Water Resources Research

RESEARCH ARTICLE
10.1029/2019WR024892

Special Section:
Big Data & Machine Learning in Water Sciences: Recent Progress and Their Use in Advancing Science

Identifying Subsurface Drainage using Satellite Big Data and Machine Learning via Google Earth Engine

Eunsang Cho[1,2], Jennifer M. Jacobs[1,2], Xinhua Jia[3], and Simon Kraatz[1,4]

[1]Department of Civil and Environmental Engineering, University of New Hampshire, Durham, NH, USA, [2]Earth Systems Research Center, Institute for the Study of Earth, Oceans, and Space, University of New Hampshire, Durham, NH, USA, [3]Department of Agricultural and Biosystems Engineering, North Dakota State University, Fargo, ND, USA, [4]Now at Department of Electrical and Computer Engineering, University of Massachusetts, Amherst, MA, USA

Abstract Human-induced landscape changes affect hydrologic responses (e.g., floods) that can be detected from a suite of satellite and model data sets. Tapping these vast data sets using machine learning algorithms can produce critically important and accurate insights. In the Red River of the North Basin in the United States, agricultural subsurface drainage (SD; so-called tile drainage) systems have greatly increased since the late 1990s. Over this period, river flow in the Red River has markedly increased and 6 of 13 major floods during the past century have occurred in the past two decades. The impact of SD systems on river flow is elusive because there are surprisingly few SD records in the United States. In this study, Random Forest machine learning (RFML) classification method running on Google Earth Engine's cloud computing platform was able to capture SD within a field (30 m) and its expansion over time for a large watershed (>100,000 km^2). The resulting RFML classifier drew from operational multiple satellites and model data sets (total 14 variables with 36 layers including vegetation, land cover, soil properties, and climate variables). The classifier identified soil properties and land surface temperature to be the strongest predictors of SD. The maps agreed well with SD permit records (overall accuracies of 76.9–87.0%) and corresponded with subwatershed-level statistics (r = 0.77–0.96). It is expected that the maps produced with this data-intensive machine learning approach will help water resource managers to assess the hydrological impact from SD expansion and improve flood predictions in SD-dominated regions.

Plain Language Summary Farmers install subsurface drainage pipes (so-called tile drainage) to improve crop yields on poorly drained soils, which impacts hydrological response (e.g., floods). Consistent records of subsurface drainage expansion are needed to understand its impacts on water resources. In the Red River of the North Basin in the United States, subsurface drainage systems have increased since the late 1990s. Over this period, river flow in the Red River has markedly increased and 6 of 13 major floods during the past century have occurred in the past two decades. Because the current National Oceanic and Atmospheric Administration's National Weather Service flood forecasting model does not include subsurface drainage information, they sometimes overpredict or underpredict flood flows. We developed high-resolution (30 m) subsurface drainage maps by combining multiple satellite "big" data and model products using a Random Forest machine learning classification via Google Earth Engine's cloud computing platform. The maps showed good agreement with available subsurface permit records. It is expected that the machine learning-based subsurface drainage maps will help water resource managers and flood forecasters to improve flood prediction in agricultural dominated regions.

Key Points:
- High-resolution subsurface drainage maps were developed using satellite big data and random forest machine learning via Google Earth Engine
- Reliable subsurface drainage records are needed for sustainable water resource management, but such records are very limited in the United States
- While soil variables are important to identify potential drainage areas, land surface temperature distinguishes where drainage has occurred

Supporting Information:
- Supporting Information S1

Correspondence to:
E. Cho,
ec1072@wildcats.unh.edu

Citation:
Cho, E., Jacobs, J. M., Jia, X., & Kraatz, S. (2019). Identifying subsurface drainage using satellite big data and machine learning via Google Earth Engine. *Water Resources Research*, *55*. https://doi.org/10.1029/2019WR024892

Received 31 JAN 2019
Accepted 24 AUG 2019
Accepted article online 30 AUG 2019

1. Introduction

In the northcentral United States, the amount of streamflow has greatly increased and floods have occurred more frequently during the last 20 years. In the Red River of the North Basin (RRB), 6 of the 13 major floods over the past 100 years have occurred since the late 1990s (Rannie, 2015; Todhunter, 2001; Tuttle et al., 2017). Numerous studies have been conducted to determine the major causes for the hydrologic changes in the northcentral United States (Foufoula-Georgiou et al., 2015; Frans et al., 2013; Raymond et al., 2008; Schilling et al., 2010). Potential factors include changes in climate, land use and land cover, including agricultural subsurface drainage installation. Subsurface drainage (SD) expansion in agricultural landscapes resulting in an increase in cultivated areas is a key cause of regional water balance change (Rogger et al., 2017; Schottler et al., 2014). In the past two decades, SD systems have exponentially expanded over the

agricultural areas in the northcentral United States. In the RRB, SD areas have dramatically increased from 2000 to the present (e.g., in North Dakota, 1.26, 114, and 892 km^2 for 2002, 2008, and 2016, respectively; Finocchiaro, 2014, 2016; Dollinger et al., 2013).

SD systems are used to remove excess surface water and to lower water tables through subsurface pipe networks installed below the ground surface. When the drainage pipes are installed at a certain depth and spacing, the pressure head at the pipes is approximately the atmospheric pressure and the pressure distributions in soil profile horizons change to an equilibrium profile. Thus, the original water tables in the undrained condition are lowered to the equivalent depth of the drainage systems, especially during spring and fall. The effective infiltration rates would be changed by drainage installations due to the altered hydraulic gradient of the upper soil layer above drained pipes (Rodgers et al., 2003; Shokri & Bardsley, 2015; Youngs, 1975).

SD impacts on runoff, soil moisture dynamics, and evapotranspiration have been studied at a range of spatiotemporal scales (Eastman et al., 2010; Frans et al., 2013; Kelly et al., 2017; King et al., 2014; Kladivko et al., 2004; Lenhart et al., 2011; Rahman et al., 2014; Randall et al., 2003; Schottler et al., 2014; Williams et al., 2015). At a field scale, Kladivko et al. (2004) showed that SD-induced water yields were 8% to 26% of annual rainfall in southeastern Indiana, depending on the year and the drain spacing. Eastman et al. (2010) found that the subsurface-drained field discharged four times more water than the naturally drained field for their clay loam sites. At a watershed scale, King et al. (2014) reported that about 21% of annual precipitation and 47% of total watershed discharge was generated from SD in central Ohio. Williams et al. (2015) concluded that SD discharge contributed 56% of the annual watershed discharge in the same Ohio watershed. At a larger scale, Frans et al. (2013) showed that SD increased annual streamflow up to 40% locally in the Upper Mississippi River basin. Schottler et al. (2014) compared a change in water yield between two historical periods (1940–1974 and 1975–2009) in watershed scale. They found that SD expansion is likely the major driver of increased streamflow in 21 Minnesota agricultural watersheds. Kelly et al. (2017) also concluded that the extensive SD systems in agricultural basins have contributed to the increase in river flow at the large basin scale.

Despite the increased water yield, it is possible that SD could mitigate downstream flooding by allowing surface runoff to infiltrate and be released at a slower rate. As recently as 2013, the National Oceanic and Atmospheric Administration's (NOAA) National Weather Service North Central River Forecast Center (NCRFC) predicted a peak flow that exceeded the observed by 70% in the RRB (Tuttle et al., 2017). Because the current flood forecasting system does not consider SD information, it is still an open question as to the dominant processes that are affected by SD in the region. However, it has been observed that as SD systems have expanded, operational flood forecasting has become more difficult due to limited information about spatial and temporal SD expansion (personal communications with Mike DeWeese and Pedro Restrepo, NOAA NCRFC).

Due to the paucity of SD data, the results of the previous studies also had considerable uncertainties. Schottler et al. (2014) indicated that the unexplained portion of evapotranspiration change in the long-term water balance approach is due to SD change but did not have the supporting SD data. While Kelly et al. (2017) had county-level drainage data for five census years to assess SD impact on runoff patterns, they noted inconsistencies and errors of the census data with concerns about limited SD records in the United States.

Most previous studies that have attempted to map SD systems focused on delineating subsurface drained lines (or areas) at a field or catchment scales and used Geographical Information System (GIS)-based analyses and/or aerial image processing techniques (Naz et al., 2009; Naz & Bowling, 2008; Sugg, 2007; Tetzlaff et al., 2009; Tetzlaff et al., 2009; Zhang et al., 2014). The 1992 National Resource Inventory (NRI) data set provided potential extents of subsurface drains in continental United States (Wieczorek, 2004). The NRI data set was created with GIS and database management tools using collections at more than 800,000 sample sites over the United States. Sugg (2007) estimated SD percentage for each county based on the GIS-based soil drainage class. They compared their results with the NRI drainage map and developed a SD map at the county scale. Sui (2007) also used a GIS-based analysis of land cover, soil, and slope data sets to classify the SD areas for cropland in Indiana where the soils are poorly drained, and the slope is less than 2%. However, the SD mapping studies could not validate their results due to the lack of SD data (Naz et al., 2009; Sugg, 2007). Infrared aerial photographs have been used to map subsurface drain lines and to delineate wet and drained areas in a field (Verma et al., 1996). Soils over subsurface drained areas have higher

reflectance in the infrared spectrum because these areas tend to dry faster than the soil at other regions. Previous studies found that the best time to take imagery to be used for SD delineation is within 3 days after a 25 mm or greater rainfall event (Northcott et al., 2000; Varner et al., 2002). A combination of high resolution (1-m) color (or black and white) infrared aerial images with land cover, soil, and topography data provided a map of individual drainage lines in westcentral Indiana (Naz et al., 2009). Tlapáková et al. (2015) provided an example of manifestations of SD systems in color aerial images and suggested best land conditions for the optimal SD identification. Using an optical camera and unmanned aerial vehicle system, Zhang et al. (2014) developed a mosaiced SD map from infrared color composite imagery. While the aerial imagery approaches allow targeted study of watersheds, they are expensive and may be limited by weather and the availability of resources.

Satellite remote sensing data offers the ability to observe temporal changes in surface conditions due to SD at large spatial extents. Gökkaya et al. (2017) and Møller et al. (2018) provide evidence of SD induced surface changes using Landsat satellite images. However, they had few satellite observations due to limited cloud-free images and data processing requirements. Jacobs et al. (2017) showed that Moderate Resolution Image Spectroradiometer (MODIS) land surface temperature and Advanced Microwave Scanning Radiometer for Earth Observing System soil moisture products could detect physical effects of SD systems on soil thermal-moisture dynamics. In addition to these products, there are many other satellite products that might show the SD signature. However, traditional analysis techniques, such as image processing techniques and the GIS-based decision tree classification commonly used in previous studies (Gökkaya et al., 2017; Naz & Bowling, 2008; Sugg, 2007), are not well suited to manage and analyze terabyte-size satellite remote sensing data sets for SD detection. In these cases, machine learning (ML) techniques have demonstrated value (Belgiu & Drăguţ, 2016; McCabe et al., 2017; Møller et al., 2018; Shen, 2018; Tao et al., 2016).

Random forest machine learning (RFML) is a supervised classification algorithm that constructs a multitude of decision trees and predicts class labels, using a random subset of training samples and variables (Breiman, 2001). The RFML has become popular within the remote sensing and hydrology communities due to its accuracy (Belgiu & Drăguţ, 2016; Gómez et al., 2016; Petty & Dhingra, 2018). For land surface and crop type monitoring, the RFML has been shown to produce higher accuracies than other ML techniques such as Maximum Likelihood Classifier, Neural Network, and Support Vector Machine (Gómez et al., 2016; Ma et al., 2017; Ok et al., 2012). Also, it has been widely used in the field of hydrological predictions due to its capacity to determine variable importance, its robustness to data reduction, and that does not overfit (Petty & Dhingra, 2018; Shortridge et al., 2016; Wang et al., 2015). Compared to other techniques, however, the RFML method has inherent limitations including (1) complexity which makes less straightforward to understand the relationship in the input data and (2) significant and timing-consuming of computational requirements to construct the algorithm.

The Google Earth Engine's (GEE) cloud computing platform (Gorelick et al., 2017) provides the ability to manage very large satellite and model data sets to analyze them using ML techniques. The GEE is designed to provide access to high-performance computing resources for processing massive geospatial data sets, without technical hurdles (e.g., data download and storage, handling obscure file formats, and managing databases). Because a variety of geospatial data sets including historical and current satellite and aerial imaging systems can be freely accessed and analyzed, the GEE has been widely used in computationally expensive hydrological, agricultural and socio-economic studies (Deines et al., 2017; Ge et al., 2019; Jin et al., 2019; Xie et al., 2019).

Here, we focus on developing SD maps to improve the capability of flood forecasting in agricultural landscapes across the RRB. The RFML algorithm is used to develop annual SD maps in the GEE computing platform. We also seek to understand which of the related, globally available vegetation, thermal, moisture, and climate data sets from multiscale satellites and models can be used to identify SD areas and with what accuracy. The accuracy of these maps is assessed using SD permit records in the Bois de Sioux Watershed (BdSW) in Minnesota and the North Dakota portion of the RRB region (ND-RRB).

2. Study Area

The Red River of the North Basin overlies portions of North Dakota, South Dakota, and Minnesota (Figure 1). Its main stem marks the border between North Dakota and Minnesota. The river flows north

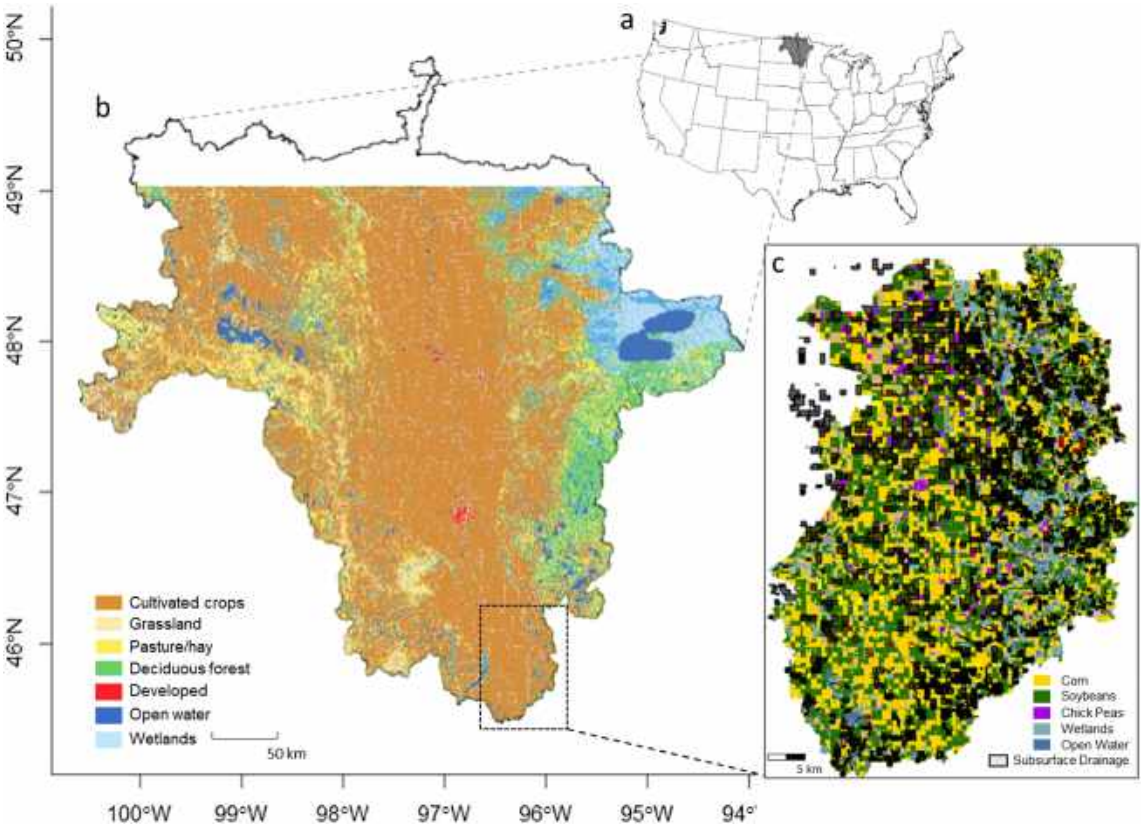

Figure 1. Study area location and land cover map. (a) Red River of the North Basin; (b) Land cover classification from U.S. Geological Survey National Land Cover Database 2011; and (c) Cropland data layers with subsurface drained area in 2017 noted in Bois de Sioux Watershed.

from Wahpeton, ND, to the U.S.-Canada border, and then through Winnipeg, Manitoba, Canada. The basin drainage area is about 112,200 km^2, with 885-km long from U.S. Geological Survey (USGS) Watershed Boundary Dataset (HUC04). Along the distance of the main stem, it drops only 72 m, for an average gradient of 0.08 m/km. In the RRB, agricultural SD systems have increasingly used to drain fields since the late 1990s due to the region's flat topography and low-permeability soils. The NOAA flood forecasters and water resource experts in RRB identified the rapid increase in the SD systems as a likely culprit for the changed river flows and floods because SD alters direct runoff, soil moisture, and evaporation seasonally (Rijal et al., 2012; Schottler et al., 2014).

3. Method

3.1. Data Sets

Working in the GEE cloud computing platform, the data sets from multisource satellite and model assimilation products were used (total 1.4 terabytes). Table 1 lists the 36 seasonal and annual layers that were generated including 16 vegetation layers (top 16 lines at the table), 8 soil-climate variable layers (next 8 lines), and 12 thermal-moisture layers (12 lines from the bottom) for 2009, 2011, 2014, and 2017. All 36 input layers were disaggregated to 30-m pixel resolution. The data sets generally fit into three categories: vegetation, thermal-moisture, and climate-land variables. The four years were selected based on land surface conditions (dry/wet) particularly in spring based on spring mean precipitation and soil moisture. Years 2009 and 2011 were selected because they have experienced spring snowmelt floods over the RRB. Even though years 2014 and 2017 were not as wet as 2009 and 2011, they were included to examine whether the RFML method can capture continuous SD expansion over the years that have occurred in RRB. The magnitude of year-over-year hydrologic variability for 2009 through 2017 is shown via a hydrograph at Fargo, North Dakota (USGS: 0505400), which a major streamflow gage in the RRB (Figure S1 in the supporting information).

Table 1
Summary of Variables Used in RFML Including Time Period, Resolution, and Data Source

Variable (full name)	Short name	Time period	Resolution (m^2)	Data source
Spring mean EVI	EVI_spr_mean	1 May to 30 Jun	30	Landsat
Spring mean GI	GI_spr_mean	1 May to 30 Jun	30	Landsat
Spring mean NDVI	NDVI_spr_mean	1 May to 30 Jun	30	Landsat
Spring mean NDWI	NDWI_spr_mean	1 May to 30 Jun	30	Landsat
Spring range in EVI	EVI_spr_range	1 May to 30 Jun	30	Landsat
Spring range in GI	GI_spr_range	1 May to 30 Jun	30	Landsat
Spring range in NDVI	NDVI_spr_range	1 May to 30 Jun	30	Landsat
Spring range in NDWI	NDWI_spr_range	1 May to 30 Jun	30	Landsat
Summer mean EVI	EVI_sum_mean	1 Jul to 30 Sep	30	Landsat
Summer mean GI	GI_sum_mean	1 Jul to 30 Sep	30	Landsat
Summer mean NDVI	NDVI_sum_mean	1 Jul to 30 Sep	30	Landsat
Summer mean NDWI	NDWI_sum_mean	1 Jul to 30 Sep	30	Landsat
Summer range in EVI	EVI_sum_range	1 Jul to 30 Sep	30	Landsat
Summer range in GI	GI_sum_range	1 Jul to 30 Sep	30	Landsat
Summer range in NDVI	NDVI_sum_range	1 Jul to 30 Sep	30	Landsat
Summer range in NDWI	NDWI_sum_range	1 Jul to 30 Sep	30	Landsat
Spring mean soil moisture	SM_spr_mean	1 May to 30 Jun	25,000	SMOS (NASA-USDA)
Spring range soil moisture	SM_spr_range	1 May to 30 Jun	25,000	SMOS (NASA-USDA)
Spring mean LST	LST_spr_mean	1 May to 30 Jun	1,000	Terra MODIS
Spring range LST	LST_spr_range	1 May to 30 Jun	1,000	Terra MODIS
Spring mean STR 1	STR1_spr_mean	1 May to 30 Jun	30	Landsat
Spring mean STR 2	STR2_spr_mean	1 May to 30 Jun	30	Landsat
Spring range STR 1	STR1_spr_range	1 May to 30 Jun	30	Landsat
Spring range STR 2	STR2_spr_range	1 May to 30 Jun	30	Landsat
Summer mean STR 1	STR1_sum_mean	1 Jul to 30 Sep	30	Landsat
Summer mean STR 2	STR2_sum_mean	1 Jul to 30 Sep	30	Landsat
Summer range STR 1	STR1_sum_range	1 Jul to 30 Sep	30	Landsat
Summer range STR 2	STR2_sum_range	1 Jul to 30 Sep	30	Landsat
Growing season precipitation	Preci_grow	1 May to 30 Sep	4,000	GRIDMET
Early season precipitation	Preci_early	1 Dec to 30 Apr	4,000	GRIDMET
Annual precipitation	Preci_ann	1 Dec to 30 Sep	4,000	GRIDMET
Aridity	Aridity	1 May to 30 Sep	4,000	GRIDMET
Cropland Data Layers	Cropland	NA	30	USDA NASS
Clay percentage	clay_perc	NA	30	POLARIS
Available soil water content	awc	NA	30	POLARIS
Saturated hydraulic conductivity	ksat	NA	30	POLARIS

Note. All input variables were accessed through the GEE's data archive, except for the three 30-m soil property data sets from POLARIS (available at www.polaris.earth; Chaney et al., 2016, 2019), which were manually uploaded to the GEE for RFML classification. 16 vegetation layers appear in the top 16 rows (EVI, GI, NDVI, and NDWI), 12 thermal-moisture layers follow the vegetation layers (SM, LST, STR1, and STR2), and 8 soil-climate variable layers are the remaining 8 rows in the table (preci, aridity, cropland, and three soil properties). GEE = Google Earth Engine; RFML = random forest machine learning; EVI = enhanced vegetation index; GI = greenness index; NDVI = normalized difference vegetation index; NDWI = normalized difference water index.

Seasonal mean and range (maximum-minimum) composites of four vegetation indices were produced using spectral reflectance products from Landsat 7 Enhanced Thematic Mapper Plus and Landsat 8 Operational Land Imager and Thermal Infrared Sensor (30-m resolution): (1) the normalized difference vegetation index (NDVI); (2) the normalized difference water index (NDWI), which is highly sensitive to vegetation water content (Jackson, 2004); (3) the enhanced vegetation index (EVI), which is an improved vegetation index with decoupling of the background signal of canopy (Huete et al., 2002); and (4) the greenness index that is sensitive to the irrigation signal (Deines et al., 2017). The vegetation indices were divided seasonally for the spring (April–June) and summer (July–September) periods to include vegetation growth and their

seasonal changes into the RFML model. The detailed variable equations are included in the supporting information (Text S1).

For thermal-moisture variables, two shortwave infrared transformed reflectances (STR) from Landsat 7 and 8 were used, which have a linear relationship with soil moisture content (Sadeghi et al., 2015). Land surface temperature (LST) from MODIS (1-km resolution) and surface soil moisture from Soil Moisture Ocean Salinity (SMOS) satellite (25-km resolution) were also used, but the soil moisture data were only available from 2010 (Kerr et al., 2010).

Climate-land variables can improve classification accuracy by refining wet versus dry year patterns and including crop type and soil property effects. Total precipitation for the growing (May to October) and nongrowing (December in the previous year to April) seasons, and aridity (precipitation scaled by reference evapotranspiration, May to August) were assembled from the University of Idaho Gridded Surface Meteorological Dataset (4-km resolution; Abatzoglou, 2013). Annual crop types from Cropland Data Layers were provided by the USDA National Agricultural Statistics Service. Three soil property maps, available water content, saturated hydraulic conductivity, and clay percent of the soils at 0–5 cm, from PLARIS database (30-m spatial resolution; Chaney et al., 2016, 2019), were also used in the RFML classification. Land cover and slope information were used to make the non-SD area (e.g., nonagricultural and high slope area). We identified low gradient cultivated crop areas (slope <2%) using the USGS National Land Cover Dataset and the USGS National Elevation Dataset (Naz et al., 2009). The input products with coarse resolutions (e.g., 1-, 4-, or 25-km grid) were disaggregated/resampled to the finer resolution (30-m grid) using a nearest neighbor resampling by default in the GEE (https://developers.google.com/earth-engine/resample).

3.2. Subsurface Drainage Permit Records for Training and Validation Data

Two separate SD permit records were used to develop training points and to validate the RFML maps, assuming the permit records are ground "truth" SD measurements: (1) a subbasin SD records obtained from the BdSW district in Minnesota (http://www.bdswd.com) and (2) the USGS records obtained from the North Dakota State Water Commission (Finocchiaro, 2016). The BdSW SD permit records were collected from 1999 to the present over the BdSW region in Minnesota (Figure 1c). The annual SD records contain locations of subsurface permit lines and the request and approved dates as GIS shape files as well as engineering design specifications. SD installation is estimated to occur within three months of permit approval. Because the BdSW SD record is a line shape file, the SD lines were buffered to provide an effective extent. A 30-m buffer (15-m buffer on either side of the line) was used based on typical SD separation and guidance from the region's agricultural engineers (Naz et al., 2009). The USGS SD records (https://www.sciencebase.gov) were issued by the ND State Water Commission and collected by the USGS over the North Dakota from 1993 to 2016 (Finocchiaro, 2016). The USGS SD records provide polygon outlines of the permit areas and approval dates.

Previous studies used the U. S, Census of Agriculture drainage data (Kelly et al., 2017; Krapu et al., 2018; USDA National Agricultural Statistics Service, 2014). The Census data are extremely limited because the five available census years only provide a single SD value for each county and year in several U.S. states, do not include areas less than 2 km^2 (Kelly et al., 2017). In contrast to previously used Census SD data, the BdSW and USGS SD permit records provide greatly improved information (e.g., exact SD locations and approval dates) and are well suited for developing training and validation data.

That said, the BdSW and USGS SD records are not perfect. Errors in the records may occur if farmers did not install the system or if they were installed them later than originally planned. The permit records also depend on an institution's policy. The North Dakota SD permit policy was changed in 2011, likely resulting in uncertainties about the SD permit records (North Dakota Century Code; https://www.legis.nd.gov/cencode/t61.html). After 2011, farmers in SD no longer needed to obtain a permit to install SD systems if the SD footprint is less than 0.32 km^2 (80 acres). Thus, in small fields, SD is underreported.

The RFML uses the satellite products to identify changes in surface vegetation and soil water state that result from SD. However, even within a single field, SD effects depend on the soils, slope, and vegetation as well as the distance from the SD. The satellite product's spatial resolution (30 m) is relatively fine compared to a field scale and captures within field variations of SD effects. Additionally, farmers install SD systems over their fields with different SD intensities (e.g., depth and spacing) depending on field-specific soils, crop type,

and cost (Blann et al., 2009). Thus, matching satellite detected effects of SD to permitted SD locations is somewhat problematic. Neither the USGS polygon outlines of fields with SD nor the static 30-m buffered SD lines provided for the BdSW SD, areas can be expected to perfectly capture the portion of the field that is affected hydrologically by SD as resolved by 30-m satellite observations.

In this study, the annual accumulated SD permit records were used to classify SD and undrained (UD) areas. The low-slope cropland areas (slope <2%) without the SD permit areas were defined as the UD areas. Pixels were then randomly selected from the buffered SD and UD areas using a random sample generator in the R package. For the BdSW, the total number of sample pixels is 2164, 2150, 4710, and 4746 for 2009, 2011, 2014, and 2017, respectively. For the ND-RRB, training sample pixels were directly selected from the accumulated SD and UD areas for each year. There were total 9016, 8880, 8766, and 8754 sample pixels for 2009, 2011, 2014, and 2017, respectively. For each region and year, half of the sample pixels were randomly selected as training pixels and the remaining 50% were used to validate the model outputs.

3.3. RFML Classification

RFML is an ensemble ML classification method comprised of a collection of tree-structured classifiers (Breiman, 2001). The major principle behind ensemble learning methods is that a group of weak classifiers (or learner) can be joined to form a strong classifier. In ensemble learning, two well-known methods are boosting (Freund et al., 1999) and bootstrap aggregation (or "bagging"; Breiman, 1996) of classification trees. Compared to a single classification tree, the bagging method is used to reduce the variance of the tree. The method creates several subsets of bootstrapped samples from original training data set chosen randomly with replacement. Each collection of subset samples is used to independently train a classification tree. In the end, an ensemble of all different trees (models) is constructed. A simple majority vote is taken for prediction which is more robust than a single classification tree. However, bagging method as an ensemble learning often do not work because classification trees in bagging are developed independently by using all variables. The bagging method is allowed to look through all variables to choose the best split point (specific variable and its value) at each node in each tree. If there exists one very strong variable for prediction, most or all of the bagged trees use the strong predictor in the top split. In this case, most bagged trees look very similar and their predictions also highly correlated. This means that the results from the highly correlated trees does not accomplish a substantial reduction in variance over a single tree.

To overcome the limitation, the RFML is an improved extension over the bagging which applies randomness to the procedure when taking a subset of variables rather than using all variables to grow trees. In other word, while in decision tree each node is split using the best among all variables, in a random forest each node is split by the best among the subset of variables (Liaw & Wiener, 2002). For example, the first tree in a random forest is constructed using a few variables, not all 36 variables, and the other trees can be developed by using different numbers of variables until each node is "pure." The development procedure in RFML model and classification processes are illustrated in Figure 2.

With the growth of satellite "big" data in hydrology, the RFML was widely used for tasks such as streamflow prediction and flood risk assessment, which have been notoriously difficult with traditional approaches (Belgiu & Drăguţ, 2016; Ma et al., 2017; Petty & Dhingra, 2018; Wang et al., 2015). In this study, to determine if the RFML SD outputs are affected by spatial scale (basin versus watershed), we developed and ran the RFML model using the same input variables for two regions with different scales, separately.

For each of the training pixels, values were extracted from the 36 input layers to train the RFML algorithm. The full training data set was used to train RF classifiers for each year in the GEE. An RF classifier was performed with 300 trees. We applied the annual classifier to the corresponding year, 2009, 2011, 2014, or 2017. After the initial classification, a 3×3 majority filter was applied to remove isolated SD pixels which were sparsely scattered on maps, because SD systems are usually installed in fields (e.g., a few hundred meters). In RFML, the outcome of implicit feature relevance for each variable is visualized by the Gini index (Breiman, 2001). A Gini index analysis was conducted using R (Liaw & Wiener, 2002) because the GEE does not provide relative importance metrics. The mean decrease in Gini index is a measure of how each variable contributes to the RFML classification.

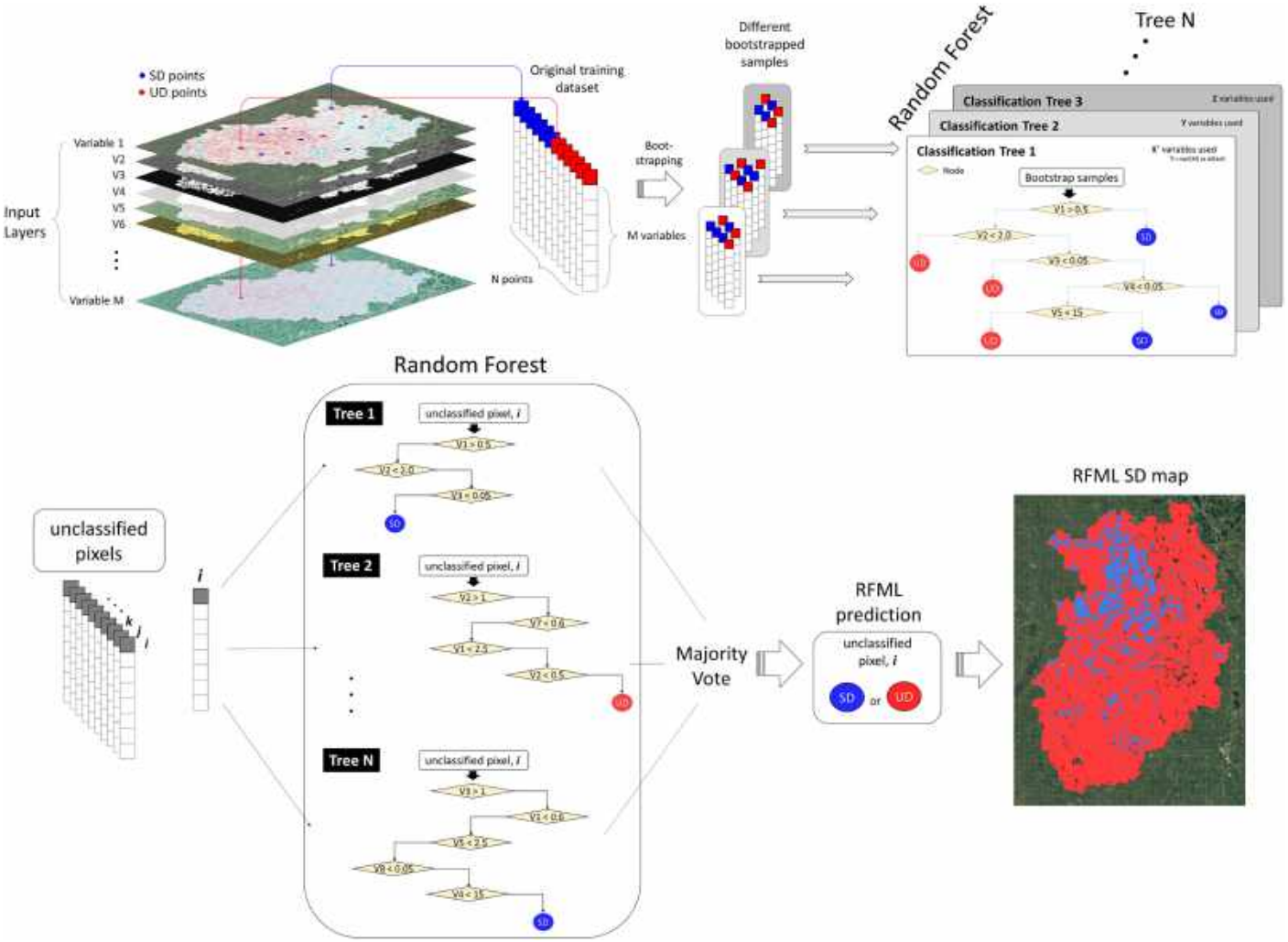

Figure 2. Scheme of construction of the random forest machine learning (RFML) model using training data and classification processes using the RFML model for classifying subsurface drainage (SD)/undrained (UD) areas

The Gini index, $i(\tau)$, at each node (τ) within a tree (T) of the RFML is defined as

$$i(\tau) = 1 - \sum_{j=1}^{n} P_j^2 \quad \text{within the tree } T \tag{1}$$

where P_j is the fraction of the N_j samples from class j out of the total of N samples at node τ in T. For a binary class $j = \{\text{SD, UD}\}$ like the current study, the Gini index is calculated by

$$i(\tau) = 1 - P_{\text{SD}}^2 - P_{\text{UD}}^2 \tag{2}$$

The decrease in Gini index, $\Delta i(\tau)$, that results from splitting the samples to two subnodes τ_{SD} and τ_{UD} (with respective sample fractions $P_{SD} = \frac{N_{\tau\text{SD}}}{N_\tau}$ and $P_{\text{UD}} = \frac{N_{\tau\text{UD}}}{N_\tau}$) by threshold t_θ on variable θ is defined as

$$\Delta i_\theta(\tau) = i(\tau) - P_{\text{SD}} \cdot i(\tau_{\text{SD}}) - P_{\text{UD}} \cdot i(\tau_{\text{UD}}) \tag{3}$$

Mean decrease in Gini index for a variable θ is the average of a variable's total decrease in node impurity over all trees N_T in the forest, weighted by the proportion of samples for all nodes τ where variable θ is used.

$$\text{Mean decrease in Gini index}(\theta) = \frac{1}{N_T}\sum^{T}\sum^{\tau} pr(\tau)\cdot\Delta i_{\theta}(\tau, T) \quad (4)$$

where $pr(\tau)$ is the proportion N_τ/N of samples reaching τ.

3.4. Accuracy Assessment (Validation)

The BdSW and USGS SD permit records were used separately to perform an accuracy assessment based on a pixel-level confusion matrix and subwatershed- and subbasin-level statistics. For the BdSW, a pixel-by-pixel comparison was conducted. The number of correct and incorrect predictions was summarized as a confusion matrix using the validation pixels, 1082, 1075, 2355, and 2373 pixels for years 2009, 2011, 2014, and 2017, respectively. For the subwatershed-level accuracy assessment within BdSW, the RFML SD area and the SD permit area were aggregated for each of the 34 subwatersheds after masking all training pixels. For the larger scale analysis, a pixel-level comparison was conducted in the same way with the BdSW analysis, but using a larger numbers of validation pixels, 4508, 4440, 4383, and 4377 pixels for years 2009, 2011, 2014, and 2017, respectively. For the subbasin-level accuracy assessment, RFML SD areas and the USGS SD permit data were aggregated and compared using the 48 NOAA river forecasting subbasins.

4. Results and Discussion

4.1. Classification Performance

The RFML classifier, using a combination of satellite-based vegetation, thermal, and soil moisture products, along with soil property and climate variables, produced annual SD maps for BdSW and ND-RRB in 2009, 2011, 2014, and 2017. Using 2,240 SD and 4,630 UD validation pixels, the pixel-level evaluation at BdSW had an overall accuracy of 77% (True positive: 1,018 SD pixels and True negative: 4,262 UD pixels) for the four years with accuracies ranging 72% to 84% for individual years (Table 2). For undrained pixels, the RFML model was 92% accurate with a range of 88% to 98%. SD pixels had relatively lower accuracies with 45% total accuracy. In the BdSW, there is good qualitative agreement between the SD expansion maps, SD permit areas, and RFML maps (Figure 3a). The RFML model results indicate that SD extent is 2.5%, 3.4%, 11.2%, and 16.1% of total BdSW area for 2009, 2011, 2014, and 2017, respectively. These values are quantitatively similar to the extent found using the SD permit records, 1.9%, 3.2%, 10.3%, and 14.3%, from 2009, 2011, 2014, and 2017, respectively. The RFML SD extents are slightly greater than those determined from permit data, by 0.2–1.8%.

Aggregated to the subwatershed-level (HUC12), the RFML SD estimates showed strong correlation ($r = 0.88$ 0.96) with SD permit areas (Figure 3b). However, RFML consistently overestimated subsurface drained areas in each subwatershed in BdSW. The overestimated SD was also found in other dry years (see Figure S2). A review of individual fields suggests that the RFML model may be capable of identifying SD effects even in small areas within a field where SD systems can exist, but for which there is no permit record (fields 1 and 2 in Figure 4). The RFML identified numerous small fields as having SD that were likely not included in the permit record because permits are not required when a field is smaller than 0.32 km^2. Additionally, the RFML detected the extent of the installed SD effect appears to frequently extend well beyond the 30-m buffer recommended in earlier literature and expert guidance (all fields in Figure 4).

For the ND-RRB region, the RFML model achieved an overall accuracy of 87%. Class specific SD and UD accuracies ranged from 20% to 59% and 98% to 99% with overall accuracies of 40% and 98%, respectively. In both regions, overall accuracies in the early years (2009 and 2011) are higher than those in recent years (2014 and 2017). SD systems were originally installed at those sites that needed them most. Therefore, training points developed in early years may retain stronger SD/UD character in this region. A subbasin-level comparison between the RFML maps and the USGS SD permit areas conducted for the NOAA river forecasting subbasins found r values ranged from 0.77 to 0.84 for the four years (Figure 5).

Again, the RFML overestimated the subsurface drained area, especially in the few subbasins that have dense SD areas. There are very few SD areas in the northern part of the RRB. SD areas are concentrated in the southern part of the RRB (Figure 6a). In North Dakota, the 2017 predicted SD map near Sheyenne National Grassland showed good spatial agreement with the SD permit map (Figures 6b and c). However, the RFML maps appear to underestimate SD areas in Minnesota areas compared to previous findings

Table 2
Point-Based Accuracy Assessment for the Four Years (2009, 2011, 2014, and 2017) Between RFML Predicted Values and SD-Permit Based Data in the BdSW and the ND-RRB.

		BdSW			ND-RRB		
Year	Class	RFML SD	RFML UD	Overall accuracy	RFML SD	RFML UD	Overall accuracy
2009	SD	19.8%	4.3%	79.0% (850/1,076)	59.4%	1.0%	90.7% (4,170/4,596)
	UD	80.2%	95.7%		40.6%	99.0%	
2011	SD	35.9%	2.4%	83.9% (894/1,066)	40.3%	1.3%	86.6% (3,909/4,512)
	UD	64.1%	97.6%		59.7%	98.7%	
2014	SD	51.9%	8.7%	77.3% (1,820/2,355)	26.5%	2.1%	82.8% (3,693/4,461)
	UD	48.1%	91.3%		73.5%	97.9%	
2017	SD	48.7%	12.4%	72.3% (1,716/2,373)	19.8%	2.0%	81.6% (3,632/4,453)
	UD	51.3%	87.6%		80.2%	98.0%	
Overall accuracy		45.4% (1,018/2,240)	92.1% (4,262/4,630)	76.9% (5,280/6,870)	39.9% (1,380/3,460)	98.4% (14,024/14,247)	87.0% (15,204/17,708)

Note. RFML = random forest machine learning; SD = subsurface drainage; UD = undrained; BdSW = Bois de Sioux Watershed; ND-RRB = North Dakota portion of the RRB region.

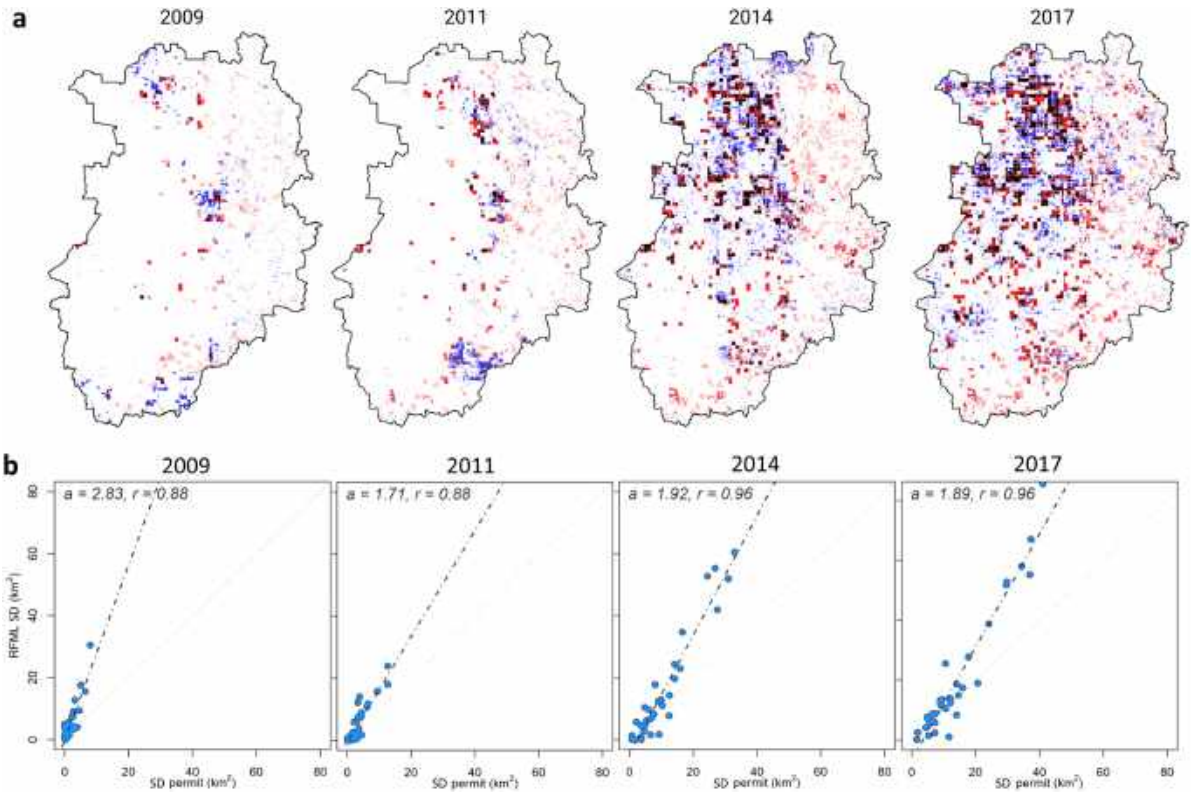

Figure 3. (a) Subsurface drainage expansion in Bois de Sioux watershed, Minnesota in 2009, 2011, 2014, and 2017 from subsurface drainage (SD) permit records (red color) and predicted SD areas (blue color) derived by random forest machine learning (RFML) classification in the Google Earth Engine. Black color indicates overlapped SD areas of the two sources. (b) Subwatershed (HUC12)-level accuracy assessment over Bois de Sioux Watershe (BdSW), Minnesota (N = 34). Subsurface drained permit area from the BdSW district permit records compared with subsurface drained area from RFML classified maps against a 1:1 line (light dashed). Agreement between the two data sets was assessed with correlation coefficient (r) metrics from simple linear regression (trend line = thick dashed line, a = slope).

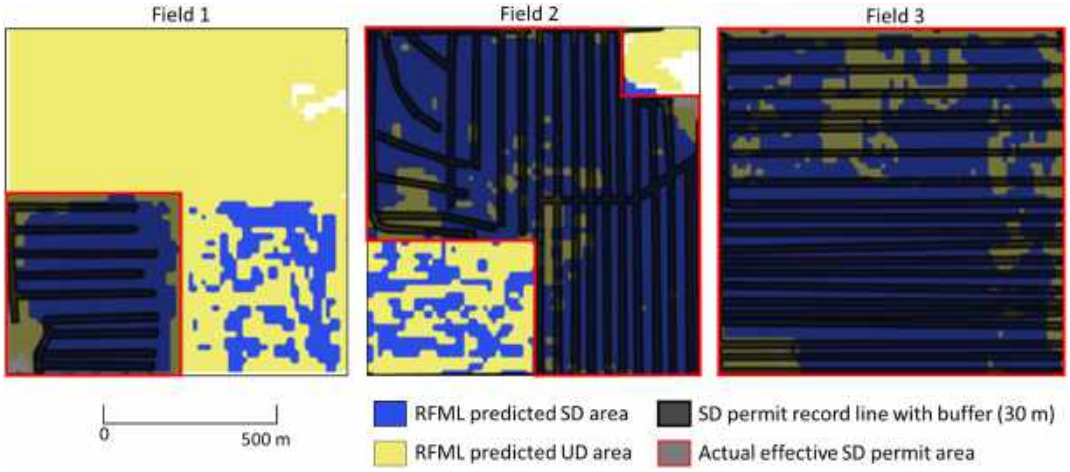

Figure 4. Examples of fields showing areal difference between subsurface drainage (SD) permit area using buffer function and actual SD effective area in Bois de Sioux watershed, Minnesota. These examples indicate that SD permit buffered areas in this study were underestimated in these fields compared to actual SD effective areas. RFML = random forest machine learning; UD = undrained.

(Kelly et al., 2017; Nakagaki & Wieczorek, 2016). This indicates that additional training points in Minnesota are required to develop more accurate RFML models.

4.2. Variable Importance

The mean decrease in Gini index was used to determine the relative contribution of each of the 36 input variables for the SD classification. Larger mean decreases in Gini index indicate variables that play a greater role in partitioning the data into the SD/UD classification. Soil properties (available water content, awc, clay percentage, clay_perc, and saturated hydraulic conductivity, ksat, in this study) ranked the highest for both regions (Figure 7). Climate variables, precipitation, and aridity also were important, especially for the larger scales. For both regions, LST contributed strongly to the classification. Soil moisture showed minimal importance even though subsurface drains are intended to enhance drainage. This may be due to the coarse resolution (25 km) from the SMOS satellite observations. The importance of spring thermal and wetness variables (e.g., LST and STR2) is noted. These indices warrant further study for use in SD/UD classification in other agricultural regions. Interestingly, no vegetation-related variables were in the top 10. NDWI scored relatively high among the four vegetation indices, indicating only water-related vegetation variables may enhance accuracy in this region.

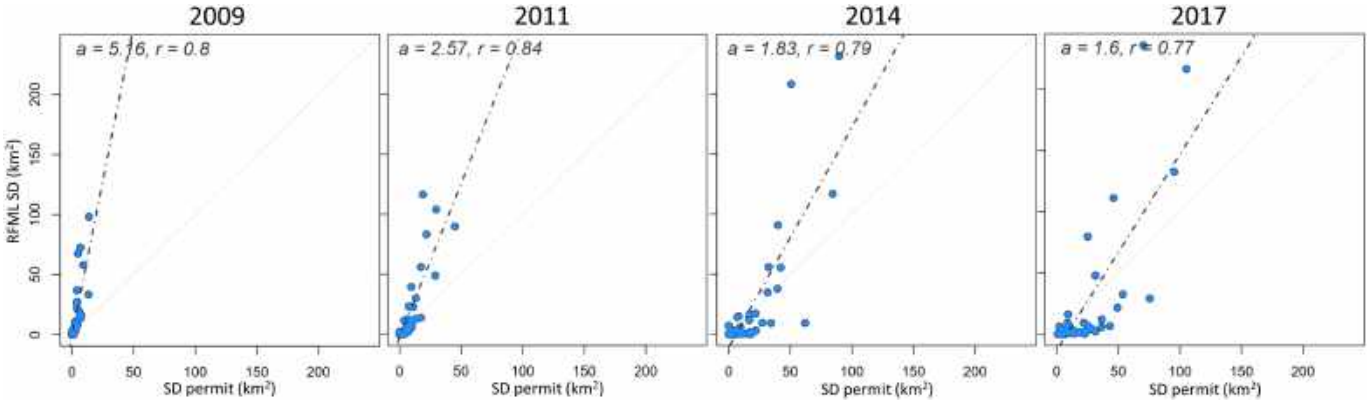

Figure 5. National Oceanic and Atmospheric Administration (NOAA) subbasin-level accuracy assessment over North Dakota portion of the RRB region ($N = 48$). NOAA subbasin is hydrological unit to operate the river forecasting system, NOAA River Forecast Centers. Subsurface drained permit area from the Bois de Sioux Watershed district permit records compared with subsurface drained area from random forest machine learning classified maps against a 1:1 line (light dashed). Agreement between the two data sets was assessed with correlation coefficient (r) metrics from simple linear regression (trend line = thick dashed line, a = slope). Note that the ranges of y axis are different.

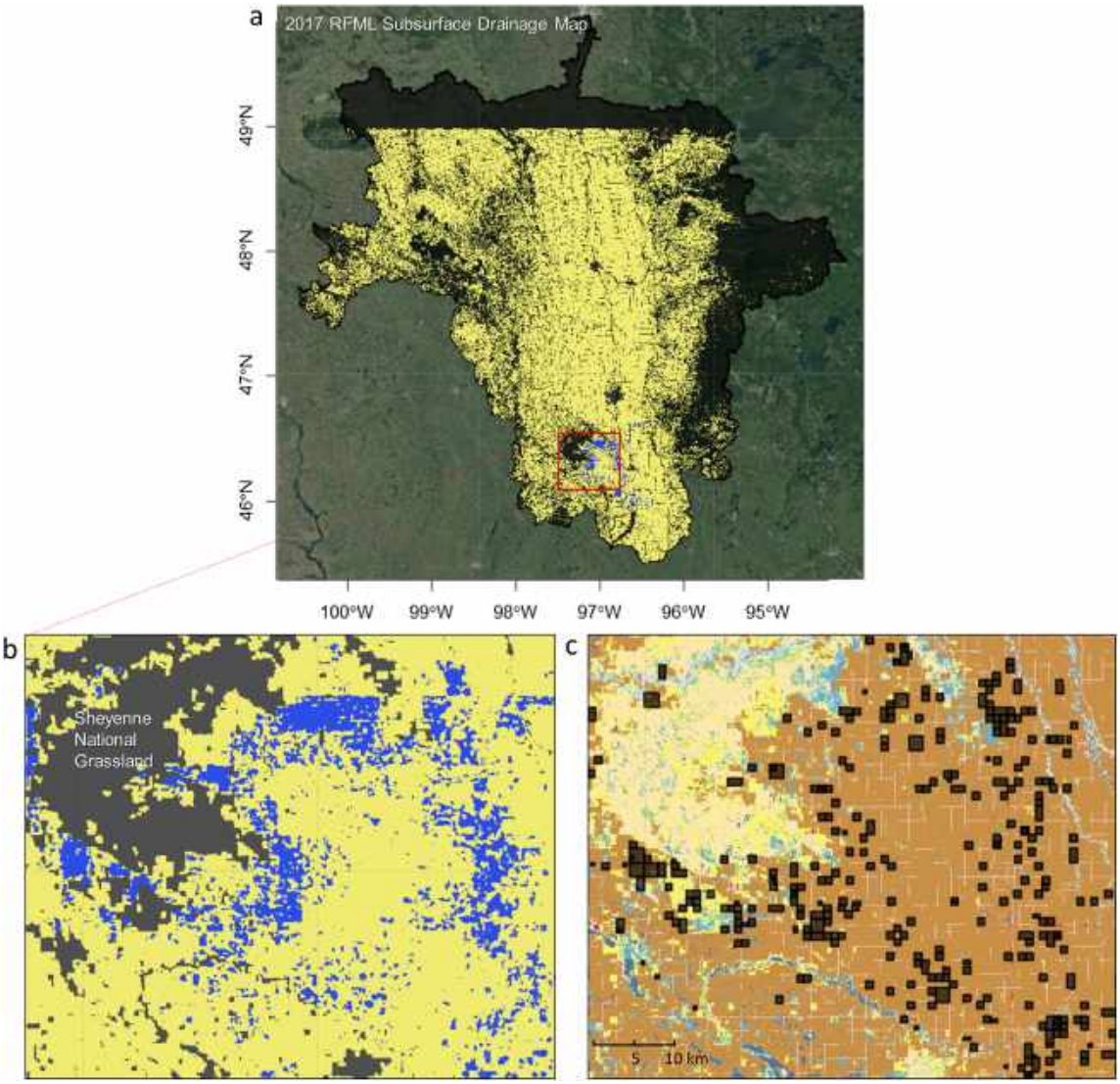

Figure 6. (a) Subsurface drainage map from random forest machine learning (RFML) over the RRB in 2017. (b) A close-up map near Sheyenne National Grassland in North Dakota. Blue colors indicate predicted subsurface drainage areas. Yellow colors indicate undrained area. (c) U.S. Geological Survey subsurface drainage permit records (Finocchiaro, 2016) overlaying the National Land Cover Database 2011 (Yang et al., 2018) with same legends in Figure 1.

It is possible that the accuracies in the RFML SD map are improved with new relevant data as an input variable. To test this, Sentinel-1 Synthetic Aperture Radar (SAR) Ground Range Detected C-band backscatter data (VV polarization, ImageCollection ID: COPERNICUS/S1_GRD in GEE) was included in current RFML model as additional input variables (two spring mean and range layers) in 2017. In BdSW, the RFML SD map with the Sentinel-1 SAR information shows slightly better accuracies than the original SD without Sentinel-1 SAR (Table 3). The point-based accuracies in RFML SD and UD predictions were improved by 0.3% (from 48.7% to 49%) and 0.9% (from 87.6% to 88.5%), respectively (the overall accuracy from 72.3% to 73.0%). In the subwatershed-level assessment, the two SD maps with/without Sentinel-1 SAR have the same correlations ($r = 0.96$) with similar slopes (Figure 8). However, in the ND-RRB, there is no clear improvement in SD map accuracies based on the both point-based and subbasin-level assessments. Given that the Sentinel-1 SAR backscattering signal is directly related to surface soil moisture, we expect that any improvements of the SD prediction map by Sentinel-1 data would be much clearer in a wet year. This also suggests that the current RFML SD model can be steadily improved by including (or replacing) new SD-related variable information. The Sentinel-1 SAR and RFML SD maps were

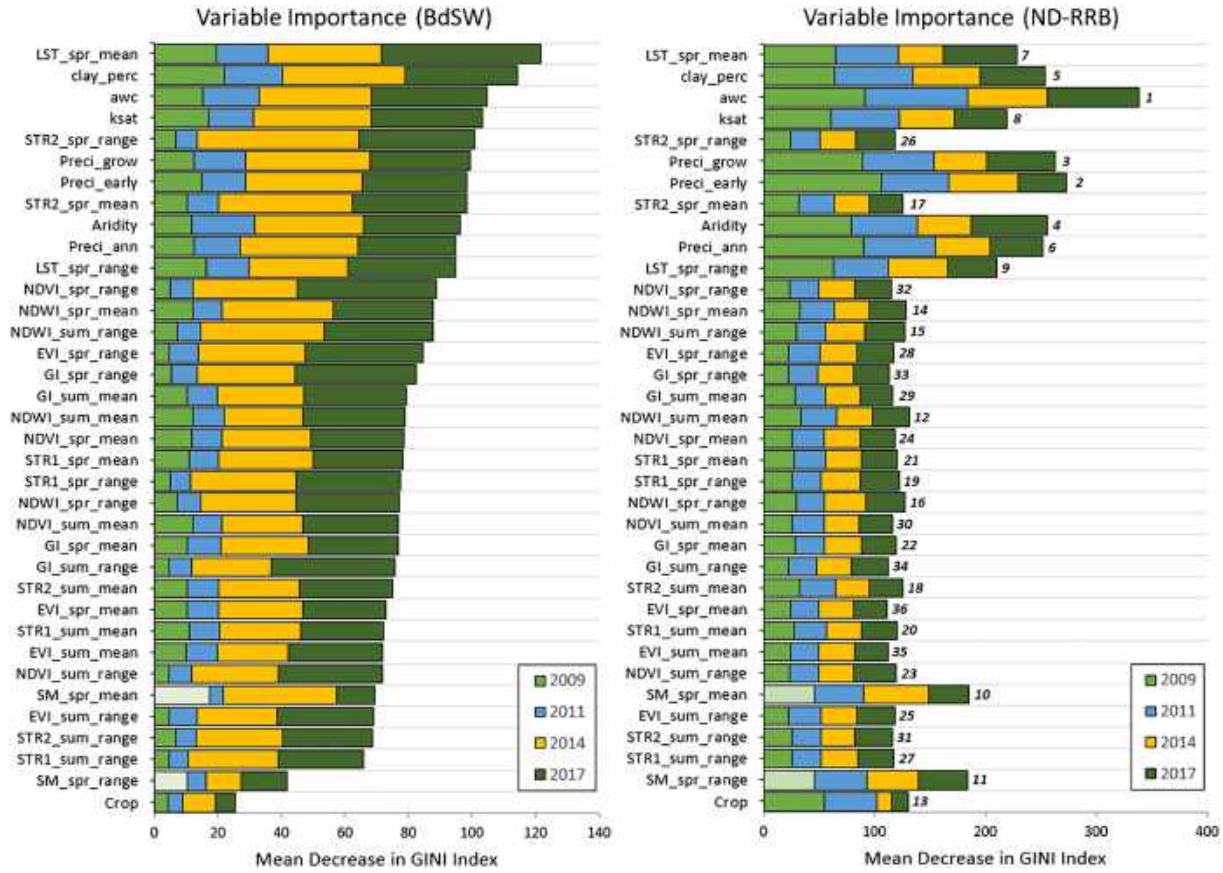

Figure 7. Variable importance in the random forest machine learning classification for two regions with different spatial scale (a) BdSW and (b) ND-RRB. For BdSW, variables with their short names were arranged from largest (top) to smallest (bottom) of the accumulated mean decrease in Gini index. Variables in RRB was arranged in same order to those of BdSW. The numbers at the edge of the bar indicate the ranks of each variable. Due to the absence of Soil Moisture Ocean Salinity soil moisture in 2009, we calculated mean decreases in Gini index of the spring soil moisture mean and range by averaging the other three years' values. Their full names were given in Table 1.

provided in supporting information (Figure S3). (Note: Subwatershed-level accuracy assessments over the BdSW using the 10 most important variables only are provided in Figure S4.)

4.3. Comparison With Recent Studies

In the RRB, Kelly et al. (2017) reported that the 2012 SD area was 1,340 km^2, 2.0% of the entire basin area, using the county-level agricultural census drainage data (USDA National Agricultural Statistics Service, 2014). This is larger than our predicted SD areas (916 km^2) for 2011. There are two potential reasons

Table 3
Comparison of RFML SD Maps Between With and Without Sentinel-1 Synthetic Aperture Radar (SAR) C-Band Backscatter Data Based On Point-Based Accuracy Assessments in 2017

		Without Sentinel-1			With Sentinel-1		
Year	Class	RFML SD	RFML UD	Overall Accuracy	RFML SD	RFML UD	Overall Accuracy
BdSW	SD	48.7%	12.4%	72.3% (1716/2373)	49.0%	11.5%	73.0% (1732/2373)
	UD	51.3%	87.6%		51.0%	88.5%	
ND-RRB	SD	19.8%	2.0%	81.6% (3,632/4,453)	19.6%	1.8%	81.7% (3,639/4,453)
	UD	80.2%	98.0%		80.4%	98.2%	

Note. RFML = random forest machine learning; SD = subsurface drainage; UD = undrained; BdSW = Bois de Sioux Watershed; ND-RRB = North Dakota portion of the RRB region.

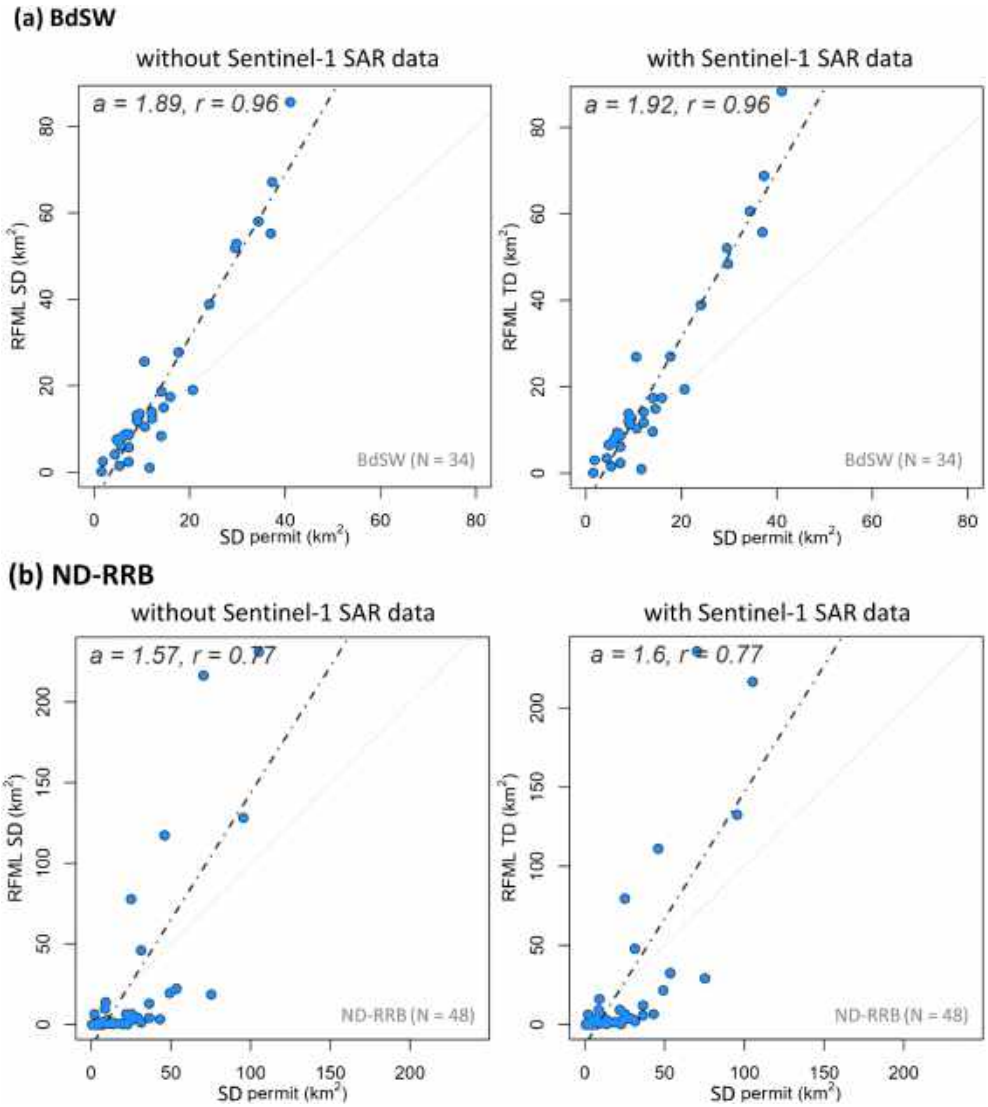

Figure 8. Comparison of random forest machine learning (RFML) subsurface drainage (SD) maps between with and without Sentinel-1 Synthetic Aperture Radar (SAR) C-band backscatter data based on (a) Subwatershed (HUC12)-level accuracy assessment over Bois de Sioux Watershed, Minnesota ($N = 34$) and (b) National Oceanic and Atmospheric Administration subbasin-level accuracy assessment over North Dakota portion of the RRB region (ND-RRB; $N = 48$).

for the difference. They defined the "RRB region" as being upstream of Grand Forks, North Dakota in United State (67,005 km^2), which is the southern part of our RRB. We used the entire RRB region except for the area in Canada (101,500 km^2) where the northern part of the RRB is almost entirely undrained. There is also a year gap between our results in 2011 and SD estimates in 2012 from Kelly et al. (2017). The USGS SD permit records for the RRB region in North Dakota indicated an increase of 114 km^2 between the two years. There is likely a similar increase in Minnesota (Dollinger et al., 2013).

Most previous studies were conducted at a smaller scale (e.g., field or catchment scale) than the current study and used stepwise GIS-based analyses and aerial image processing techniques (Naz et al., 2009; Naz & Bowling, 2008; Tetzlaff, Kuhr, Vereecken, & Wendland, 2009; Zhang et al., 2014). They showed spatial agreement with overall accuracies of 78% (Tetzlaff, Kuhr, Vereecken, & Wendland, 2009) and 85% (Naz & Bowling, 2008), which are similar to the performance of the current study (76–86%). Zhang et al. (2014) and Naz and Bowling (2008) partially explained the causes of discrepancies in SD estimates within fields in the current study (e.g., Figure 4). In Canadian subsurface drained fields, Zhang et al. (2014) used

unmanned aerial vehicle-based NDVI and found within field NDVI differences due to SD line locations. Naz and Bowling (2008) also found that within-field soil variability can lead to SD misclassification. Satellite data were also used by Møller et al. (2018) to identify subsurface drained areas in a 43,000-km^2 region in Denmark using an ensemble of ML models with similar input variables to the current study. Møller et al. (2018) is the sole previous study applying ML methods to SD detection. However, they only used one month (March 2014) of Landsat 8 imagery. Their final ensemble contained 36 unique models that predicted SD areas with an accuracy of 76.5%. The results from our current study have better accuracies of 76.9% to 87.0%. This suggests that using an ensemble of multisource satellite data including seasonal thermal, reflectance, and vegetation input variables could improve results. They also found soil property (e.g., clay content) to be the most important variable, followed by precipitation. This corresponds with our finding that available water content of the soil is the most important variable. Clay percentage ranked in the top five in the RRB region. Climate variables are important at larger scales (Møller et al., 2018; Tetzlaff, Kuhr, & Wendland, 2009). Additionally, we found that LST is the most important variable at a relatively small scale. This seems reasonable considering that drainage systems have significant impacts on surface heat flux and land surface water dynamics. Jacobs et al. (2017) found that spring LST, obtained by subtracting past mean values (2002–2006) from recent values (2013–2017), has significant relationships (r^2 = 0.85 and 0.83, respectively) with the SD density based on a subwatershed-level analysis.

Previous studies also identified uncertainties. Tetzlaff, Kuhr, and Wendland (2009) noted the difficulty of acquiring aerial images at the right time associated with rainfall events and vegetation growth for a large area. Sugg's (2007) GIS analysis based on soil drainage class and land cover in the Midwest United States overestimated total SD in Minnesota by 3,643 km^2 compared to the 1992 National Resource Inventory (NRI) including inflated estimates of SD for the RRB region. Their GIS method identified large areas in northwest Minnesota as SD areas because they are poorly drained soils and cultivated lands. However, the actual SD installations result from not only geophysical characteristics but also socio-economic demand for drainage (Blann et al., 2009). Care must be taken to differentiate between models that predict potential areas requiring SD systems based on drainage properties versus those that are able to discern areas where SD has been installed.

Belgiu and Drăguţ (2016) found that the RFML method can handle multisource satellite data dimensionality and multicollinearity with fast processing and insensitivity to overfitting. However, it tends to be sensitive to training samples (Colditz, 2015), which correspond with our finding in the process of this analysis. We found that the outputs from the RFML method of the current study were sensitive to the proportion of SD/UD training samples in several trials (not shown). The proportional allocation of SD/UD training samples to each class based on SD permit records achieved the best results because the UD class has much larger areas and requires more training samples than the SD class that occupies limited areas. Further investigations are needed to better understand sample proportioning for RFML. Furthermore, studies are needed that compare the performance among multiple ML methods.

5. Conclusion and Future Perspectives

Subsurface drainage systems were mapped at 30-m resolution by leveraging a ML technique and multisource "big" data sets from operational satellites, Landsat-based vegetation indices (NDVI, EVI, NDWI, and GI) and STR, MODIS LST, and SMOS soil moisture, along with USGS National Land Cover and Slope Datasets, USDA Cropland Data Layer, soil properties from POLARIS, and climate variables from GRIDMET over the RRB region. RFML was conducted in the GEE cloud computing platform and used SD permit records from the USGS and the BdSW district for training and validation. The RFML maps showed spatial agreement with SD permit records and correlated well with HUC12 subwatershed statistics. The RFML maps appear to be capable of identifying within field variations in SD effects and capturing the overall SD expansion over time including for those fields whose acreage was less than that required to be permitted. Soil properties, climate variables, and LST are the strongest predictors of SD. Predictor variables differed between the two scales, suggesting that SD models are sensitive to the spatial scale. Using the Sentinel-1 SAR data, we demonstrated the RFML SD model could be further improved with new relevant data. This ML technique can be readily applied to other regions and future years to provide updated information about SD expansion to regional water managers and flood forecasters. However, this technique

relies on the availability of baseline data sets (e.g., permit records) and more of these data sets may be needed for other regions.

There are future opportunities to further improve the SD classification (or similar work with demanding land use and land cover detection/classification) using ML algorithms. As a limitation of the current RFML method like other nondeep learning algorithms, the input layers must be developed from raw data with formulas or retrieval algorithms provided by experts for each input data and can be labor intensive. In this context, deep learning (DL) has substantial potential to overcome this weakness. The DL method, a layered structure of advanced artificial neural network algorithm, allows the automatic extraction of features from raw data by capturing abstract spatial or temporal structures hidden in data (Bengio et al., 2013; Shen, 2018). Also, the use of new remote sensing platforms such as CubeSat and Unmanned Aerial Vehicles can add value for enhanced SD identification (McCabe et al., 2017; NASA, 2017; NASA CubeSat Launch Initiative, 2018; Planet Team, 2018). For example, more than 130 CubeSats launched by Planet (http://www.planet.com) currently provide daily visible (Red-Green-Blue) and near-infrared imagery with ultrahigh resolutions (e.g., 3 m and 72 cm), capturing daily near-global coverage (Planet Team, 2018). This imagery could potentially greatly improve SD identification with ML or DL methods.

Acknowledgments
The authors gratefully acknowledge support from NASA Water Resources Applied Sciences Program (NNX15AC47G). The RFML SD maps from this study are available on Hydroshare at http://www.hydroshare.org/resource/f2f7a9cfbae1451f85b5c0dc3938b9a1. We are grateful to the Martyn Clark (Editor-in-Chief), Chaopeng Shen (Associate Editors), and three anonymous reviewers who have contributed to improving this manuscript. We thank Ronny Schroder (UNH) for constructive discussions; Carrie Vuyovich (NASA), Mike Cosh (USDA), Samuel Tuttle (Mount Holyoke College), Pedro Restrepo, Mike DeWeese, and Brian Connelly (NOAA NCRFC) for their comments at the early stage of this research through the NASA RRB project. We are also grateful to Jillian Deines (Stanford University) who provided her codes and guidance on analyses. The BdSW SD permit records were obtained from the BdSW district in Minnesota (http://www.bdswd.com) and the USGS SD records are publicly available from the USGS Science Base website (https://www.sciencebase.gov). All satellite and model data used as inputs are available through the Google Earth Engine code editor (https://code.earthengine.google.com), except for soil properties (freely available at www.polaris.earth).

References

Abatzoglou, J. T. (2013). Development of gridded surface meteorological data for ecological applications and modelling. *International Journal of Climatology*, *33*(1), 121–131. https://doi.org/10.1002/joc.3413

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, *114*, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798–1828. https://doi.org/10.1109/TPAMI.2013.50

Blann, K. L., Anderson, J. L., Sands, G. R., & Vondracek, B. (2009). Effects of agricultural drainage on aquatic ecosystems: A review. *Critical Reviews in Environmental Science and Technology*, *39*(11), 909–1001. https://doi.org/10.1080/10643380801977966

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123–140.

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

Chaney, N. W., Minasny, B., Herman, J. D., Nauman, T. W., Brungard, C. W., Morgan, C. L. S., et al. (2019). POLARIS soil properties: 30-m probabilistic maps of soil properties over the contiguous United States. *Water Resources Research*, *55*, 2916–2938. https://doi.org/10.1029/2018WR022797

Chaney, N. W., Wood, E. F., McBratney, A. B., Hempel, J. W., Nauman, T. W., Brungard, C. W., & Odgers, N. P. (2016). POLARIS: A 30-meter probabilistic soil series map of the contiguous United States. *Geoderma*, *274*, 54–67. https://doi.org/10.1016/j.geoderma.2016.03.025

Colditz, R. (2015). An evaluation of different training sample allocation schemes for discrete and continuous land cover classification using decision tree-based algorithms. *Remote Sensing*, *7*(8), 9655–9681. https://doi.org/10.3390/rs70809655

Deines, J. M., Kendall, A. D., & Hyndman, D. W. (2017). Annual irrigation dynamics in the U.S. Northern High Plains derived from Landsat satellite data. *Geophysical Research Letters*, *44*, 9350–9360. https://doi.org/10.1002/2017gl074071

Dollinger, D., Lundeen, B., Stroom, K., Anderson, P., Monson, B., Nelson, S., et al. (2013). Bois de Sioux River watershed monitoring and assessment report, Minnesota Pollution Control Agency, Saint Paul, MN.

Eastman, M., Gollamudi, A., Stämpfli, N., Madramootoo, C. A., & Sarangi, A. (2010). Comparative evaluation of phosphorus losses from subsurface and naturally drained agricultural fields in the Pike River watershed of Quebec, Canada. *Agricultural Water Management*, *97*(5), 596–604. https://doi.org/10.1016/j.agwat.2009.11.010

Finocchiaro, R. G. (2014). Agricultural subsurface drainage tile locations by permits in South Dakota. U.S. Geological Survey data release. doi:https://doi.org/10.5066/F7KS6PNW

Finocchiaro, R. G. (2016). Agricultural subsurface drainage tile locations by permits in North Dakota. U.S. Geological Survey data release. doi:https://doi.org/10.5066/F7QF8QZW

Foufoula-Georgiou, E., Takbiri, Z., Czuba, J. A., & Schwenk, J. (2015). The change of nature and the nature of change in agricultural landscapes: Hydrologic regime shifts modulate ecological transitions. *Water Resources Research*, *51*, 6649–6671. https://doi.org/10.1002/2015WR017637

Frans, C., Istanbulluoglu, E., Mishra, V., Munoz-Arriola, F., & Lettenmaier, D. P. (2013). Are climatic or land cover changes the dominant cause of runoff trends in the Upper Mississippi River Basin? *Geophysical Research Letters*, *40*, 1104–1110. https://doi.org/10.1002/grl.50262

Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting *Journal-Japanese Society For Artificial Intelligence*, *14*(771–780), 1612.

Ge, Y., Hu, S., Ren, Z., Jia, Y., Wang, J., Liu, M., et al. (2019). Mapping annual land use changes in China's poverty-stricken areas from 2013 to 2018. *Remote Sensing of Environment*, *232*, 111285. https://doi.org/10.1016/j.rse.2019.111285

Gökkaya, K., Budhathoki, M., Christopher, S. F., Hanrahan, B. R., & Tank, J. L. (2017). Subsurface tile drained area detection using GIS and remote sensing in an agricultural watershed. *Ecological Engineering*, *108*, 370–379. https://doi.org/10.1016/j.ecoleng.2017.06.048

Gómez, C., White, J. C., & Wulder, M. A. (2016). Optical remotely sensed time series data for land cover classification: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *116*, 55–72. https://doi.org/10.1016/j.isprsjprs.2016.03.008

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, *202*, 18–27. https://doi.org/10.1016/j.rse.2017.06.031

Huete, A., Didan, K., Miura, T., Rodriguez, E. P., Gao, X., & Ferreira, L. G. (2002). Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment*, *83*(1-2), 195–213. https://doi.org/10.1016/S0034-4257(02)00096-2

Jackson, T. J., Chen, D., Cosh, M., Li, F., Anderson, M., Walthall, L. G., et al. (2004). Vegetation water content mapping using Landsat data derived normalized difference water index for corn and soybeans. *Remote Sensing of Environment*, *92*(4), 475–482. https://doi.org/10.1016/j.rse.2003.10.021

Jacobs, J. M., Cho, E., & Jia, X. (2017). Tile drainage expansion detection using satellite soil moisture dynamics. In AGU Fall Meeting Abstracts.

Jin, Z., Azzari, G., You, C., Di Tommaso, S., Aston, S., Burke, M., & Lobell, D. B. (2019). Smallholder maize area and yield mapping at national scales with Google Earth Engine. *Remote Sensing of Environment*, *228*, 115–128. https://doi.org/10.1016/j.rse.2019.04.016

Kelly, S. A., Takbiri, Z., Belmont, P., & Foufoula-Georgiou, E. (2017). Human amplified changes in precipitation-runoff patterns in large river basins of the Midwestern United States. *Hydrology and Earth System Sciences*, *21*(10), 5065–5088. https://doi.org/10.5194/hess-21-5065-2017

Kerr, Y. H., Waldteufel, P., Wigneron, J. P., Delwart, S., Cabot, F., Boutin, J., et al. (2010). The SMOS mission: New tool for monitoring key elements of the global water cycle. *Proceedings of the IEEE*, *98*(5), 666–687. https://doi.org/10.1109/jproc.2010.2043032

King, K. W., Fausey, N. R., & Williams, M. R. (2014). Effect of subsurface drainage on streamflow in an agricultural headwater watershed. *Journal of Hydrology*, *519*, 438–445. https://doi.org/10.1016/j.jhydrol.2014.07.035

Kladivko, E. J., Frankenberger, J. R., Jaynes, D. B., Meek, D. W., Jenkinson, B. J., & Fausey, N. R. (2004). Nitrate leaching to subsurface drains as affected by drain spacing and changes in crop production system. *Journal of Environmental Quality*, *33*(5), 1803–1813. https://doi.org/10.2134/jeq2004.1803

Krapu, C., Kumar, M., & Borsuk, M. (2018). Identifying wetland consolidation using remote sensing in the North Dakota Prairie Pothole Region. *Water Resources Research.*, *54*, 7478–7494. https://doi.org/10.1029/2018WR023338

Lenhart, C. F., Peterson, H., & Nieber, J. (2011). Increased streamflow in agricultural watersheds of the Midwest: implications for management. *Watershed Science Bulletin*, *2*, 25–31.

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

Ma, L., Li, M., Ma, X., Cheng, L., Du, P., & Liu, Y. (2017). A review of supervised object-based land-cover image classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, *130*, 277–293. https://doi.org/10.1016/j.isprsjprs.2017.06.001

McCabe, M. F., Rodell, M., Alsdorf, D. E., Miralles, D. G., Uijlenhoet, R., Wagner, W., et al. (2017). The future of Earth observation in hydrology. *Hydrology and Earth System Sciences*, *21*(7), 3879–3914. https://doi.org/10.5194/hess-21-3879-2017

Møller, A. B., Beucher, A., Iversen, B. V., & Greve, M. H. (2018). Predicting artificially drained areas by means of a selective model ensemble. *Geoderma*, *320*, 30–42. https://doi.org/10.1016/j.geoderma.2018.01.018

Nakagaki, N. & Wieczorek, M. E. (2016). Estimates of subsurface tile drainage extent for 12 Midwest states, 2012: U.S. *Geological Survey data release*, https://doi.org/10.5066/F7W37TDP

NASA (2017), CubeSat 101: Basic concepts and processes for first-time CubeSat developers, https://www.nasa.gov/sites/default/files/atoms/files/nasa_csli_cubesat_101_508.pdf (Last access: 5 August 2019)

NASA CubeSat Launch Initiative (2018), https://www.nasa.gov/directorates/heo/home/CubeSats_initiative (Last access:5 August 2019).

Naz, B. S., Ale, S., & Bowling, L. C. (2009). Detecting subsurface drainage systems and estimating drain spacing in intensively managed agricultural landscapes. *Agricultural water management*, *96*(4), 627–637. https://doi.org/10.1016/j.agwat.2008.10.002

Naz, B. S., & Bowling, L. C. (2008). Automated identification of tile lines from remotely sensed data. *Transactions of the ASABE*, *51*(6), 1937–1950. https://doi.org/10.13031/2013.25399

Northcott, W. J., Verma, A. K., & Cooke, R. A. (2000). Mapping subsurface drainage systems using remote sensing and GIS. Mapping subsurface drainage systems using remote sensing and GIS., 1-10.

Ok, A. O., Akar, O., & Gungor, O. (2012). Evaluation of random forest method for agricultural crop classification. *European Journal of Remote Sensing*, *45*(1), 421–432. https://doi.org/10.5721/EuJRS20124535

Petty, T. R., & Dhingra, P. (2018). Streamflow hydrology estimate using machine learning (SHEM). *JAWRA Journal of the American Water Resources Association*, *54*(1), 55–68. https://doi.org/10.1111/1752-1688.12555

Planet Team (2018). Planet application program interface: In space for life on Earth. San Francisco, CA. https://api.planet.com.

Rahman, M. M., Lin, Z., Jia, X., Steele, D. D., & DeSutter, T. M. (2014). Impact of subsurface drainage on streamflows in the Red River of the North basin. *Journal of Hydrology*, *511*, 474–483. https://doi.org/10.1016/j.jhydrol.2014.01.070

Randall, G. W., Vetsch, J. A., & Huffman, J. R. (2003). Nitrate losses in subsurface drainage from a corn-soybean rotation as affected by time of nitrogen application and use of nitrapyrin. *Journal of Environmental Quality*, *32*(5), 1764–1772. https://doi.org/10.2134/jeq2003.1764

Rannie, W. (2015). The 1997 flood event in the Red River basin: Causes, assessment and damages. *Canadian Water Resources Journal/Revue canadienne des ressources hydriques*, *41*(1-2), 45–55. https://doi.org/10.1080/07011784.2015.1004198

Raymond, P. A., Oh, N. H., Turner, R. E., & Broussard, W. (2008). Anthropogenically enhanced fluxes of water and carbon from the Mississippi River. *Nature*, *451*(7177), 449–452. https://doi.org/10.1038/nature06505

Rijal, I., Jia, X., Zhang, X., Steele, D. D., Scherer, T. F., & Akyuz, A. (2012). Effects of subsurface drainage on evapotranspiration for corn and soybean crops in southeastern North Dakota. *Journal of Irrigation and Drainage Engineering*, *138*(12), 1060–1067. https://doi.org/10.1061/(ASCE)IR.1943-4774.0000508

Rodgers, M., Mulqueen, J., & McHale, J. (2003). A model study of mole drain spacing and performance. *Agricultural Water Management*, *60*(1), 33–42. https://doi.org/10.1016/s0378-3774(02)00153-1

Rogger, M., Agnoletti, M., Alaoui, A., Bathurst, J. C., Bodner, G., Borga, M., et al. (2017). Land use change impacts on floods at the catchment scale: Challenges and opportunities for future research. *Water Resources Research*, *53*, 5209–5219. https://doi.org/10.1002/2017WR020723

Sadeghi, M., Jones, S. B., & Philpot, W. D. (2015). A linear physically-based model for remote sensing of soil moisture using short wave infrared bands. *Remote Sensing of Environment*, *164*, 66–76. https://doi.org/10.1016/j.rse.2015.04.007

Schilling, K. E., Chan, K. S., Liu, H., & Zhang, Y. K. (2010). Quantifying the effect of land use land cover change on increasing discharge in the Upper Mississippi River. *Journal of Hydrology*, *387*(3-4), 343–345. https://doi.org/10.1016/j.jhydrol.2010.04.019

Schottler, S. P., Ulrich, J., Belmont, P., Moore, R., Lauer, J. W., Engstrom, D. R., & Almendinger, J. E. (2014). Twentieth century agricultural drainage creates more erosive rivers. *Hydrological Processes*, *28*(4), 1951–1961. https://doi.org/doi:10.1002/hyp.9738

Shen, C. (2018). A trans-disciplinary review of deep learning research and its relevance for water resources scientists. *Water Resources Research*, *54*, 8558–8593. https://doi.org/10.1029/2018WR022643

Shokri, A., & Bardsley, W. E. (2015). Enhancement of the Hooghoudt drain-spacing equation. *Journal of Irrigation and Drainage Engineering*, *141*(6). https://doi.org/10.1061/(asce)ir.1943-4774.0000835

Shortridge, J. E., Guikema, S. D., & Zaitchik, B. F. (2016). Machine learning methods for empirical streamflow simulation: A comparison of model accuracy, interpretability, and uncertainty in seasonal watersheds. *Hydrology and Earth System Sciences*, *20*(7), 2611–2628. https://doi.org/10.5194/hess-20-2611-2016

Sui, Y. (2007). Potential Impacts of Controlled Drainage in Indiana Watersheds, (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (Accession Order No. 3307491). West Lafayette, IN: Purdue University.

Sugg, Z. (2007). Assessing U.S. farm drainage: Can GIS lead to better estimates of subsurface drainage extent. World Resources Institute, Washington, DC, 20002

Tao, Y., Gao, X., Hsu, K., Sorooshian, S., & Ihler, A. (2016). A deep neural network modeling framework to reduce bias in satellite precipitation products. *Journal of Hydrometeorology*, *17*(3), 931–945. https://doi.org/10.1175/JHM-D-15-0075.1

Tetzlaff, B., Kuhr, P., Vereecken, H., & Wendland, F. (2009). Aerial photograph-based delineation of artificially drained areas as a basis for water balance and phosphorus modelling in large river basins. *Physics and Chemistry of the Earth, Parts A/B/C*, *34*(8-9), 552–564. https://doi.org/10.1016/j.pce.2009.02.002

Tetzlaff, B., Kuhr, P., & Wendland, F. (2009). A new method for creating maps of artificially drained areas in large river basins based on aerial photographs and geodata. *Irrigation and Drainage: The journal of the International Commission on Irrigation and Drainage*, *58*(5), 569–585. https://doi.org/10.1002/ird.426

Tlapáková, L., Žaloudík, J., Kulhavý, Z., & Pelíšek, I. (2015). Use of remote sensing for identification and description of subsurface drainage system condition. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, *63*(5), 1587–1599. https://doi.org/10.11118/actaun201563051587

Todhunter, P. E. (2001). A hydroclimatological analysis of the Red River of the North snowmelt flood catastrophe of 1997. *JAWRA Journal of the American Water Resources Association*, *37*(5), 1263–1278. https://doi.org/10.1111/j.1752-1688.2001.tb03637.x

Tuttle, S. E., Cho, E., Restrepo, P. J., Jia, X., Vuyovich, C. M., Cosh, M. H., & Jacobs, J. M. (2017). Remote sensing of drivers of spring snowmelt flooding in the North Central US. In *Remote sensing of hydrological extremes* (pp. 21–45). Springer, Cham. https://doi.org/10.1007/978-3-319-43744-6_2

United States Department of Agriculture National Agricultural Statistics Service (2014). Special tabulation: County-level land drained by tile and ditches, 2012 Census Agric.

Varner B. L., Gress, T., Copenhaver, K., White, S. (2002). The effectiveness and economic feasibility of image based agricultural tile maps. Inst. of Tech., Champaign, IL. Final Report to NASA ESAD 2002.

Verma, A. K., Cooke, R. A., Wendte, L. (1996). Mapping subsurface drainage systems with color infrared aerial photographs. American Water Resource Association's 32nd Annual Conference and Symposium "GIS and Water Resources", September 22–26, Ft. Lauderdale, Florida.

Wang, Z., Lai, C., Chen, X., Yang, B., Zhao, S., & Bai, X. (2015). Flood hazard risk assessment model based on random forest. *Journal of Hydrology*, *527*, 1130–1141. https://doi.org/10.1016/j.jhydrol.2015.06.008

Wieczorek, M. (2004). Subsurface drains on agricultural land in the conterminous United States, 1992: National Resource Inventory Conservation Practice 606, raster digital data, Version 1.1, http://water.usgs.gov/lookup/getspatial?nri92_cp606

Williams, M. R., King, K. W., & Fausey, N. R. (2015). Contribution of tile drains to basin discharge and nitrogen export in a headwater agricultural watershed. *Agricultural Water Management*, *158*, 42–50. https://doi.org/10.1016/j.agwat.2015.04.009

Xie, Z., Phinn, S. R., Game, E. T., Pannell, D. J., Hobbs, R. J., Briggs, P. R., & McDonald-Madden, E. (2019). Using Landsat observations (1988–2017) and Google Earth Engine to detect vegetation cover changes in rangelands-A first step towards identifying degraded lands for conservation. *Remote Sensing of Environment*, *232*, 111317. https://doi.org/10.1016/j.rse.2019.111317

Yang, L., Jin, S., Danielson, P., Homer, C., Gass, L., Bender, S. M., et. al. (2018). A new generation of the United States National Land Cover Database: Requirements, research priorities, design, and implementation strategies. *ISPRS Journal of Photogrammetry and Remote Sensing*, *146*, 108–123.

Youngs, E. (1975). The effect of the depth of an impermeable barrier on water-table heights in drained homogeneous soils. *Journal of Hydrology*, *24*(3-4), 283–290. https://doi.org/10.1016/0022-1694(75)90086-4

Zhang, C., Walters, D., & Kovacs, J. M. (2014). Applications of low altitude remote sensing in agriculture upon farmers' requests—A case study in northeastern Ontario, Canada. *PloS One*, *9*(11). https://doi.org/10.1371/journal.pone.0112894