is required. The methodology chosen is very similar to that used in Beeson et al. (2020) and expanded in Hunt et al. (2022). In Beeson et al. (2020), the polygons are referred to as crop management units (CMUs). To summarize, the method includes:

1. Cleaning each CDL with minimal filtering
2. Stacking the chosen years into unique combinations of crop types
3. Converting them to polygons
4. Cleaning the noise.

Two changes to the original CMU algorithm are made. The first is that an eight-year window is used instead of an eleven-year window. The reason for this change is that the eight-year window enables comparisons to previous work, which forecasts 2016, 2017, and 2018. An eleven-year window only allows forecasts for 2019 and later. The second modification includes eliminating small polygons left on the edges of larger fields instead of buffering. This was done because buffering dilates then erodes large polygons, which causes a loss in acreage. Like the CMUs, the new polygons are set to be 2.5 acres or larger to reduce noise from the CDL. Note that most fields in the United States tend to be larger than 2.5 acres, with the median field-size increasing over time (White and Roy, 2015).

The modified CMUs described above will be referred to as crop sequence boundaries (CSBs). The new CSBs represent fields of homogenous cropping area. Each CSB contains an attribute table with the required information for a field-level crop prediction model. This information includes the crop type planted for years one to eight, the location of the polygon, and the area of the polygon.

These CSBs were originally created for multiple purposes including summarizing cropland into zones of persistent cropping sequences as required in this study. In Jennings (2022), an example of their varied use is described in a georeferencing process seeking to identify agricultural fields that are not registered with any FSA CLU program. The CSBs were developed to represent complete coverage of the contiguous U.S. which aids in identifying these types of fields. As the CSB project develops these data layers are being prepared for distribution as outlined in Hunt et al. (2022) with the intention of a product released in the public domain in the future (Fig. 2).

2.4.3. Tabular data set from field polygons

Once the CSBs are created, they need to be represented in a format that is suitable for use in a model. This can be done by giving each polygon a unique ID. The set of CSBs can then be represented as a table, where each row is an ID, and each column is an attribute of the field

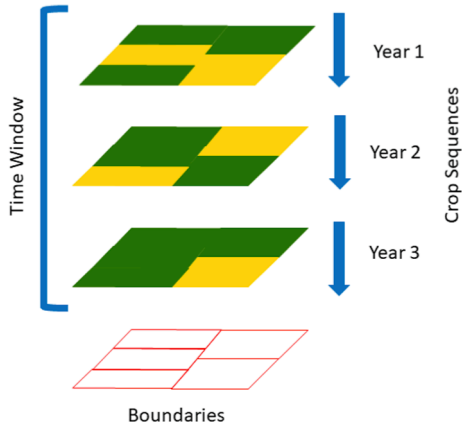


Fig. 2. A small example of crop sequence boundaries (red) from a hypothetical three-year time series of corn (yellow) and soybean (green) pixels. Note that the "L" shaped soybean boundary in year three is made up of four different fields. Close examination shows that each field boundary polygon includes pixels with the same within-year crop type, while the between year crop type does not need to be the same. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

represented by the ID. The attributes in this study include the crop type planted in the polygon for each year in the time series and the area of the CSB polygon. An example is given for the setup implied by Fig. 2 (Table 2). Other attributes are also possible under this framework, such as the location coordinates of the field centroid and any other potentially explanatory data the analyst has on hand.

Once in a tabular format (Table 2), the CSB-level data can be input into any standard multi class classification algorithm to train a model. Predictions from the trained model can be used to predict the future crop type planted to each field. Note that this all works because, as mentioned in Section 2.4.1, the CDL pixels in the CSB polygons can only contain one crop type each year. This means the predictor and response variables are unique within each CSB.

2.4.4. Machine learning with CSBs for Pre-Season forecasting

Given the CSB based historical crop rotations described in sections 2.4.1-2.4.3, the objective is to forecast the next crop type to be planted in each boundary. In particular, the question is whether the current years' crop type within a CSB can be forecasted using the rotation history within that CSB. Many potential options exist to complete this task, including the machine learning based approach described in this section.

One option to complete this task is to create a rules-based classifier. For example, if a CSB follows a basic rotation like corn-soybean-corn-soybean, one would expect the crop in year five to be corn. There are several drawbacks to this approach. The first is that not all rotations are as simple as the corn-soybean rotation. For example, consider a rotation that looks like corn-corn-corn-soybeans-corn. There are two reasonable rules here. The first is that this is a monocrop rotation with noise (maybe the farmer was prevented from planting corn in year four due to the weather, for example), so the reasonable rule for year six would be corn. It is also possible, however, that the farmer is switching from a monocrop rotation to a corn-soybean rotation, which would make the reasonable rule soybeans. The second issue with rule-based classifiers is that the number of potential rules increases exponentially with the number of crop types and the length of the rotation history. For example, a six-year rotation with the only possibilities being corn, soybeans, and other would require 729 rules. A third issue is that rule-based classifiers do not provide measures of uncertainty. Even a stable corn-soybean rotation has some probability of breaking, so even if it is possible to make a guess with the rule-based approach, one could not express a quantitative level of certainty to that guess. Finally, rule-based approaches do not handle the addition of extra continuous data well. For example, when using the CSBs the size and location of the fields are available as potential covariates. It is not clear how one would integrate these data into a rule-based approach.

Table 2
Tabular dataset implied by CSBs from hypothetical three year time series of corn and soybean pixels.

CSB ID	Crop Year 1	Crop Year 2	Crop Year 3	Area (Acres)
1	soybeans	soybeans	soybeans	1
2	corn	soybeans	soybeans	1
3	soybeans	corn	soybeans	1
4	soybeans	corn	soybeans	1.5
5	corn	soybeans	corn	1.5

The drawbacks of the manual rule-based approach suggest the need for an alternative. An ideal algorithm would avoid manual user defined rules by taking the tabular CSB based rotation data as input and using it to automatically derive forecasts. This is exactly the purview of supervised machine learning. A machine learning model is fed CSB level training data. The training data contains predictor variables and a response variable. The predictors include the cropping history in the CSB for years $1, 2, \cdots, T-1$. The response variable is the crop type planted in year T. Using this labeled training data, the machine learning model can automatically generate rules that map a given rotation to a forecasted crop type. The automation of the generation of rules using machine learning means the user does not need to make a manual decision for every possible rotation. Furthermore, most machine learning models can generate probabilities as measures of uncertainty and can easily incorporate continuous data like field area. All of these benefits point toward machine learning as the best approach for making CSB level crop type forecasts.

The first step to taking a machine learning based approach to CSB based crop type forecasting is to select a model. There are a variety of modeling options available, in fact, any multiclass classification algorithm that accepts tabular data can be used to obtain a predictive model from the CSB-level dataset. Previous work in crop type forecasting have used artificial neural networks Zhang et al. (2019a), random forests Johnson and Mueller (2021), and deep learning Yaramasu et al. (2020).

Another option is to use gradient boosting, which is the machine learning model selected for this study. In particular, the gradient boosting decision function for crop type k is:

$$F_k\left(\vec{x}_i\right) = \sum_{m-1}^{M} f_{km}\left(\vec{x}_i\right)$$

The data $\vec{x}_i$ in this case are the cropping history of a given CSB (the i^{th} CSB) and potentially other useful variables like the size of the CSB. In other words, $\vec{x}_i = \langle C_{i1}, C_{i2}, \cdots, C_{iT-1}, A_i \rangle$ with C_{it} referring to the crop type planted in CSB i during year t and A_i referring to the area of CSB i. The decision function for each crop type F_k is a sum of weak learners (typically decision trees) f_{km}. The probability that the future crop type planted in the CSB is k^* is:

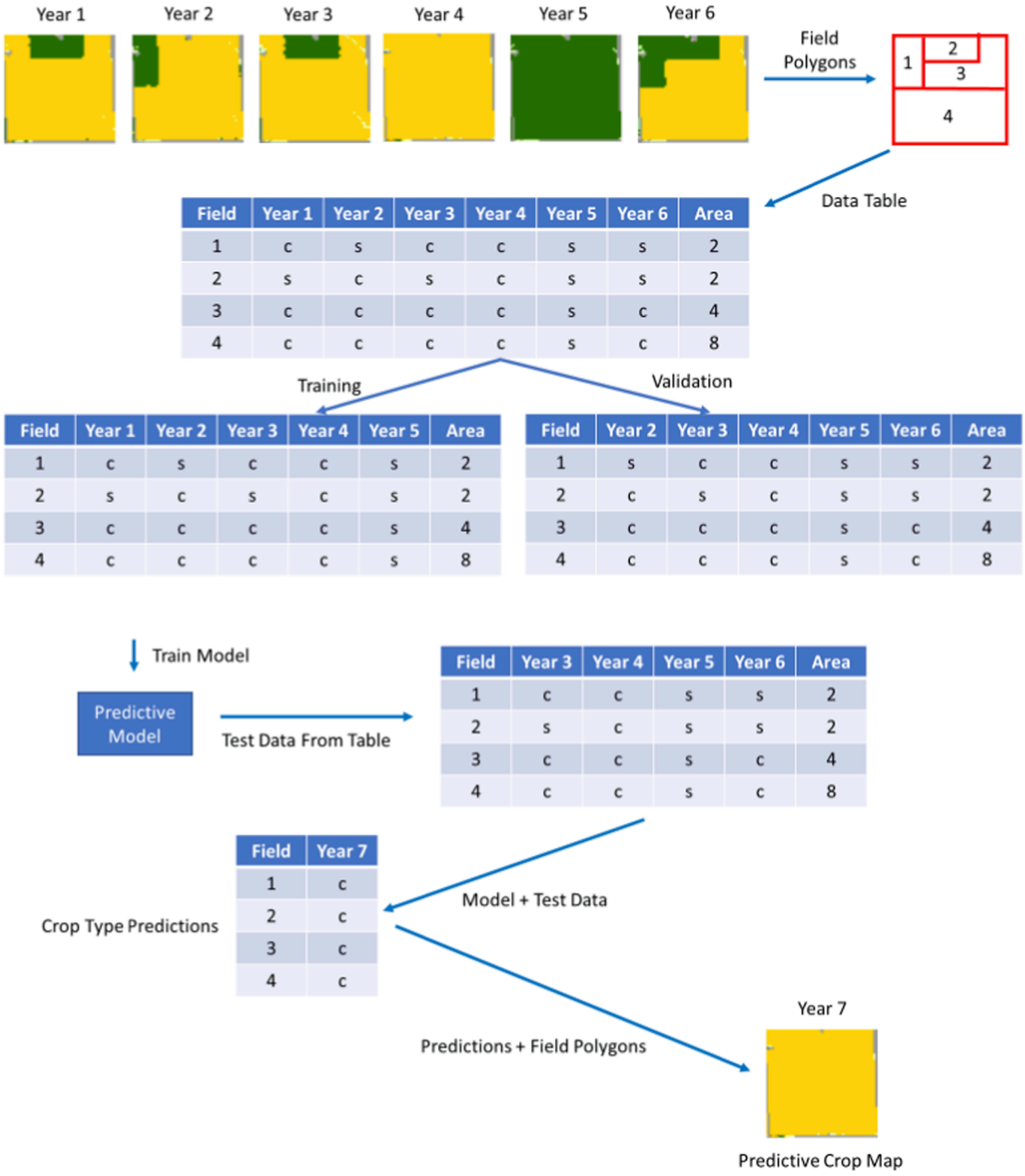
$$p(C_{iT} = k) = \frac{e^{F_{k^*}\left(\vec{x}_i\right)}}{\sum_{k=1}^{K} e^{F_k\left(\vec{x}_i\right)}}$$

The derivation of the weak learners f_{km} depends on the gradient boosting algorithm used. For this study, we use the LightGBM package as the boosting method (Ke et al., 2017). Gradient boosting has been demonstrated to be an effective classification algorithm for tabular data (Shwartz-Ziv and Armon, 2022). Moreover, LightGBM has favorable computing speed when compared to competing gradient boosting packages along with easy handling of categorical features.

Given the selected LightGBM model, the next step is to select the predictor variables included in the training data. The first set of variables is the rotation history. Six years of rotation history are included,

with the allowable crop types taken from Table 1. In other words, six categorical variables taking values within the set of possible crop types are included as predictors. In addition to the CSB's rotation history, the area of the CSB was also included in the model. The rationale being that larger fields may have more stable rotations than smaller ones. The last attribute included was the agricultural statistics district (ASD) in which the field was located. Note that ASDs, as defined by USDA NASS, are regions within a state that contain similar agriculture. See USDA-NASS (2007) for the exact counties within each ASD. Finally, the response variable was the crop type planted in the seventh year.

Most machine learning models, LightGBM included, have several tuning parameters, so hyperparameter selection via cross validation is required. The eight-year structure of the CSB polygons (Section 2.4.1) and the use of a six-year rotation model imply the procedure described next. A set of candidate models were trained, each with a different hyperparameter combination, by using years one through six to predict year seven. The performance of these models was evaluated by using years two through seven to predict year eight. Finally, the best performing hyperparameters were used to train the final model on the same set of years (rotation history coming from years two through seven with year eight as the response variable). Given the final model, years three through eight were used to forecast the crop type for the unknown year nine. An example of the cross-validation and forecasting procedure is included in Fig. 3 in Section 2.4.5.



Field	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Area
1	c	s	c	c	s	s	2
2	s	c	s	c	s	s	2
3	c	c	c	c	s	c	4
4	c	c	c	c	s	c	8

Field	Year 1	Year 2	Year 3	Year 4	Year 5	Area
1	c	s	c	c	s	2
2	s	c	s	c	s	2
3	c	c	c	c	s	4
4	c	c	c	c	s	8

Field	Year 2	Year 3	Year 4	Year 5	Year 6	Area
1	s	c	c	s	s	2
2	c	s	c	s	s	2
3	c	c	c	s	c	4
4	c	c	c	s	c	8

Field	Year 3	Year 4	Year 5	Year 6	Area
1	c	c	s	s	2
2	s	c	s	s	2
3	c	c	s	c	4
4	c	c	s	c	8

Field	Year 7
1	c
2	c
3	c
4	c

Fig. 3. Workflow for CSB-based crop type prediction for hypothetical corn (yellow, c) and soybean (green, s) regions derived from six-year CDL time series. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

If a predictive map is desired, the year nine predictions can be added as an attribute to the original table. The polygon IDs and model predictions can then be used to color code the original polygons and produce the map. Predicted planted areas can also be obtained by summing the field predictions contained within the area of interest. For example, to predict next year's corn planted area in Illinois, the area of the fields in Illinois that are predicted to be corn would be summed.

2.4.5. Summary

To summarize, the procedure is to first use the time series of CDLs to derive CSBs, translate the CSB polygons and their attributes to a tabular dataset, split the table into training and validation time series, train a model using the training and validation data sets, and use the model to make pre-season predictions on what crop type will be planted in each CSB. The entire procedure is summarized for a small area CDL time series (Fig. 3). Also note the potential for data compression in Fig. 3, which is one of the benefits of the field-based model. There are 16 CDL pixels per year in this example, but only four fields. In other words, to achieve an equivalent data size between the field-based and pixel-based models, a 25% sample would need to be taken for the latter.

2.5. Competing models

To demonstrate the effectiveness of this new CSB-based crop type prediction algorithm, a comparison was made with two existing modeling approaches from the literature. The first is an artificial neural network (ANN) and the second is a random forest. Both approaches train their models using a subsample of the CDL pixels.

The ANN approach of Zhang et al. (2019a) involves several tuning parameters. For the comparison, the same tuning parameters mentioned in Zhang et al. (2019a) were used. Parameters that were left ambiguous were selected using a validation set. A separate ANN was fit to each ASD using the same sampling scheme (Zhang et al., 2019a). Finally, the same training, validation, and testing setup as defined in (Zhang et al., 2019a) was used.

The Random Forest approach considered here is by Johnson and Mueller (2021). Cross validation was not used in the original paper, so it was not performed in the experiments in Section 3. The same default parameters and training procedure as mentioned in the original paper were used. A separate Random Forest model was fit to each county using the same sampling scheme (Johnson and Mueller, 2021).

2.6. Model evaluation

To quantify the performance of the models, model predictions are compared to the rasterized FSA ground reference data (Section 2.3) at the 900 square meter pixel level. The following metrics are used to evaluate model predictions against the FSA ground reference data:

$$OA = \frac{\sum_{FSAPixels} I(modelprediction = FSAvalue)}{\sum_{FSAPixels}\{I(modelprediction = FSAvalue) + I(modelprediction \neq FSAvalue)\}}$$

$$F_{1t} = \frac{2 \times P_t \times R_t}{P_t + R_t}$$

$$P_t = \frac{\sum_{FSAPixels} I(model = tandFSA = t)}{\sum_{FSAPixels} I(model = t)}$$

Table 3
Dataset Sizes for CDL and CSBs by State.

State	CDL Pixels	CSB Polygons
Illinois	162,140,000	680,000
Indiana	104,130,000	440,000
Iowa	161,940,000	800,000
Louisiana	134,460,000	270,000
Minnesota	242,820,000	1,010,000
Missouri	200,590,000	650,000
Nebraska	222,600,000	790,000
Ohio	118,730,000	510,000

$$R_t = \frac{\sum_{FSAPixels} I(model = tandFSA = t)}{\sum_{FSAPixels} I(FSA = t)}$$

The function I is an indicator function, equal to 1 if the inside condition is met and 0 otherwise. The overall accuracy, OA, is the number of times the model prediction matches the FSA ground reference data, normalized by the FSA planted area. It attests to the global accuracy of the model. The F_1 score for each crop type t is the harmonic mean of the precision P_t and recall R_t. It ranges from zero (worst) to one (best). It can be used to assess the model accuracy for each individual crop type.

3. Results

In this section, the scalability of the field-based approach was evaluated by testing its data compression ability and predictive accuracy in each state mentioned in Table 1 in Section 2.1. As mentioned in Section 2.4.4, use of the CSB polygons as the unit of analysis theoretically allows for substantial data compression when compared to the raw CDL pixel dataset. Models using the CSB polygons in place of the CDL pixels benefit from a more than 200-fold data size reduction for every state (Table 3). Louisiana saw the biggest reduction by nearly 500 times.

Aside from the data-compression ability, the CSB-based approach also provides predictive accuracy. This was demonstrated by comparing the predictive accuracy of the CSB-based approach with pixel sampling-based alternatives from the existing literature (Tables 4, 5, 6). In Tables 4, 5 and 6, overall refers to the overall accuracy and each crop type refers to the F1 score for that crop.

The CSB-based field model was generally more accurate than the pixel-based models with respect to the FSA ground reference data. This accuracy advantage was consistent, as in the 24 scenarios outlined in Tables 4 through 6, the CSB-based model has the highest overall accuracy in 23 tests. The CSB-based field model was also the most accurate at identifying most individual crops. The CSB-based model has the highest F1 for corn (23/24), soybeans (20/21), sugarcane (3/3), sugar beets (3/3), cotton (2/3), and other (20/24) (Tables 4, 5, 6). For the underperforming cases, the field-level model ties with the pixel random forest in rice (3/6) and ties with both pixel-level models for spring wheat (1/3) (Tables 4, 5, 6).

All models tend to perform best in Illinois, Indiana, and Iowa with overall accuracies and major crop F1 scores greater than 80% (Table 4). In these states the CSB-based field model tends to have the highest score, except for two cases of lower "other" F1 in Indiana. The next highest accuracies are achieved in Nebraska and Ohio, with overall accuracies and major crop F1 scores ranging from 72 to 84% (Table 6). Again, the CSB-based field model tends to have the highest score, except for one case of lower "other" F1 in Ohio. Note that all these states have corn and

Table 4
Performance of Models: Illinois, Indiana, and Iowa. Most Accurate Predictions are Bolded.

	2016			2017			2018		
	FIELD LGB	CDL ANN	CDL RF	FIELD LGB	CDL ANN	CDL RF	FIELD LGB	CDL ANN	CDL RF
Illinois									
Overall	**83.5%**	82.2%	81.8%	**83.9%**	83.2%	82.0%	**84.9%**	84.1%	83.1%
Corn	**86.1%**	85.1%	85.1%	**85.9%**	85.4%	84.6%	**86.7%**	86.1%	85.4%
Soy	**83.2%**	81.8%	80.4%	**83.9%**	83.1%	81.4%	**85.3%**	84.4%	83.1%
Other	**52.7%**	49.4%	50.6%	**58.2%**	57.2%	56.1%	**51.3%**	49.5%	49.2%
Indiana									
Overall	**81.6%**	80.7%	79.4%	**82.0%**	81.0%	79.0%	**81.4%**	81.3%	80.3%
Corn	**83.1%**	82.5%	81.6%	**82.9%**	82.1%	80.6%	**82.6%**	82.5%	81.5%
Soy	**83.4%**	81.9%	80.3%	**83.6%**	82.4%	79.8%	**83.7%**	82.6%	81.5%
Other	55.4%	**56.7%**	55.3%	**58.8%**	58.0%	56.4%	53.2%	**57.5%**	57.0%
Iowa									
Overall	**85.9%**	85.3%	85.2%	**86.0%**	85.5%	84.6%	**87.4%**	86.8%	85.9%
Corn	**88.5%**	88.4%	88.3%	**88.4%**	87.8%	87.4%	**89.5%**	89.1%	88.5%
Soy	**83.5%**	82.1%	81.9%	**83.7%**	83.4%	81.6%	**85.5%**	84.9%	83.3%
Other	**71.0%**	68.5%	68.3%	**72.2%**	69.5%	69.6%	**70.3%**	67.5%	68.9%

Table 5
Performance of models: Louisiana and Minnesota. Most accurate predictions are bolded.

	2016			2017			2018		
	FIELD LGB	CDL ANN	CDL RF	FIELD LGB	CDL ANN	CDL RF	FIELD LGB	CDL ANN	CDL RF
Louisiana									
Overall	**79.1%**	74.5%	77.3%	**82.3%**	78.9%	80.8%	**80.8%**	76.5%	78.1%
Corn	41.0%	16.6%	**42.9%**	**41.8%**	26.7%	30.4%	**53.5%**	46.4%	41.1%
Sugarcane	**87.0%**	85.9%	84.8%	**86.8%**	86.5%	84.7%	**86.2%**	85.6%	83.7%
Rice	**79.7%**	71.1%	74.9%	**79.1%**	68.4%	75.6%	**78.2%**	64.7%	73.3%
Other	**84.4%**	81.7%	83.3%	**86.6%**	84.5%	86.2%	**85.6%**	82.9%	84.6%
Minnesota									
Overall	**78.6%**	76.9%	76.6%	**78.8%**	77.6%	76.4%	**79.6%**	78.8%	78.1%
Corn	**83.2%**	82.1%	82.0%	**82.9%**	82.2%	81.4%	**84.0%**	83.4%	82.6%
Soy	**80.2%**	79.0%	77.7%	**80.9%**	80.2%	78.1%	**81.4%**	80.8%	79.7%
Spring Wheat	66.4%	**79.0%**	64.7%	**62.6%**	61.1%	59.7%	64.1%	67.0%	**67.1%**
Sugar Beets	**60.7%**	54.2%	51.4%	**50.6%**	45.3%	39.7%	**61.8%**	55.3%	53.7%
Other	**61.2%**	55.0%	58.7%	**61.9%**	57.7%	59.6%	**62.8%**	58.9%	61.0%

Table 6
Performance of Models: Missouri, Nebraska, and Ohio. Most Accurate Predictions are Bolded.

	2016			2017			2018		
	FIELD LGB	CDL ANN	CDL RF	FIELD LGB	CDL ANN	CDL RF	FIELD LGB	CDL ANN	CDL RF
Missouri									
Overall	69.0%	32.7%	**72.1%**	**77.3%**	40.2%	74.5%	**78.9%**	41.4%	76.5%
Corn	**72.5%**	60.4%	68.3%	**75.2%**	75.0%	71.4%	**76.9%**	76.2%	73.4%
Cotton	54.0%	10.8%	**67.0%**	**70.6%**	9.9%	66.9%	**76.1%**	12.6%	69.9%
Rice	60.0%	10.4%	**61.3%**	**61.1%**	15.0%	**61.1%**	67.6%	20.2%	**71.6%**
Soy	71.5%	32.4%	**76.1%**	**80.0%**	34.6%	78.1%	**81.0%**	33.7%	79.3%
Other	62.6%	30.7%	**70.5%**	**76.8%**	29.8%	72.5%	**78.2%**	28.3%	75.9%
Nebraska									
Overall	**80.9%**	80.4%	79.2%	**80.6%**	80.1%	77.6%	**81.7%**	81.1%	78.9%
Corn	**83.8%**	83.6%	82.9%	**83.2%**	82.6%	81.1%	**84.1%**	83.8%	82.4%
Soy	**76.5%**	75.5%	72.3%	**77.0%**	76.8%	71.4%	**79.5%**	78.3%	73.8%
Other	**78.6%**	77.1%	76.7%	**77.9%**	77.3%	75.5%	**77.1%**	76.7%	75.3%
Ohio									
Overall	**78.6%**	77.8%	76.0%	**79.3%**	78.6%	76.8%	**78.7%**	78.3%	76.5%
Corn	**77.1%**	75.5%	73.9%	**76.4%**	75.7%	73.6%	**75.7%**	75.6%	73.0%
Soy	**83.3%**	82.7%	81.0%	**84.1%**	83.4%	81.8%	**83.4%**	82.9%	81.5%
Other	**60.8%**	60.7%	57.8%	**62.0%**	61.7%	59.1%	59.6%	**61.0%**	59.5%

soybeans as their major crops.

All models perform less well in states with more than two major crops. In Louisiana and Minnesota, overall accuracies were acceptable, ranging from the mid-70 s to low-80 s (Table 5). However, some crop specific F1 scores can drop much lower, especially for the ANN model. Louisiana corn was the most notable example where, for the most part, all models simultaneously achieve F1 scores under 50%. For all the previously mentioned states, with few exceptions, the CSB-based field model tends to have the highest score. Regardless, the CSB-based field model was still the most accurate, except for one case of Louisiana corn and two cases of Minnesota wheat.

The only outlier state, where the CSB-based model underperformed was Missouri, particularly in 2016 (Table 6). This was the state and year with the lowest overall accuracy (69%), much lower than all the other examples. During 2016 the random forest tended to be the top performer. The CSB-based model did recover for Missouri in 2017 and

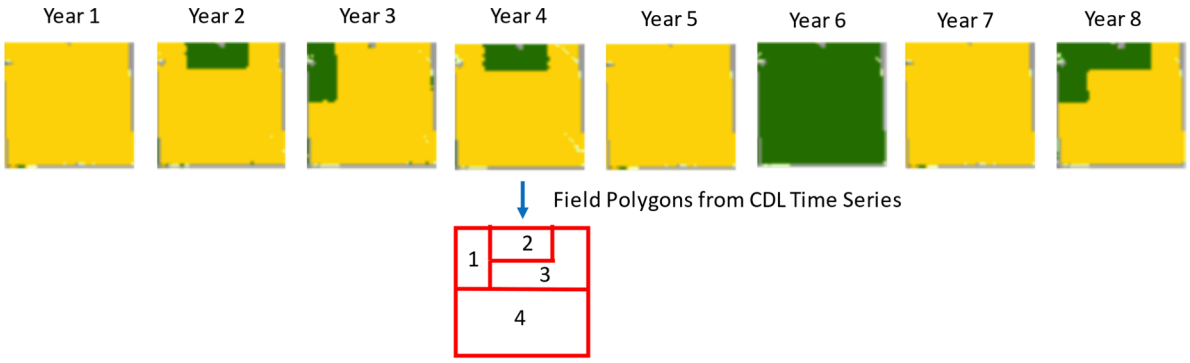
2018 however, where it generally regained top performance (excluding rice). The ANN tended to be the worst performing in all years for this state, with some F1 scores ranging as low as 10%.

4. Discussion

In the experiment in Section 3 it was demonstrated that field-based models were generally more accurate than their pixel-based counterparts with respect to accuracy and F1 score. In this section other qualitative benefits of modeling at the field level are discussed.

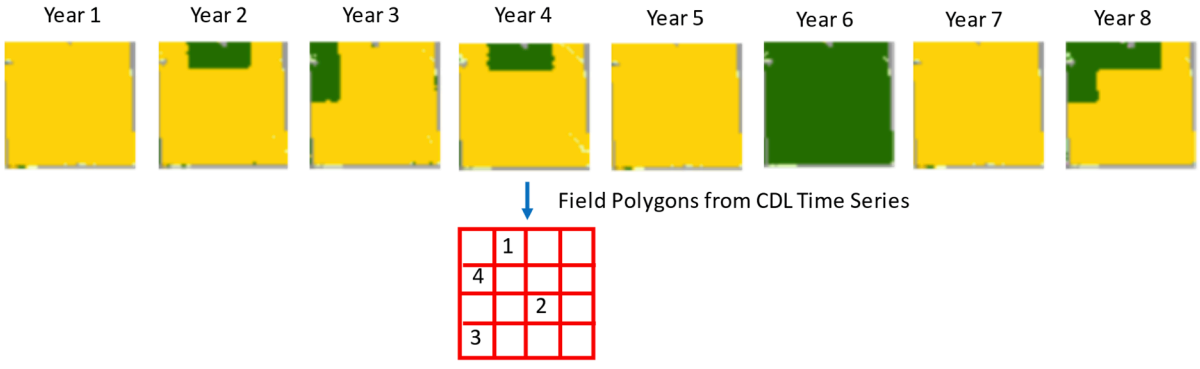
4.1. Field polygon vs pixel-level modeling

The primary difference in CSB and pixel-based models occurs at the first stage in the modeling process. Instead of deriving fields from the CDL, a pixel-based model samples CDL pixels from the full CDL time series. This sample is typically small, 10% and 0.25% for the pixel-based models discussed in Section 3. In other words, the CSB dataset is a deterministic set of field polygons while the sampled CDL dataset is a random subset of pixels. An example of the difference between the two approaches when applied to a small land area is provided in Fig. 4.



Field	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7	Year 8	Area
1	c	c	s	c	c	s	c	s	2
2	c	s	c	s	c	s	c	s	2
3	c	c	c	c	c	s	c	c	4
4	c	c	c	c	c	s	c	c	8

CSB



Sample	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7	Year 8	Area
1	c	s	c	s	c	s	c	s	1
2	c	c	c	c	c	s	c	c	1
3	c	c	c	c	c	s	c	c	1
4	c	c	s	c	c	s	c	s	1

Pixel

Fig. 4. Creating CSBs (top, red) from CDL corn (yellow) and soybean (green) pixels vs sampling CDL pixels (bottom, red). Note that CSBs 1,2,3,4 (top, red) are deterministic while sampled pixels 1,2,3,4 (top, red) are an outcome of a random draw. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

One question for pixel sampling is how many samples to take. The ANN model of (Zhang et al., 2019a) uses 10% of the data, while the random forest approach of (Johnson and Mueller, 2021) only uses 0.25%. The small sample used to train the random forest likely explains why it tends to underperform the field-level models and pixel-level ANN in most states in Section 3.2. Alternatively, the large sample of the ANN means it takes much longer to train than the random forest and may be harder to scale to many states, years, and crops. These factors suggest an accuracy vs computation time tradeoff that any pixel-based model would encounter.

In contrast to pixel-level modeling, the CSB-based field polygon-level model does not require sampling at all. This is true because a single field tends to contain many pixels, so summarizing the data at the CSB level may lead to a substantial dataset size reduction. This is indeed the case, as shown in Table 3. While the raw CDLs tend to have on the order of hundreds of millions of pixels per state, the summarized field-level datasets tend to have hundreds of thousands of fields. This order of magnitude reduction means the entire CSB dataset can be used in modeling, as it is not difficult to find hardware and algorithms that can handle datasets of this size. Therefore, the accuracy vs computation cost dilemma that exists with pixel-based modeling is absent with the CSBs.

4.2. Other benefits of CSB-Level modeling

The CSB-level dataset ensures that all the data are used in the analysis in proper proportion. This will not always be true in the sampling case, as the samples may not respect the underlying correlation structure resulting from within field homogenous planting practices. For example, note that for the given random sample in Fig. 4, field 4 is double counted while field 3 is missed. For larger area pixel-level models and small sample size, the information loss due to sampling is likely to be the more impactful on model performance because the probability of sampling more than one pixel from the same field is likely small. On the other hand, for small area models or large samples where the sample size may be closer to the number of pixels, double counting may become more of an issue. Regardless, these problems only manifest for the pixel-level models so field-level modeling is preferable.

Another drawback of pixel-level modeling is potential higher susceptibility to CDL noise. The CDL is not error free (Boryan et al., 2011; USDA-NASS, 2022) and thus misclassified CDL pixels may add unneeded noise into a predictive model. For example, a field that has a six-year corn only rotation but has a few erroneous soybean pixels inside may cause any purely pixel sampling derived models to make an unneeded mistake on the prediction. This is not a problem for the field-based approach as small islands in a larger field will be assigned to the majority crop type.

4.3. Limitations

The goal of this study was to demonstrate the effectiveness of field polygon modeling in place of pixel-based modeling. In general, while the CSB models were more accurate, there was one major outlier. This outlier, as mentioned in Section 3 and Table 6, occurred in Missouri in 2016. During this period, the CSB-based field model tended to underperform when compared to the random forest pixel-based model. As machine learning models tend to be black box algorithms, the reason for this one major case of underperformance cannot be known with certainty.

One possible cause for the Missouri 2016 outlier may be anomalous planting season weather that occurred during the prior year (2015). Unusually heavy precipitation in 2015 disrupted soybean, sorghum, and cotton planting, keeping progress well below historical averages (USDA-NASS, 2015a, 2015b, 2015c). This disrupted planting may have caused some areas to break their rotations in 2015, causing a rotation reset outlier which was not accounted for by the CSB or ANN models. This may also explain why the random forest achieves the highest performance, as it only uses four-year rotations. Being trained on smaller rotation histories may allow the random forest model to account for rotation resets caused by outliers in the previous planting year. Indeed, the ANN, which uses the longest histories, performs the worst, which also adds credence to the idea of poor performance of long rotation histories in the presence of recent rotational outliers.

Because the creation of CSBs relies on the CDL as input, the accuracy of the CDLs can impact any CSB based prediction model. This may increase prediction error if this model is extended to states and crops beyond this study. In other words, the predictive accuracies of the CSB based model will likely be lower than what is seen from Tables 4-6 when CDL accuracy is low. This is also a limitation of the pixel-based approaches, as they use the CDL pixels directly as inputs. Therefore, more work is needed to determine all possible states and crop types where meaningful pre-season predictions can be generated from both CSB and pixel-based methods.

Because the experiments in this study did not focus on optimizing the machine learning algorithm, there is no guarantee that LightGBM is the best model for pre-season crop type forecasting. While the accuracies in Tables 4-6 were good enough to be competitive with the pixel-based models, it is possible that they could be improved further by use of an alternative machine learning model. This limitation applies to the pixel-based approaches as well, as a single machine learning model (e.g., multi layer perceptron, random forest, deep learning) is typically used with model selection not considered.

Finally, a critical assumption for this study is that farmers do not break their rotation pattern too often. Because the machine learning model uses historical the crop rotation pattern to predict the future crop type, any break in the pattern may result in an inaccurate prediction. This may be resolvable by the addition of data available early in the year that may help predict rotation breaking. Some potential examples include commodity prices and futures, ethanol plant locations and capacities, seed varieties and prices, and input costs. The acquisition of such data, not all of which may be publicly available, still presents a challenge. Furthermore, pre-season prediction cannot address later season developments like inclement weather that may affect planting and break the rotation.

4.4. Future work

In this study, the primary concern was pixel-based modeling vs field polygon-based modeling. In the future, other factors affecting model performance could be examined. One factor is the length of the rotation history used in each model. A six-year history was used for the CSB-based model, the pixel-based ANN approach (Zhang et al., 2019a) used an eight-year history, and the pixel-based random forest approach (Johnson and Mueller, 2021) used a four-year history. It is possible that the difference in rotation history could impact the result. Another factor that could affect model performance is the use of recursive training sets. To form the training data, a single historical time series (last year) was used for the CSB-based model, (Zhang et al., 2019a) used three subsets for the ANN, and (Johnson and Mueller, 2021) used five for the random forest. All of these have different assumptions for the underlying model and may or may not be the best choice. Choosing the optimal number of time series to pool together is another direction for future work. Finally, the choice of machine learning model (LightGBM, ANN, Random Forest, other) may also impact the results. No one model is superior for every application, so the best choice for pre-season modeling is still an open question. The best model may also vary by region and crop type. In practice, one would tune all three of these hyperparameters for each choice (pixel and field) using the validation set. This is very computationally intensive and is left for future work.

Another avenue for future work is to explore the impact of how the fields are created. The CSB algorithm, modified from Beeson et al. (2020), has several tuning parameters. Varying these may provide better (or worse) quality fields. Finding the best parameters for each state,

county, etc. is a study in and of itself and can be looked at in future work. Furthermore, other field creation algorithms may also exist and may improve prediction quality. The review and comparison of different field creation algorithms for the crop type prediction application can also be a subject of future research.

Regarding scope expansion, performance could be measured for alternative states, years, and crop types not considered here. Because the study was designed to use scenarios where the CDL was most accurate, expanding to different states and crops inherently means encountering more CDL classification error. This could be a difficult issue as all models may underperform for crops and states where errors in the CDL are more prevalent. On the other hand, the smoothing effect of CSB polygons may allow the CSB-based models to be more robust to CDL classification error than the pixel-based alternatives. Examination of model performance in this general setting with higher CDL noise can be a subject of future investigation.

Finally, while this research was conducted in the U.S. using the annual CDL as the primary input, the proposed method can be used by researchers and practitioners in other countries which produce annual crop classifications based on remote sensing technology. Researchers in Canada, can potentially utilize the Annual Crop Inventory (Fisette et al., 2014; Fisette et al., 2013), produced since 2009, to evaluate the proposed methodology to generate synthetic field boundaries and the resulting crop predictions for similar applications in Canada. Researchers in England, can potentially utilize the Crop Map of England (CROME) (Rural Payments Agency, 2020), produced since 2016, to produce crop type predictions and synthetic field boundaries using the proposed method based on a shorter time frame. The research and eventual production of crop type classifications, based on remote sensing, are expanding in other countries including developing countries (FAO, 2022; Servicio de Información Agroalimentaria y Pesquera, 2020). Over time researchers, studying agricultural production around the world will be able to evaluate the proposed method for preseason crop type prediction using synthetic field polygons for their specific applications.

5. Conclusions

Reliable pre-season crop-type prediction has many applications, including crop mapping, planted acreage prediction, crop yield prediction, disaster response, area sample design, crop survey imputation, and more. In NASS, pre-season crop type predictions are used operationally for pre-season crop acreage estimation and large-scale operational crop survey imputation. These applications will likely be expanded in the future.

This study introduced a novel machine-learning based pre-season crop prediction approach using field-level polygons called CSBs as inputs. This adds to previous literature by being the first CDL derived parcel-based machine learning model for pre-season forecasting. This contrasts with other existing CDL based machine learning pre-season forecasting approaches, which are pixel based. Parcels are useful because they may be closer to the actual decision-making units, as opposed to pixels, which are arbitrarily sized and shaped.

Parcel level machine learning predictions using CSBs also offer an efficient way to use the full dataset without relying on pixel sampling, which is the standard in the current literature. Because there are far fewer CSBs than there are CDL pixels, one can use the entire CSB dataset for predictive modeling. This avoids the need to consider the sample-size accuracy tradeoff that may occur for pixel-based machine learning. As such the proposed CSB based machine learning approach offers an alternative method to processing large CDL datasets.

In addition to the novel parcel level predictions, the CSB based machine learning model also improves upon the performance of current pixel-based approaches. The CSB-level model results were compared to CDL pixel-based alternatives in nine states. It was shown that the CSB-level model was competitive and achieved the highest overall accuracy, with a median of 80.85% versus medians of 78.90% for the ANN model and 78.50% for the RF model with respect to agreement with FSA ground reference data (medians computed from Tables 4-6). The CSB model was also competitive on a crop-by-crop basis for all crops except spring wheat, achieving median F1 scores ranging from 0.5% to 6% higher than the best pixel-based alternative.

In the future, the limitations of this approach will continue to be addressed. Exploration will be done into expanding the approach to new states, years, and crops. This will allow investigation into the effect of higher CDL error on prediction performance as well as potentially gaining more understanding of negative performance outliers like Missouri in 2016. Furthermore, as more data is obtained over time, it may become possible to add new pre-season variables such as economic indicators to make the model more robust to rotation breakages.

CRediT authorship contribution statement

Jonathon Abernethy: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, Data curation. **Peter Beeson:** Methodology, Software, Writing – review & editing, Visualization. **Claire Boryan:** Resources, Data curation, Writing – review & editing, Supervision, Project administration. **Kevin Hunt:** Resources, Data curation, Writing – review & editing, Visualization, Supervision, Project administration. **Luca Sartore:** Software, Data curation, Writing – review & editing.

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.

Data availability

The data that has been used is confidential.

Acknowledgements

We would like to acknowledge Valbona Bejleri, Bruce Craig, Zhengwei Yang, and Linda Young for their work in reviewing this paper prior to submission. This research was supported by the intramural research program of the U.S. Department of Agriculture, National Agricultural Statistics Service (58-3AEU-9-0025) and Economic Research Service (58-6000-0-0055R).

References

Aurbacher, J., Dabbert, S., 2011. Generating crop sequences in land-use models using maximum entropy and Markov chains. Agric. Syst. 104, 470–479. https://doi.org/10.1016/J.AGSY.2011.03.004.

Ballestores, F., Qiu, Z., 2012. An integrated parcel-based land use change model using cellular automata and decision tree. Proc. Int. Acad. Ecol. Environ. Sci. 2, 53–69.

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. https://doi.org/10.1080/10106049.2011.562309 26, 341–358. https://doi.org/10.1080/10106049.2011.562309.

Boryan, C.G., Yang, Z., Sandborn, A., Willis, P., Haack, B., 2018. Operational agricultural flood monitoring with sentinel-1 synthetic aperture radar. Int. Geosci. Remote Sens. Symp. 2018-July, 5831–5834. https://doi.org/10.1109/IGARSS.2018.8519458.

Boryan, C.G., Yang, Z., 2017. Integration of the Cropland Data Layer Based Automatic Stratification Method into the Traditional Area Frame Construction Process. Surv. Res. Methods 11, 289–306. https://doi.org/10.18148/SRM/2017.V11I3.6725.

National Council, 2013. Advancing Land Change Modeling: Opportunities and Research Requirements, Advancing Land Change Modeling. National Academies Press. https://doi.org/10.17226/18385.

Servicio de Información Agroalimentaria y Pesquera, 2020. Mapa con la estimación de superficie sembrada de cultivos básicos [WWW Document]. URL https://www.gob.mx/siap/documentos/mapa-con-la-estimacion-de-superficie-sembrada-de-cultivos-basicos.

Dornbierer, J., Wika, S., Robison, C., Rouze, G., Sohl, T., 2021. Prototyping a methodology for long-term (1680–2100) historical-to-future landscape modeling for the conterminous United States. Land 10. https://doi.org/10.3390/LAND10050536.

FAO, 2022. Crop mapping using remote sensing in Malawi site, Egypt.
Fisette, T., Rollin, P., Aly, Z., Campbell, L., Daneshfar, B., Filyer, P., Smith, A., Davidson, A., Shang, J., Jarvis, I., 2013. AAFC annual crop inventory: Status and challenges. 2013 2nd Int. Conf. Agro-Geoinformatics Inf. Sustain. Agric. Agro-Geoinformatics 2013, 270–274. https://doi.org/10.1109/ARGO-GEOINFORMATICS.2013.6621920.
Fisette, T., Davidson, A., Daneshfar, B., Rollin, P., Aly, Z., Campbell, L., 2014. Annual space-based crop inventory for Canada: 2009–2014. Int. Geosci. Remote Sens. Symp. 5095–5098 https://doi.org/10.1109/IGARSS.2014.6947643.
Han, W., Yang, Z., Di, L., Mueller, R., 2012. CropScape: A Web service based application for exploring and disseminating US conterminous geospatial cropland data products for decision support. Comput. Electron. Agric. 84, 111–123. https://doi.org/10.1016/J.COMPAG.2012.03.005.
Heald, J., 2002. USDA Establishes a Common Land Unit.
Irwin, E.G., Geoghegan, J., 2001. Theory, data, methods: developing spatially explicit economic models of land use change. Agric. Ecosyst. Environ. 85, 7–24. https://doi.org/10.1016/S0167-8809(01)00200-6.
Johnson, D.M., 2014. An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the United States. Remote Sens. Environ. 141, 116–128. https://doi.org/10.1016/J.RSE.2013.10.027.
Johnson, D.M., Mueller, R., 2021. Pre- and within-season crop type classification trained with archival land cover information. Remote Sens. Environ. 264, 112576 https://doi.org/10.1016/J.RSE.2021.112576.
Lark, T.J., Mueller, R.M., Johnson, D.M., Gibbs, H.K., 2017. Measuring land-use and land-cover change using the U.S. department of agriculture's cropland data layer: Cautions and recommendations. Int. J. Appl. Earth Obs. Geoinf. 62, 224–235. https://doi.org/10.1016/J.JAG.2017.06.007.
Orynbaikyzy, A., Gessner, U., Conrad, C., 2019. Crop type classification using a combination of optical and radar remote sensing data: a review. https://doi.org/10.1080/01431161.2019.1569791 40, 6553–6595. https://doi.org/10.1080/01431161.2019.1569791.
Osman, J., Inglada, J., Dejoux, J.F., 2015. Assessment of a Markov logic model of crop rotations for early crop mapping. Comput. Electron. Agric. 113, 234–243. https://doi.org/10.1016/J.COMPAG.2015.02.015.
Rauf, U., Qureshi, W.S., Jabbar, H., Zeb, A., Mirza, A., Alanazi, E., Khan, U.S., Rashid, N., 2022. A new method for pixel classification for rice variety identification using spectral and time series data from Sentinel-2 satellite imagery. Comput. Electron. Agric. 193, 106731 https://doi.org/10.1016/J.COMPAG.2022.106731.
Rural Payments Agency, 2020. Crop Map of England (CROME) 2020 - data.gov.uk [WWW Document]. URL https://www.data.gov.uk/dataset/be5d88c9-acfb-4052-bf6b-ee9a416cfe60/crop-map-of-england-crome-2020 (accessed 2.28.23).
Sohl, T., Dornbierer, J., Wika, S., Sayler, K., Quenzer, R., 2017. Parcels versus pixels: modeling agricultural land use across broad geographic regions using parcel-based field boundaries. https://doi.org/10.1080/1747423X.2017.1340525 12, 197–217. https://doi.org/10.1080/1747423X.2017.1340525.
Sohl, T., Dornbierer, J., Wika, S., Robison, C., 2019. Remote sensing as the foundation for high-resolution United States landscape projections – The Land Change Monitoring, assessment, and projection (LCMAP) initiative. Environ. Model. Softw. 120, 104495. https://doi.org/10.1016/J.ENVSOFT.2019.104495.
Shwartz-Ziv, R., Armon, A., 2022. Tabular data: Deep learning is not all you need. Inf. Fusion 81, 84–90. https://doi.org/10.1016/J.INFFUS.2021.11.011.
USDA-FSA, 2017. Common Land Unit Information Sheet.
USDA-NASS, 2015a. June Acreage.
USDA-NASS, 2015b. July Production.
USDA-NASS, 2015c. August Production.
USDA-NASS, 2022. CropScape and Cropland Data Layers - FAQs.
Wang, X., Berg, E., Zhu, Z., Sun, D., Demuth, G., 2018. Small Area Estimation of Proportions with Constraint for National Resources Inventory Survey. J. Agric. Biol. Environ. Stat. 2018 234 23, 509–528. https://doi.org/10.1007/S13253-018-0329-6.
White, E.V., Roy, D.P., 2015. A contemporary decennial examination of changing agricultural field sizes using Landsat time series data. Geo Geogr. Environ. 2, 33–54. https://doi.org/10.1002/GEO2.4.
Xiao, Y., Mignolet, C., Mari, J.F., Benoît, M., 2014. Modeling the spatial distribution of crop sequences at a large regional scale using land-cover survey data: A case from France. Comput. Electron. Agric. 102, 51–63. https://doi.org/10.1016/J.COMPAG.2014.01.010.
Yan, L., Roy, D.P., 2014. Automated crop field extraction from multi-temporal Web Enabled Landsat Data. Remote Sens. Environ. 144, 42–64. https://doi.org/10.1016/J.RSE.2014.01.006.
Yang, L., Jin, S., Danielson, P., Homer, C., Gass, L., Bender, S.M., Case, A., Costello, C., Dewitz, J., Fry, J., Funk, M., Granneman, B., Liknes, G.C., Rigge, M., Xian, G., 2018. A new generation of the United States National Land Cover Database: Requirements, research priorities, design, and implementation strategies. ISPRS J. Photogramm. Remote Sens. 146, 108–123. https://doi.org/10.1016/J.ISPRSJPRS.2018.09.006.
Yao, A., Di, L., 2021. Machine Learning-based Pre-season Crop Type Mapping: A Comparative Study. 2021 9th Int. Conf. Agro-Geoinformatics, Agro-Geoinformatics 2021. https://doi.org/10.1109/AGRO-GEOINFORMATICS50104.2021.9530356.
Yaramasu, R., Bandaru, V., Pnvr, K., 2020. Pre-season crop type mapping using deep neural networks. Comput. Electron. Agric. 176, 105664 https://doi.org/10.1016/J.COMPAG.2020.105664.
You, N., Dong, J., 2020. Examining earliest identifiable timing of crops using all available Sentinel 1/2 imagery and Google Earth Engine. ISPRS J. Photogramm. Remote Sens. 161, 109–123. https://doi.org/10.1016/J.ISPRSJPRS.2020.01.001.
Zhang, C., Di, L., Hao, P., Yang, Z., Lin, L., Zhao, H., Guo, L., 2021a. Rapid in-season mapping of corn and soybeans using machine-learned trusted pixels from Cropland Data Layer. Int. J. Appl. Earth Obs. Geoinf. 102, 102374 https://doi.org/10.1016/J.JAG.2021.102374.
Zhang, C., Di, L., Yang, Z., Lin, L., Yu, E.G., Yu, Z., Rahman, M.S., Zhao, H., 2019b. Cloud environment for disseminating NASS cropland data layer. 2019 8th Int. Conf. Agro-Geoinformatics, Agro-Geoinformatics 2019. https://doi.org/10.1109/AGRO-GEOINFORMATICS.2019.8820465.
Zhang, C., Di, L., Lin, L., Guo, L., 2019a. Machine-learned prediction of annual crop planting in the U.S. Corn Belt based on historical crop planting maps. Comput. Electron. Agric. 166, 104989. https://doi.org/10.1016/J.COMPAG.2019.104989.
Zhang, C., Yang, Z., Di, L., Lin, L., Hao, P., Guo, L., 2021b. Applying machine learning to cropland data layer for agro-geoinformation discovery. Int. Geosci. Remote Sens. Symp. 1149–1152 https://doi.org/10.1109/IGARSS47720.2021.9554628.