Identifying Causes of Irrigated Cropland Misclassification

Cella Schnabel
Srikrishnan Research Group @ Cornell BEE

Last updated: April 19th, 2024

Abstract

A reliable and complete global dataset of irrigated land would be ideal for understanding the impact of agricultural practices on sustainable water management – a crucial aspect of responding to shifts in land and water use due to climate change. While global agricultural land area has been largely well characterized, the distinction between irrigated and rainfed lands is often ignored. Several datasets to identify irrigated land in the United States have been developed using remote sensing, but it is unclear how well those methods might generalize to other regions. A critical question for expanding these datasets globally is what environmental conditions may cause misclassification. In this project, I conducted a thorough literature review to understand the limitations of irrigated cropland analysis, due primarily to distinguishing irrigation systems in coarser satellite imagery. My work follows the 2017 Landsat-based Irrigation Dataset for the United States (LANID-US) produced by Yanhua Xie et al.[1], in which they provide sample coordinates of ground-truth land identified as irrigated or rainfed in the conterminous United States (CONUS). I developed a random forest classifier in Google Earth Engine (GEE) to identify irrigated vs. rainfed cropland across CONUS by attaching the LANID coordinates to a Landsat pixel to train the classifier with geographical data. In preliminary analysis, the classified image is highly consistent with published irrigation datasets for the U.S., and the classification is most influenced by geographic location and Short-Wave Infrared Band 1 (SWIR 1). This is reasonable because SWIR 1 is sensitive to soil moisture, and geographic location plays a pivotal role in shaping environmental conditions like climate and soil properties that impact irrigation strategies. This classifier may generalize poorly to extremely moist or arid regions, but more feature manipulation is required. Quantifying the probability of misclassification and systematic biases can point to breakdowns in generalizing irrigated cropland classification, to provide a starting point for a more skilled global dataset of irrigated cropland.

Introduction

Global change is putting extraordinary pressure on croplands. While agricultural land across the world has been well distinguished from other land covers, the distinction between irrigated cropland and rainfed agricultural land—in this case, defined as cropland that relies solely on rainfall as a water source—is often ignored. Quantifying global irrigated cropland at high resolution can inform policy-making, minimize overuse or misuse of water, and contribute to overall environmental and economic sustainability, as well as climate resilience of agricultural systems [2].

Irrigation consumes roughly 90% of global freshwater resources [3] and, as of 2018, approximately 40% of the world's food is cultivated on artificially irrigated land. This number is expected to increase by nearly 20% by 2050 [4]. Due to the increasingly limited availability of freshwater, a major challenge in global agriculture and global water use is to mitigate the misuse of water while continuing to meet increasing demands for food exacerbated by climate change and population growth [5]. High resolution information about where irrigated land is located, the distribution of local water demand, and how this pattern is changing over regions and time works to inform management policies and climate smart agriculture solutions.

Several datasets to distinguish irrigated land in the United States using census statistics [6] or remote sensing data [1] or both [7] have been developed. In this project, we are particularly interested in studying patterns of remote sensing based misclassification, which could help identify which areas of the world might be particularly challenging to classify in this manner, as it is currently unclear how well these methods may generalize when ground-based data isn't available. A critical question for generalizing remote sensing based classifiers is which environmental conditions result in misclassification. In equal and opposite force, this begs the question: what land features are most important in irrigated land classifying?

This study uses Shapley values to break down random forest classification of Landsat-based images. Shapley values yield class-level feature importance to identify the most prominent features that determine the class value [8]. At this point, it's a binary classification: 1 for irrigated, 0 for rainfed. I used the Shapley additive explanations (SHAP) package in Python [9] to visualize feature importance in cropland classification. The intention is to discover which variables consistently stand out in particular climates.

Materials and Methods

The region of interest in this study is the conterminous United States (CONUS), with more than 58 million acres of irrigated cropland, according to The 2017 Census of Agriculture [10].

In GEE, I have examined a Landsat 8 image from USGS Landsat 8 Collection 2 Tier 1 TOA Reflectance that was manipulated with median composites and

sorting to optimize cloud coverage. I used the date range from June 15th to July 15th, 2017. This year was selected because the LANID-US map provides ground reference locations in the eastern CONUS until 2017 [1], while the month was selected as a useful portion of a given growing season. I cropped the Landsat image to mask out all land covers except for cropland, as shown in Figure 1; using band 15 of the 2020 North American Land Cover 30-meter dataset [11]. I selected this land cover dataset despite the time inconsistency as it uses Landsat at 30m resolution, instead of MODIS at 250m resolution.

The LANID-US data I used is a list of coordinates that include center pivot fields (irrigated) and stable rainfed fields [12].

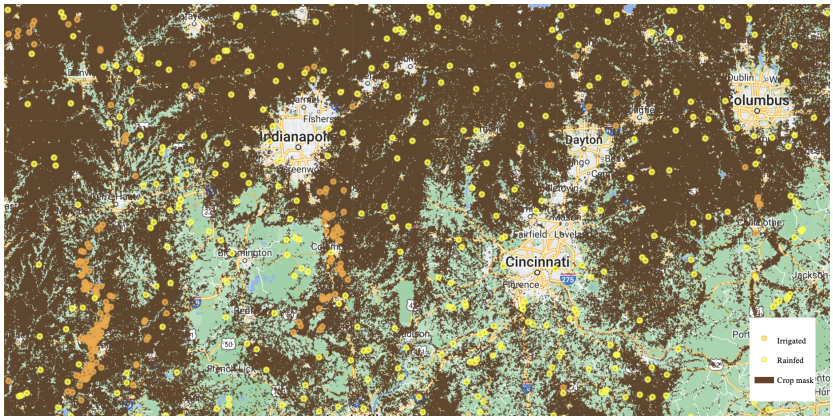

Figure 1: Random sample of materials used showing the crop mask, as well as an example of LANID-US points in Ohio, U.S.

Croplands can fluctuate immensely over a given growing season, but at any given time stamp, the heterogeneity of croplands is a major barrier to accurate classification. It's crucial to include time series data in a classifier. However, due to the brief nature of this project, so far I've only examined one median composite image.

I labeled Landsat pixels within a 30m radius of the given LANID coordinates a class: 1 for irrigated, 0 for rainfed. Using the data structure with Landsat bands B2-B7 and geographic coordinates as features, I split the 30m buffered sample coordinates randomly 80:20. I then used the training data to train a random forest classifier for pixel-level classification of CONUS.

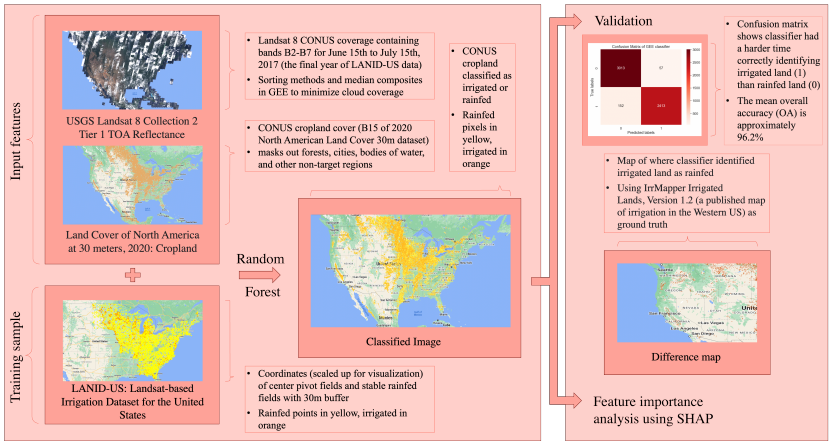

Figure 2: Flowchart of the method used.

Results

I classified CONUS cropland, as shown in Figure 3

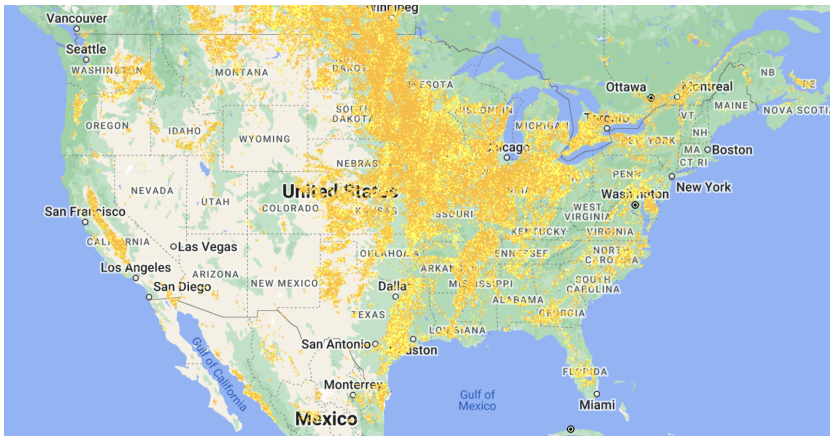

Figure 3: CONUS cropland classified as irrigated (orange) or rainfed (yellow)

In terms of validation, the random forest classifier had a harder time correctly identifying irrigated land than rainfed land, but the mean overall accuracy (OA) is approximately 96.2%. I compared this classified map to IrrMapper Irrigated Lands, Version 1.2, a published annual classification of irrigation status in the 11 Western United States made at Landsat 30 m resolution using the Random Forest algorithm [13]. For an accuracy assessment, I made a difference

map in Figure 4 where this classifier classified a pixel as rainfed and IrrMapper says irrigated.

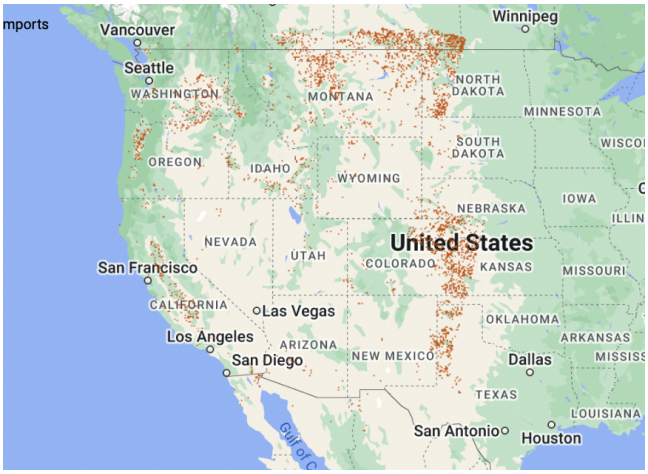

Figure 4: Difference map between this classification and IrrMapper

Using very crude, minimal inputs and still getting pretty high accuracy is inspiring for this strategy of land classification. There seem to be many avenues for adding data that may improve performance: adding more Landsat bands as features, implementing a time series, including more band metrics like Normalized Difference Water Index (NDWI), and Greenness Index (GI). LANID-US is an example of a very robust classification of CONUS. The objective of this project, however, is to understand the generalizability of this style of classification to characterize land types as easily classified or not.

I used SHAP to see what kinds of features were influencing this classification. The results reveal that certain Landsat bands played a more important role than others. In particular, geographic location, Short-Wave Infrared Band 1 (SWIR 1), and the blue sensor were the most influential in classification. See Figure 5 for a complete breakdown of feature importance.

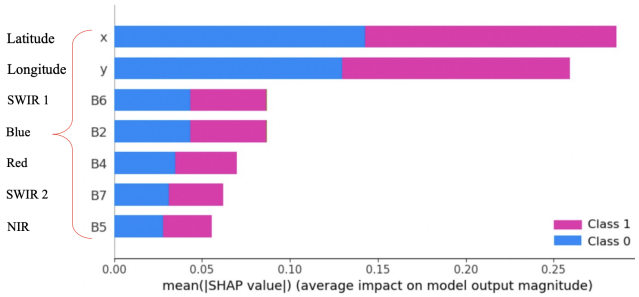

Figure 5: A bar plot indicating the relative importance of various features—the longer the bar, the larger influence the given feature had on classification for irrigated (Class 1) or rainfed (Class 0).

I also did this analysis with band metrics NDMI and NDVI, which are computed as

$$\text{NDMI} = \frac{\text{B5} - \text{B6}}{\text{B5} + \text{B6}}$$

$$\text{NVMI} = \frac{\text{B5} - \text{B4}}{\text{B5} + \text{B4}}$$

They appear to be minimally important in this analysis (see Figure 6), but have proven to be useful in more robust analysis [7], [1], [14].

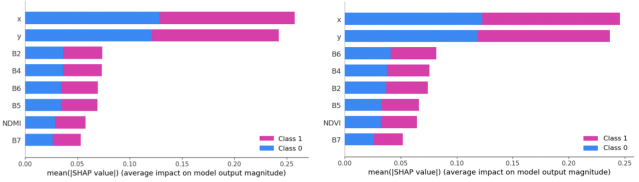

Figure 6: The same feature importance breakdown as Figure 5 but including rearranged band inputs NDMI (left) and NDVI (right) to see how vegetation and moisture were contributing to this classification. In this assessment, both are minimally influential.

Discussion

These results are reasonable. The classifier is most influenced by geographic location (x,y) likely because location is a strong indicator of climate conditions. In this case, certain U.S. regions have specific irrigation needs for high-value crops due to insufficient rainfall. This pattern may have been recognized by the classifier. A different, but related, explanation is that a clustering phenomenon is occurring, where the classifier is picking up on the probability of regional irrigation.

The most important bands—SWIR 1, blue sensor, SWIR 2, red sensor, and Near Infrared (NIR), suggest the potential sensitivity of classification to ground moisture and bodies of water, due to their specific spectral responses in these regions. Irrigated land is expected to have consistently higher soil moisture than strictly rainfed fields. Irrigated plants are also more likely to be greener for longer sections of the growing season.

I'm unsure about why the NDVI and NDMI metrics weren't more useful in classification, but I hypothesize that with a time series implemented, the reliance on spacial data will reduce and the usefulness of these metrics will emerge.

Conclusion

I look forward to the rest of this project, where I will work to make the classification more robust, and then move into misclassification analysis.

In early conclusions, I think this methodology has the potential to uncover the way irrigated land might be misclassified.

References

[1] Yanhua Xie, Tyler J. Lark, Jesslyn F. Brown, and Holly K. Gibbs. Mapping irrigated cropland extent across the conterminous united states at 30·m resolution using a semi-automatic training approach on google earth engine. *ISPRS Journal of Photogrammetry and Remote Sensing*, 155:136–149, 2019. ISSN 0924-2716. doi: https://doi.org/10.1016/j.isprsjprs.2019.07.005. URL `https://www.sciencedirect.com/science/article/pii/S0924271619301728`.

[2] Prasad S. Thenkabail, Munir A. Hanjra, Venkateswarlu Dheeravath, and Muralikrishna Gumma. A holistic view of global croplands and their water use for ensuring global food security in the 21st century through advanced remote sensing and non-remote sensing approaches. *Remote Sensing*, 2(1):211–261, 2010. ISSN 2072-4292. doi: 10.3390/rs2010211. URL `https://www.mdpi.com/2072-4292/2/1/211`.

[3] Igor A. Shiklomanov. Appraisal and assessment of world water resources. *Water International*, 25(1):11–32, 2000. doi: 10.1080/02508060008686794. URL `https://doi.org/10.1080/02508060008686794`.

[4] Hervé Plusquellec and Sanjiva Cooke. A retrospective of lending for irrigation - reflections on 70 years of bank experience (english). *Discussion Paper Washington, D.C.*, 2024. URL `http://documents.worldbank.org/curated/en/099757003282448162/IDU162ed994f17c69149bb1bcd21c19f282a40c2`.

[5] Robert I. McDonald, Pamela Green, Deborah Balk, Balazs M. Fekete, Carmen Revenga, Megan Todd, and Mark Montgomery. Urban growth, climate

change, and freshwater availability. *Proceedings of the National Academy of Sciences*, 108(15):6312–6317, 2011. doi: 10.1073/pnas.1011615108. URL https://www.pnas.org/doi/abs/10.1073/pnas.1011615108.

[6] Aaron Hrozencik, Grant Gardner, Nicholas Potter, and Steven Wallander. Irrigation organizations: Groundwater management. *Economic Research Service*, EB-34, 2023. doi: https://doi.org/10.32747/2023.7975545.ers.

[7] Jesslyn F. Brown and Md Shahriar Pervez. Merging remote sensing data and national agricultural statistics to model change in irrigated agriculture. *Agricultural Systems*, 127:28–40, 2014. ISSN 0308-521X. doi: https://doi.org/10.1016/j.agsy.2014.01.004. URL https://www.sciencedirect.com/science/article/pii/S0308521X14000134.

[8] Lloyd S Shapley et al. A value for n-person games. *Princeton University Press Princeton*, 1953.

[9] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Curran Associates, Inc.*, 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

[10] USDA. Irrigation & water use. *Economic Research Service*, 2019.

[11] Commission for Environmental Cooperation. North american land cover, 2020 (landsat, 30m). *North American Land Change Monitoring System*, 2019.

[12] Yanhua Xie and Tyler Lark. LANID-US: Landsat-based Irrigation Dataset for the United States, October 2021. URL https://doi.org/10.5281/zenodo.5548555.

[13] David Ketchum, Kelsey Jencso, Marco P. Maneta, Forrest Melton, Matthew O. Jones, and Justin Huntington. Irrmapper: A machine learning approach for high resolution mapping of irrigated agriculture across the western u.s. *Remote Sensing*, 12(14), 2020. ISSN 2072-4292. URL https://www.mdpi.com/2072-4292/12/14/2328.

[14] Tianfang Xu, Jillian M. Deines, Anthony D. Kendall, Bruno Basso, and David W. Hyndman. Addressing challenges for mapping irrigated fields in subhumid temperate regions by integrating remote sensing and hydroclimatic data. *Remote Sensing*, 11(3), 2019. ISSN 2072-4292. doi: 10.3390/rs11030370. URL https://www.mdpi.com/2072-4292/11/3/370.