Our response to the reviewer's comments is marked in blue. The context we cited from the revised clean manuscript with line numbers are marked in green.

Response to referee #1:

I have reviewed the paper, and my primary concern is whether the data make sense. National-scale crop type mapping is essentially an operational task, and conducting it at this scale poses significant challenges for a scientific team. I cannot recommend publication in ESSD before all my concerns are thoroughly addressed.

We thank the reviewer for the comments. Please see our response below.

1. Over the past two decades, numerous studies have been published on maize and soybean mapping in the United States, and CDL data (including all major crop types) have been available for more than 15 years. In contrast, the dataset presented here includes only maize and soybean. Given that cropland in the U.S. is characterized by large field sizes, the higher spatial resolution offered by this dataset does not appear to meaningfully improve national-level crop area statistics. It is unclear who the intended users of this dataset are. Do the authors envision USDA adopting it to replace the current CDL for NASS statistics? Are there specific states or counties that would use it for their own statistical work? For maize/soybean yield forecasting, which national projects would benefit from this dataset? For climate change analysis, does this 10 m maize/soybean map outperform the existing CDL products? From my perspective, the dataset would be more compelling if it were a 10 m crop type map covering all crops in the U.S., or if it provided 10 m major crop type maps for underrepresented regions such as Africa.

We appreciate the reviewer recognizing the value of CDL. Although CDL also has its own problems, much research relies on this critical dataset that it is difficult to overstate its value. However, we respectfully disagree with the reviewer that because CDL exists, national-scale crop mapping over the United States is a less useful topic to research, for several reasons. First, so far, users of CDL are enjoying the open-access nature of the dataset. However, it is a dataset produced by the United States government and therefore, it is subject to federal policy regulation. While we are not saying that access to this dataset would be limited in the future, we cannot ignore that future policy change could affect the data access. The most recent example is USDA's Common Land Units, which was once freely accessible but now restricted. Therefore, it is important to have independent datasets managed by academic institutions. Second, our map products are validated following the best practice guidelines using a stratified random sample and field visit. This is by far the most rigorous procedure for any national-scale crop mapping studies. Our validation method is also different from CDL, but the transparency nature of our validation allows other researchers to gauge the effort level and uncertainty level of implementing stratified random sampling at the country scale.

We have to emphasize that we are not claiming to replace the current CDL with our 10-m crop maps. Our objective is to introduce a robust workflow for crop mapping over large countries, and to quantify the benefits of 10-m maps compared to the 30-m products, specifically in mixed pixel reduction. Our methodology does not rely on any existing products for stratification, for training data collection, and for map validation. The comparison with CDL is used only to demonstrate how our 10-m maps improved the crop classification in fragmented fields, on field edges, or along the road networks, even in the context of industrialized agriculture for the US.

In the context of industrial agriculture in the US, the value of high resolution (finer than 30 m) crop maps is often ignored given that the majority of crop fields are in large sizes. However, as we demonstrated in this study, 10-m maps have many benefits in reducing mixed pixels for many regions. With less misclassification and mixed pixels, such maps can facilitate crop acreage estimation by reducing the uncertainties. Compared to 30-m maps, these maps can delineate smaller fields clearly (for example, the crop strips in the Appalachian regions in Pennsylvania, and the wetland/field mosaics in North Dakota, as we presented in Figure 13 in the manuscript, also the same figure Fig R.1 below). More accurate field boundaries can be derived from these maps to contribute to field boundary products, thus help investigation on field-level agricultural practices such as crop rotation.

Fig R.1 The 2022 aggregated 30-m maize and soybean map and CDL show mixed pixels in various landscapes. (a) wetland/agriculture mosaics in North Dakota; (b) center-pivot irrigated fields in Nebraska;

(c) strip fields in Pennsylvania. The coordinates of the center points are shown on the Google Earth imagery.

Producing a 10-m crop map for all crops in the US is extremely challenging. We are working towards that direction. We believe our established method can be extended to include more crops, although it is subject to more experiments particularly with regard to rare crops.

2. Line 122~142, Sentinel 2 Level 2A data is Surface Reflectance data, I wonder whether the author know what is "TOA reflectance" and what is "SR"! And I wonder whether the author understand the Sentinel-2 ARD pre-processing workflow and know what is the input.

“TOA reflectance” was neither used in our pre-processing workflow nor mentioned in the manuscript. We have to clarify that, we used L2A “**Bottom of the Atmosphere reflectance” (BOA)**, which is also known as **Surface Reflectance (SR),** as the input of the Sentinel-2 ARD pre-processing workflow, as Line 124 described: “We downloaded Sentinel-2A and -2B Level-2A Bottom of the Atmosphere reflectance (S2 L2A) images from Google Cloud”.

Reference: Copernicus Sentinel-2 Collection 1 MSI Level-2A (L2A).
https://sentinels.copernicus.eu/sentinel-data-access/sentinel-products/sentinel-2-data-products/collection-1-level-2a

3. Section 2.3: The samples are used to identify maize and soybean in the PSUs, and the resulting maize/soybean maps within the PSUs are then used to identify these crops at the national level. I find this approach questionable, particularly given that the study focuses on only two crop types. It is unclear how well this method would perform if all crop types in the U.S. were included. Moreover, any misclassification in the PSU-level mapping is likely to be amplified when scaling up to the national level. A potentially more robust solution would be to use high-resolution drone imagery in place of Sentinel data for maize/soybean mapping within the PSUs.

Based on the probability sample of field visit, our map product has >95% accuracy, which clearly demonstrates the creditability of our method.

The misclassification in the PSUs may or may not be propagated to the national-scale classification by the statistical nature of nonparametric machine learning algorithms such as random forest. Moreover, we implemented a very high standard of quality assurance for mapping at the PSU as well as national scales: 1) the PSU-level classification was conducted within each PSU individually using a decision tree classifier based on the windshield-based training data collected in the PSU; 2) conflict pixels, although being a small fraction, were excluded for the national-scale training; 3) SSU pixels with a buffer of 3 pixels were also excluded as they were solely used in validation, not in training; 4) a morphological erosion (dilation) of 5 pixels was applied to crop maps to avoid potential mixed crop (non-crop) pixels on field boundaries. These steps ensure that we can obtain high-quality training pixels for the national-scale mapping.

We produced these accurate PSU-level maps based on high-quality satellite imagery, robust data processing, representative field data for training, and domain knowledge from experienced experts. Higher-resolution drone imagery can be acquired over limited experimental sites (we do have multiple drones and have drone data over agriculture fields), but it is simply not practical to acquire drone data over a random sample over the entire United States.

4. Regarding the selection of training data, please demonstrate its representativeness by providing examples of typical vegetation index (VI) time series and corresponding meteorological data.

For the annual field surveys from 2019 to 2022, we implemented a stratified, two staged clustering sampling framework to select the target sites to collect field data (as described in Section 2.2). These four-year PSUs and SSUs are selected in a statistically rigorous manner to collect a tremendous amount of ground data across space and time representing various characteristics across the CONUS. Various factors are reflected from the ground data, including geographical locations, local crop diversities, weather conditions, crop phenological stages, irrigation, etc. These wide-range training data were then used to train generalized models for the national-scale classification.

As we presented in Figure 18 in Section 4.3 (also see the same figure Fig R.2 below), the NDVI for maize and soybean across space and time showed noticeable intra-annual and interannual variations in crop phenology, which were captured by the representative field data collected through our sampling framework. In the same year in 2022, maize and soybean in different states presented various NDVI profiles (a and b). In the same location, the NDVI profiles for maize and soybean also showed variations. Utilizing the multi-year field data for training generalized the model and produced national-scale crop maps with consistent accuracies.

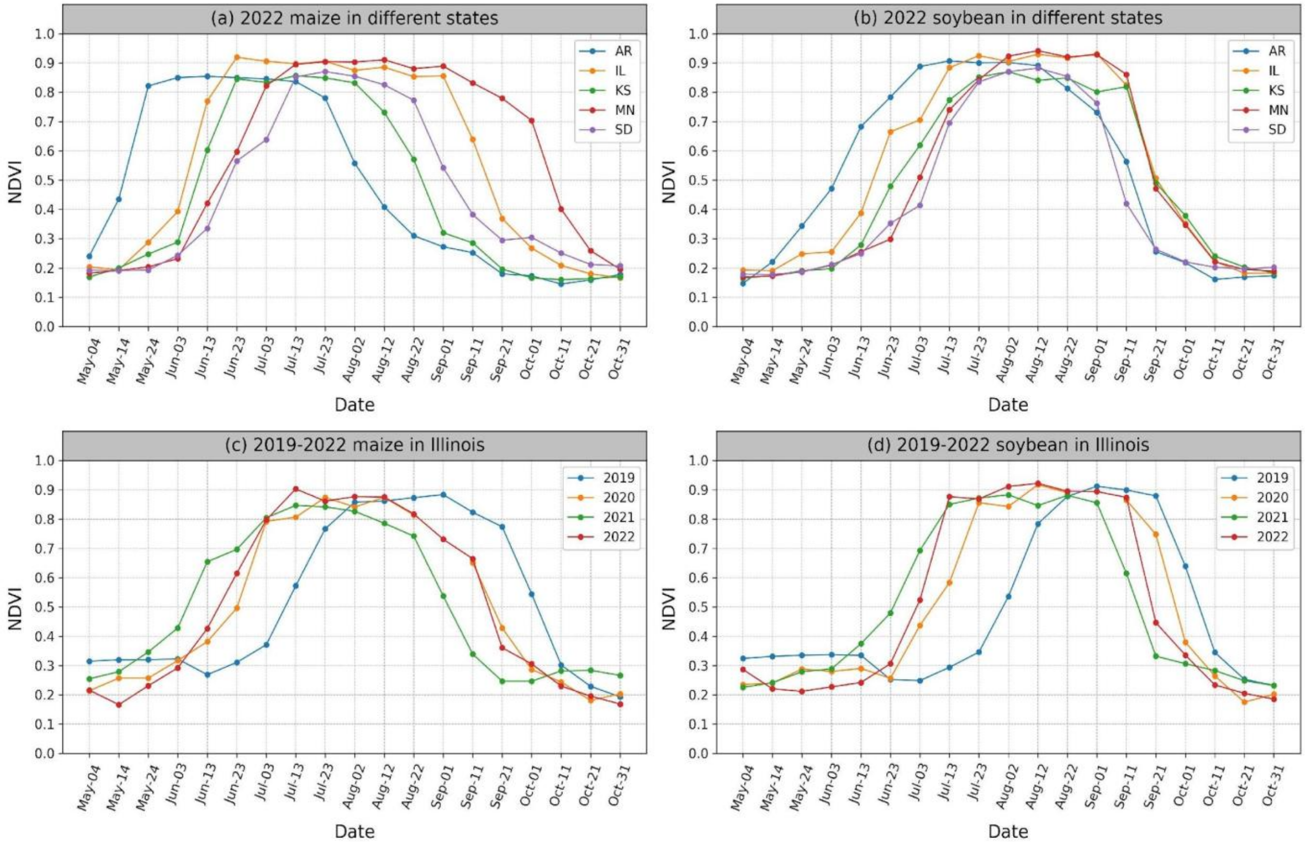

Fig R.2 NDVI time series for maize and soybean from representative sites. (a) 2022 maize NDVI in Arkansas (AR), Illinois (IL), Kansas (KS), Minnesota (MN), South Dakota (SD); (b) the same as (a) but for soybean; (c) 2019-2022 interannual NDVI variations for maize in Illinois; (d) the same as (c) but for soybean. The details about the sites are shown in Figure S4 and Table S6.

5. Validation data collection (Table S2): For each year, only 90 PSUs are used, and validation samples are collected solely from these PSUs. This sample size and distribution are not sufficiently representative. I recommend selecting at least 10 samples per PSU and including a minimum of 300 PSUs per year. The same PSUs could be used across multiple years, and the authors could consider contacting and collaborating with farmers to facilitate this process.

Pursuing field visits to a probability sample spanning a country as large and diverse as the US is indeed a costly endeavor for a research team (a safety issue as well). This is rare for remote sensing research by universities. Therefore, the financial and human cost for the sampling was substantial. Sample size planning was largely based on the tradeoff between the expected precision level and what was affordable based on the budget for the field sampling.

Cochran (and others) present sample size planning formulas, as for example Sec. 10.6 of Cochran (1977) in which optimal sampling fractions for two-stage cluster sampling (such as our sampling design) are presented. These sample size planning formulas require specification of costs (which are generally possible to approximate) and specification of variance components (in this case variance among PSUs and variance within PSUs). These variance components are generally speculative, and even more problematic as we would have needed those estimates for each of the three strata. Further, as is typical of most studies, several estimates are of interest. In our study, we required estimates of user's accuracy and producer's accuracy of the different crops, and area of the different crops. Sample size planning formulas focus on estimation of a single parameter and are difficult to apply to these multi-estimate settings.

Although we do not discount the value of sample size planning, the critical information is less about the precision one planned to obtain and more about what precision was actually achieved. Whether the sample size and allocation is adequate can be judged from the actual standard errors of the estimates reported. A larger sample size yields smaller standard errors. This would be a key consideration for agricultural ministries. From a research perspective, we report the details of our sampling design and the resulting standard errors in a transparent way. If one has a particular standard error target smaller than ours to achieve, increasing the sample size is clearly one of the possible improvements.

Finally, we re-iterate the important fact that standard errors depend on the absolute sample size, not the sample size relative to the population size (e.g., not the percent of the population sampled). We cited the following statements:

1) Cochran's (1977, p.24) Sampling Techniques book: "Provided that n/N [where N is the size of the population, n is the size of the sample] remains low, these factors are close unity [factors are (N-n)/N, the finite population correction], and "the size of the population as such has no direct effect on the standard error of the sample mean."

2) From Lohr (2019), Sampling, Design and Analysis, p.46: " Except in very small populations, precision is obtained through the absolute size of the sample, not the proportion of the population covered."

3) From Kish (1965), p.50: "Often laymen are surprised to hear that precision depends only on the size of the sample and not on the population size. But population size affects only the factor (1-n/N) and this can usually be ignored in designing the sample." [Please note that N is the population size and n is the sample size.]

Reference:

Cochran, W.G., 1977. Sampling techniques. John Wiley & Sons, New York, pp. 24.

Lohr, S.L., 2019. Sampling: design and analysis. Chapman and Hall/CRC, New York, pp. 46.

Kish, L. 1965. Survey sampling. John Wiley & Sons.

6. Figure 9, The comparison with CDL data does not make sense because the CDL result is the maize/soybean from all crop-type map, and the 10m map is just a maize/soybean map. Please conduct all crop type crop type mapping to make it comparable.

For crop mapping studies in the US, the 30-m CDL serves as the most commonly used reference. Comparing crop-specific maps with the corresponding crops derived from a subset of the CDL is not uncommon, although the majority of them are limited at the regional scale or only in selected counties (see the following references, for example). In this study, we compared our 10-m crop maps with the CDL at the national scale to demonstrate the advantage of 10-m maps in reducing mixed pixels over existing 30-m (coarser resolution) products.

Reference:

Johnson, D.M., 2019. Using the Landsat archive to map crop cover history across the United States. *Remote Sensing of Environment*, *232*, p.111286.

Wang, S., Di Tommaso, S., Deines, J.M. and Lobell, D.B., 2020. Mapping twenty years of corn and soybean across the US Midwest using the Landsat archive. *Scientific Data*, 7(1), p.307.

Cai, Y., Guan, K., Peng, J., Wang, S., Seifert, C., Wardlow, B. and Li, Z., 2018. A high-performance and in-season classification system of field-level crop types using time-series Landsat data and a machine learning approach. *Remote sensing of environment*, *210*, pp.35-47.

Johnson, D.M. and Mueller, R., 2021. Pre-and within-season crop type classification trained with archival land cover information. *Remote Sensing of Environment*, *264*, p.112576.

Skakun, S., Franch, B., Vermote, E., Roger, J.C., Becker-Reshef, I., Justice, C. and Kussul, N., 2017. Early season large-area winter crop mapping using MODIS NDVI data, growing degree days information and a Gaussian mixture model. *Remote Sensing of Environment*, *195*, pp.244-258.

You, N., Dong, J., Li, J., Huang, J. and Jin, Z., 2023. Rapid early-season maize mapping without crop labels. Remote Sensing of Environment, 290, p.113496.

7. The mixed-pixel analysis (Figures 12 and 13) does not make sense either, If more crop types were considered, the number of mixed maize pixels would likely increase. Therefore, I recommend first producing a 10 m resolution map for all crop types. Additionally, for this scale factor analysis, please include CDL data in the comparison to provide a more complete evaluation.

Mapping more crop types might have an impact on the accuracy of specific crop classes due to misclassifications that are confused with other crop types, but not necessarily increase the number of mixed pixels. Mixed pixels at a certain spatial resolution (for example, 30 m) are caused by the fact that heterogeneous land cover types coexist within the pixel, with varied fractional covers. Such cases, including the mixture of different crops, the mixture of crop/non-crop, may occur along the road network, field boundaries, and small, fragmented landscapes. Higher-resolution products, such as 10-m crop maps presented in our study, can lead to a reduction in mixed pixels owing to their ability to map various land covers in a finer spatial unit, as we have quantified in this study. We did include CDL in the comparison in Fig. 13 to demonstrate how the 10-m map helps reduce mixed pixels in Line 318 “in small, fragmented fields, on field edges, or along the road networks, where crops coexisted with other land cover (e.g., other crops, pasture, built-up, etc.)”.

We have conducted additional analysis by comparing our aggregated 30 m map with the 30 m CDL for the soybean and maize/corn class. We applied a 50% threshold to convert our aggregated 30 m map to binary soybean and maize maps and conducted a per-pixel comparison. In addition, we overlaid the CDL map on our 30 m fractional cover map, and computed the number and proportion of CDL soybean and corn pixels that are mixed pixels. We have added these analyses in Section 4.1 in Line 347-372.

As mentioned in our responses to your previous comments, mapping all crop types over the entire US is a challenging topic for future research. However, we anticipate that our methodology is applicable to generate 10-m maps with more crop types if resources are available.

Response to referee #2:

Review summary

This manuscript presents an approach to mapping maize and soybean at 10 m resolution using Sentinel-2 data and a two-step classification framework. While the topic is relevant and of potential interest to the remote sensing and agricultural monitoring communities, several aspects of the paper require clarification and refinement before it can be considered for publication. The study is, however, a refreshing example of work that combines a sound sampling design with expert-driven processing of remote sensing data, emphasizing data quality and the use of relatively simple, interpretable models. This stands in contrast to the growing tendency in the field to rely on deep learning models that ingest large volumes of raw data without adequate consideration of underlying data quality or sampling consistency. Overall, the study could become a valuable contribution if the authors strengthen the framing of their research objective, provide clear justifications for methodological choices, and align the discussion with recent developments in 10 m crop mapping.

We thank the reviewer for the comments. Our response to other comments is as below.

Major comments

- The scope of the paper is not clearly defined. The described classification methodology builds on well-established works by the same research group and therefore - by itself - lacks substantial novelty. Moreover, the claim that the Cropland Data Layer (CDL) is available only at 30 m resolution is no longer valid (see: nass.usda.gov/Research_and_Science/Cropland/SARS1a.php), as CDL products are now also distributed at 10 m. The title suggests that the study's primary aim is to produce accurate 10 m maize and soybean maps. However, a considerable portion of the analysis instead focuses on the differences between 10 m and 30 m classification maps. The authors should therefore clarify the main objective of the paper (L104-105 hints at this but it should get more focus). If the focus is on generating accurate 10 m maps, the added value of these maps compared to the now-available 10 m CDL must be articulated more convincingly. Alternatively, if the emphasis lies on analyzing the effect of spatial resolution on crop type mapping, this should be clearly stated and consistently reflected throughout the manuscript. Or if this workflow is merely a first step in potential upscaling to more challenging regions where there's less competing well-established products, it should be stated like this as well.

Thank you for the good point! When the study was designed and conducted, 10-m CDL were not available. We noticed the significant progress made by the USDA that the 2024 10-m CDL became available in February 2025. We also acknowledge that the 10-m CDL has much improved quality compared to the 30-m CDL 2024, particularly in delineating heterogeneous agricultural fields. If the USDA made 10-m CDL from 2019 to 2022, we could have made a fair comparison with 10-m CDL.

We have revised the objective statement to make it clear that generating 10-m maps is the first priority, and comparing the 10-m maps with 30-m map is the added value. Line 106-107: "The objective of this study is to develop annual 10-m crop maps with Sentinel-2 time series. We also quantify the benefits of 10-m maps compared to existing 30-m products."

- The manuscript uses data covering the period 2019–2022, but the rationale for selecting this specific time frame is not explained. It would be helpful if the authors clarified why these years were chosen, whether due to data availability or another reason. Providing this context would strengthen the transparency and reproducibility of the study design.

Thank you for the comments. By the time the research was conducted, the sample-based field surveys after 2022 were unavailable. Data preprocessing for Sentinel-2 imagery at the 10-m spatial resolution over the entire CONUS requires tremendous computational resources. Updating such a big dataset would involve a well-planned schedule. We presented our results over multiple years to illustrate the robustness of the method that it could be potentially implemented in operational settings.

- The main motivation for the two-step classification approach of first producing decision-tree-based PSU maps and then using those for training random forest algorithms for full-scale production is not clearly stated. Did the authors quantify the added value of this compared to a one-step approach where the training data is immediately used for training the random forest models? Also, why - in case of a two-step approach - do the inputs (all bands + ratios of any two bands, vs the temporal stats) and algorithms (decision tree vs random forest) have to be different? I did not find any justification for that and it is quite confusing for the reader why these classification steps are so different.

Thank you for your thoughtful comments.

The objective of this two-step classification is to first produce highly accurate PSU maps utilizing the in-season training points from the ground survey, then using these PSU maps as training labels for the national-scale classification. This augmentation approach provides much more training data compared to training a national-scale model directly on the training points.

As shown in the figure Fig R.3 below (also the same Figure 7 in the manuscript), we conduct the annual field surveys during the peak growing season in the summer. The in-season PSU-level maps are then produced by experts with domain knowledge using the "windshield survey" training points after applying a quality assurance process. A PSU-specific decision tree classifier is trained based on the in-season Sentinle-2 time series.

We have revised in the manuscript Line 230-235: "We processed all available Sentinel-2 data over the PSUs from May 1 to October 31 to produce in-season PSU-level maize and soybean maps. We trained two decision tree classifiers separately for maize and soybean classification by using all the bands and normalized ratios of any two bands, as well as the "windshield survey" points as training (Figure 7b). Applying the trained models to time-series images, we created a binary maize/non-maize map and a binary soybean/non-soybean map at 10-m resolution for each PSU (Figure 7c). These in-season PSU maps were then pooled as training labels for national-scale wall-to-wall mapping."

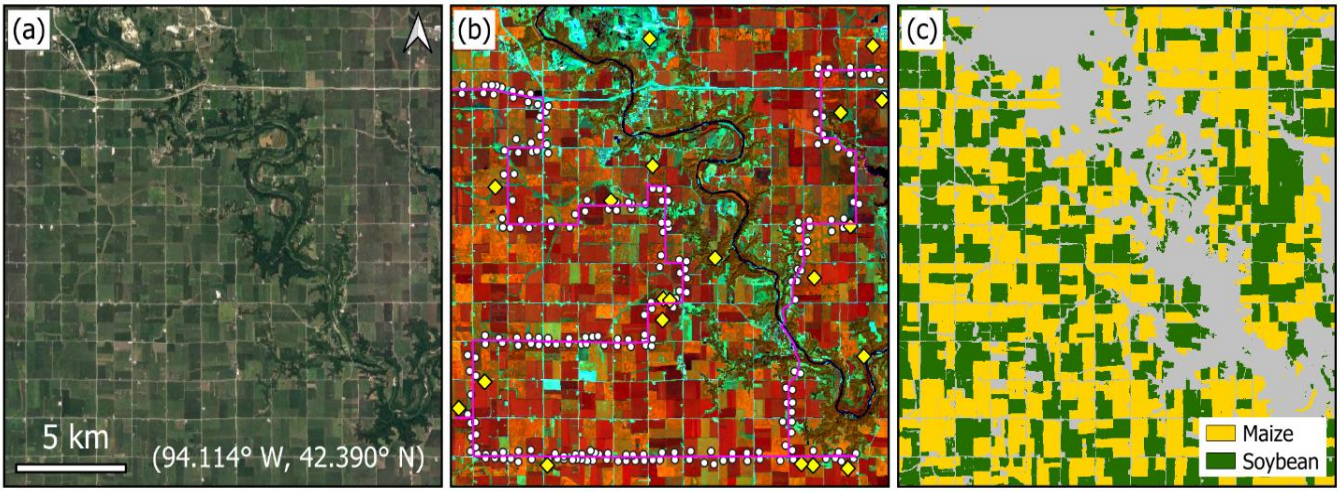

Fig R.3 An example of primary sampling unit (PSU) block-level crop mapping using field data. (a) a representative sample block in Illinois with center coordinates shown on the Google Earth imagery. (b) field data collection in the PSU. The secondary sampling units (SSUs) of pixels are shown as yellow diamonds. The "windshield survey" points are shown as white dots. The driving routes are shown in pink tracks. (c) PSU-level crop maps.

At the PSU level, all bands and all ratios of any two bands are exhaustive inputs customized for local context. Swapping this input with temporal stats would likely generate comparable results but not necessarily better. On the other hand, the exhaustive inputs of all bands and all ratios do not generalize well across large areas (e.g., a country) due to phenological variations. Instead, temporal stats have the advantage to normalize phonologies variations over different regions and years, enabling large-area land cover mapping as illustrated in previous studies (Hansen et al., 2013; Potapov et al., 2021; Song et al., 2021). Therefore, exhaustive input is better suited for local-scale PSU level and temporal stats are better suited for national scale mapping. Random forest is a theoretically advanced algorithm than decision tree, although consuming more computational resources. We chose random forest over the national scale to better handle the complexity over large areas. Decision tree is the simpler approach and works very well at local context. We hope this clarifies the data and algorithm choices.

Reference cited:

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., Thau, D., Stehman, S. V., Goetz, S. J., Loveland, T. R., Kommareddy, A., Egorov, A., Chini, L., Justice, C. O., and Townshend, J. R.: High-resolution global maps of 21st-century forest cover change, Science, 342, 850–3, https://doi.org/10.1126/science.1244693, 2013.

Potapov, P., Turubanova, S., Hansen, M. C., Tyukavina, A., Zalles, V., Khan, A., Song, X.-P., Pickens, A., Shen, Q., and Cortez, J.: Global maps of cropland extent and change show accelerated cropland expansion in the twenty-first century, Nature Food, 3, 19–28, https://doi.org/10.1038/s43016-021-00429-z, 2021a.

Song, X.-P., Hansen, M. C., Potapov, P., Adusei, B., Pickering, J., Adami, M., Lima, A., Zalles, V., Stehman, S. V., Di Bella, C. M., Conde, M. C., Copati, E. J., Fernandes, L. B., Hernandez-Serna, A., Jantz, S. M., Pickens, A. H., Turubanova, S., and Tyukavina, A.: Massive soybean expansion in South America since 2000 and implications for conservation, Nature Sustainability, 4, 784–792, https://doi.org/10.1038/s41893-021-00729-z, 2021b.

Specific comments

L29: abstract should be self-explanatory to the extent possible. This line is very difficult to understand without context ("pixels that can be reduced at field, regional, and national levels"). I suggest to rephrase to make the message clear.

We have revised the manuscript in Line 28-29: “we aggregated the 10-m maps to 30-m spatial resolution and quantified the number of mixed pixels that could be reduced by improving the mapping from 30 m to 10 m”

L95: see major comment: as of the 2024 version, CDL is now also offered at 10m resolution (source: https://www.nass.usda.gov/Research_and_Science/Cropland/SARS1a.php). Please add a note on this and the authors could also consider taking this up in their outlook based on the 10m vs 30m analysis results.

Thank you for this comment again. We revised as Line 93-96: “Recently, the 2024 CDL has been successfully released with the spatial resolution increased from 30 m to 10 m (https://www.nass.usda.gov/Research_and_Science/Cropland/SARS1a.php, accessed 26 December 2025). However, the previous years of CDL are at 30-m resolution, and have inconsistent accuracies depending on the location, and inaccurate classifications are observed in sparse or complex agricultural regions (Larsen et al., 2015).”

L98: note the recent release of European continental 10m crop type maps which should be referenced here as well (source: https://sdi.eea.europa.eu/catalogue/srv/api/records/9db29b07-5968-4ce0-8351-1e356b3d7d47?language=all)

We have added the reference in Line 100-103: “In Europe, recent 10-m crop mapping efforts include the Crop Map of England (CROME) (CROME, 2024), the parcel-level crop maps in the Netherlands (ESA, 2024), the crop maps produced by the Sentinel-2 for Agriculture (Sen2-Agri) (Defourny et al., 2019; Inglada et al., 2015), and the more recent High Resolution Layer Crop Types (CTY) (EU, 2024).”

L101: WorldCereal released global maps while the sentence reads as if these are only European maps.

Thanks for the comment. We have revised in Line 99-100: “Global 10-m crop mapping efforts are rare, although the WorldCereal provides an example (Van Tricht et al., 2023).”

L131-132: Two cloud masks are combined which results in very agressive masking. At least one source (Sen2Cor) is known to suffer from false positives over bright surfaces. Can the authors comment on the impact of potential numerous false positives in their masking procedure?

Thanks for the question. We agree that the Sen2Cor can have false positive cloud detection over bright target such as urban area, bare ground, snow and ice, etc. As we are targeting crop field in this study, such commission errors of cloud detection are likely to have less impacts on agricultural landscapes. As illustrated in the Fig R.4 below, the scene classification showed false positive cloud detection with high/medium cloud probability over bright bare ground, whereas such commission errors did not occur over the agricultural fields (in the lower part of the panel). In addition, we utilized a temporal linear interpolation method for gap-filling when missing data occurs. We calculated the 10-day median composites and derived the temporal metrics for the wall-to-wall national-scale classification. Combining these approaches can mitigate the impacts of cloud contamination on the training inputs for classifiers, even though a relative aggressive cloud might false positively identify clear-sky pixels as cloudy. In comparison, a less conservative cloud masking might result more cloud remnants that have more impacts on the data consistency.

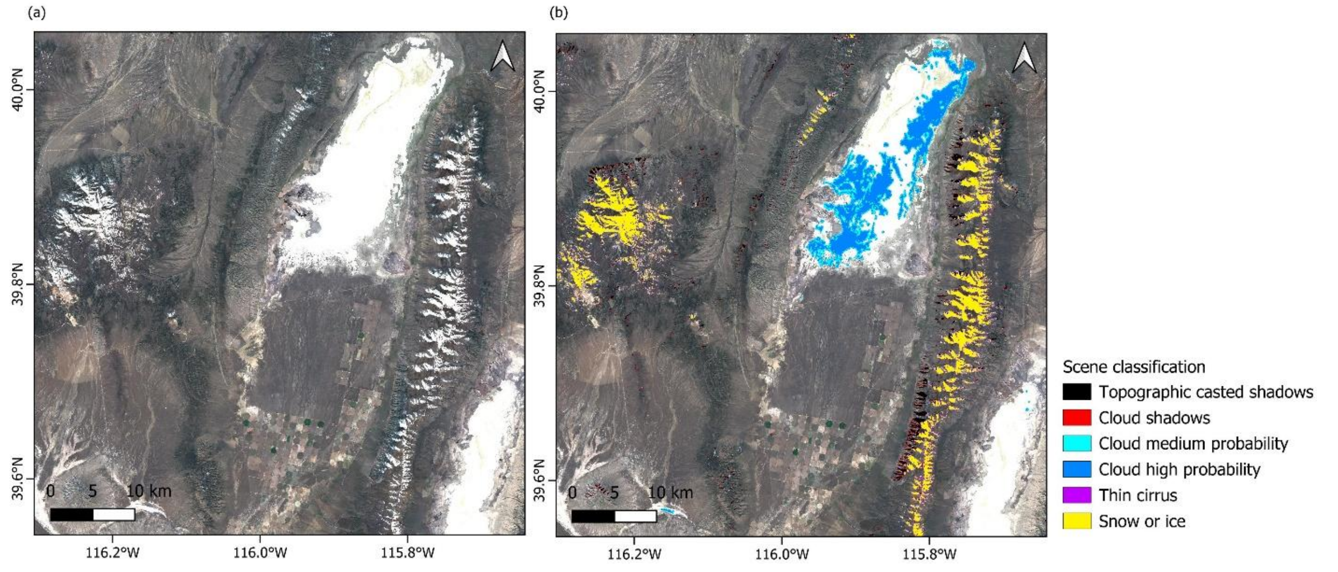

Fig R.4 Sen2cor scene classification over agricultural landscape and bright targets. (a) Sentinel-2 true composite at 10-m resolution on April 3, 2022; (b) Sen2Cor scene classification at 20-m resolution.

L155-L156: in case no interpolation was done, how was the missing value in the time series treated?

For cases in which cloud-free observations are unavailable for two months, we did not conduct interpolation. The missing values are treated as no data in the time series.

L159-L160: What is the motivation of a reprojection to lat/lon, also for the final output products, especially since that leads to variable pixel sizes? The authors should more carefully clarify their motivation for this approach.

Lat/lon projection is clearly inconvenient for pixel counting for area estimation, but pixel counting is not good practice for area estimation anyway. Although this study is focused on the United States, we aspire to do global-scale crop mapping work. We have done many case studies around the world, including China, South America and here US. Lat/lon projection is chosen such that it will make global work easier. We release all map products in lat/lon projection as it is more consistent with other geospatial datasets (e.g., weather or climate data) that some users may be interested in. This lat/lon projection is also consistent with our global Landsat ARD data.

L200-L201: which previous year's crop map? Your own final map of the preceding year? If so, how did you produce the first year?

Thanks for these questions. This is a fine point. We have been working on Landsat-based crop mapping over the US since 2015 and maintain Landsat-based crop map update at an annual frequency (unpublished work). We used our previous year's Landsat-based crop maps as stratification. This general approach has been reported in detail in our recent papers for the United States (Song et al. 2017), South America (Song et al. 2021), and China (Li et al. 2023), including statistical formulas for stratification, sample size determination and area estimation. Therefore, we don't mean to distract readers from these statistical details in this manuscript as our focus is on Sentinel 2 data processing and 10 m mapping. As you can see from Figure 5 (a), the stratification shows the spatial patterns of varying soy/corn intensity very well. We can release the stratification map as additional data, if the reviewer wants to check.

Reference:

Song, X.-P., Potapov, P. V., Krylov, A., King, L., Di Bella, C. M., Hudson, A., Khan, A., Adusei, B., Stehman, S. V., and Hansen, M. C.: National-scale soybean mapping and area estimation in the United States using medium resolution satellite imagery and field survey, Remote Sensing of Environment, 190, 383–395, https://doi.org/10.1016/j.rse.2017.01.008, 2017.

Song, X.-P., Hansen, M. C., Potapov, P., Adusei, B., Pickering, J., Adami, M., Lima, A., Zalles, V., Stehman, S. V., Di Bella, C. M., Conde, M. C., Copati, E. J., Fernandes, L. B., Hernandez-Serna, A., Jantz, S. M., Pickens, A. H., Turubanova, S., and Tyukavina, A.: Massive soybean expansion in South America since 2000 and implications for conservation, Nature Sustainability, 4, 784–792, https://doi.org/10.1038/s41893-021-00729-z, 2021b.

Li, H., Song, X.-P., Hansen, M. C., Becker-Reshef, I., Adusei, B., Pickering, J., Wang, L., Wang, L., Lin, Z., Zalles, V., Potapov, P., Stehman, S. V., and Justice, C.: Development of a 10-m resolution maize and soybean map over China: Matching satellite-based crop classification with sample-based area estimation, Remote Sensing of Environment, 294, https://doi.org/10.1016/j.rse.2023.113623, 2023.

L217-L218: it would be helpful if a table was added with total sample sizes per label (both SSU and wind shield survey numbers)

Thanks for the comment. We have added Table S3 and Table S4 in the supplementary materials document (also see the same tables below). We also have revised the manuscript in Line 219-220: "For each year from 2019 to 2022, we selected a separate stratified two-stage cluster sample following the general sampling framework and collected in-season ground data (Figure 6, Table S2-S4)."

Table S1 Number of SSUs and crop types collected during the field surveys from 2019 to 2022. “Others” includes trees, grass, hay, vegetables, orchards, bare ground, buildings, roads, etc.

Year	Crop type	Count	Year	Crop type	Count	Year	Crop type	Count	Year	Crop type	Count
2019	Maize	312	2020	Maize	307	2021	Soybean	386	2022	Soybean	352
	Soybean	273		Soybean	304		Maize	351		Maize	323
	Wheat	44		Wheat	38		Wheat	59		Alfalfa	27
	Cotton	28		Cotton	15		Sorghum	23		Wheat	19
	Sorghum	23		Alfalfa	13		Cotton	20		Cotton	13
	Rice	16		Sorghum	10		Alfalfa	13		Sorghum	10
	Alfalfa	7		Rice	6		Rice	12		Sugarcane	10
	Dry bean	6		Sugar beet	4		Sugar beet	10		Rice	8
	Canola	4		Millet	3		Canola	6		Barley	7
	Potato	4		Potato	3		Potato	5		Canola	5
	Radish	3		Dry bean	2		Peanut	4		Sugar beet	4
	Oats	3		Peanut	2		Sunflower	3		Oats	3
	Sunflower	2		Tobacco	1		Tobacco	2		Pumpkin	1
	Peanut	2		Others	1092		Oats	2		Sunflower	1
	Barley	1					Dry beans	2		Peanuts	1
	Others	1072					Sugarcane	1		Others	1056
							Others	1101			

Table S2 Number of training points and crop types collected during the field survey from 2019 to 2022. “Others” includes trees, grass, hay, vegetables, orchards, bare ground, buildings, roads, etc.

Crop type	Count	Crop type	Count	Crop type	Count
Maize	25197	Peanut	233	Radish	63
Soybean	24637	Canola	211	Millet	59
Wheat	2608	Barley	199	Pumpkin	13
Alfalfa	2156	Sunflower	168	Tomato	8
Cotton	1625	Tobacco	158	Dry beans	2
Sorghum	909	Peanuts	126	Sugar beet	1
Rice	600	Oats	68	Others	17236
Sugarcane	504	Potato	63		

L224: cfr main comment on the two-step approach

Thank you for the comment again. As we responded in the earlier comments. We utilize the training points from the ground survey to produce highly accurate in-season PSU maps. We then calculated the annual temporal metrics to relatively normalize the inter-annual crop phenology variations to ensure spatiotemporal consistency for annual mapping.

L229-L230: cfr main comment on the two-step approach

Thank you for the comment. Please see the response above.

L243: what's the rationale behind these seemingly arbitrary numbers? (0.2 vs 0.8 %)

Thank you for the question. These numbers are selected based on multiple tests to ensure a large-amount of positive (maize and soybean pixels) and negative (non-maize and non-soybean pixels) training data while balancing the consumption of computational resources and time.

L252: How can the authors be confident that the probability outputs of the independently trained maize vs soybean models can be used like this? It's a common observation that some models can be overly confident while others are not. It's perfectly possible that either the maize or soybean models are - in case of doubt - consistently outputting higher probabilities than the other model. In competition, that model would always win in this kind of aggregation rule. Could the authors comment on this potential pitfall promoting one model over the other purely based on the level of confidence?

Thank you for this great comment. It is tricky to treat the probability value as the confidence level (or the opposite of uncertainty) when the two classes are derived from two independent models.

We were very careful to implement this approach here, while recognizing that this approach cannot be generalized for other land cover types or in other study regions without careful examination of the data. The main reasons why this approach can be done in the US context are (1) at the national scale, soybean and maize are comparable class in terms of planted area (i.e. no dominant class and no rare class), (2) perhaps more importantly, soybean and maize are often planted in neighboring fields and this spatial co-location pattern of soybean and maize accounts for most of the planted areas of the two crops in the US (e.g. all Midwest states), and (3) on the technical level, when we train the soybean model, the major competing crop is maize, and when we train the maize model, the major competing crop is soybean. In other words, although the soybean model and the maize model are independent, the other crop is the dominant negative class in each model. And therefore, the two models are somewhat inversely related. As a result, when we put the soybean probability layer and corn probability layer side by side, this pattern becomes very obvious. (see Fig R.5 below).

We aggregated the per-pixel class probability layers from RF-Maize and RF-Soybean by selecting the highest probability. In Random Forest, the final output probability is the mean value across all the trees in the forest. A higher final probability for a crop class (maize vs. soybean) for a pixel indicated higher confidence across all the trees. And we measured the prediction uncertainty by calculating the standard error of the probability values across all the trees. In addition to the probability layer, we also calculated the standard deviation of the probability outputs of all the trees. When the probability values are the same for RF-Maize and RF-Soybean (accounting for ~2.7% of total pixels), the value with smaller standard deviation is selected.

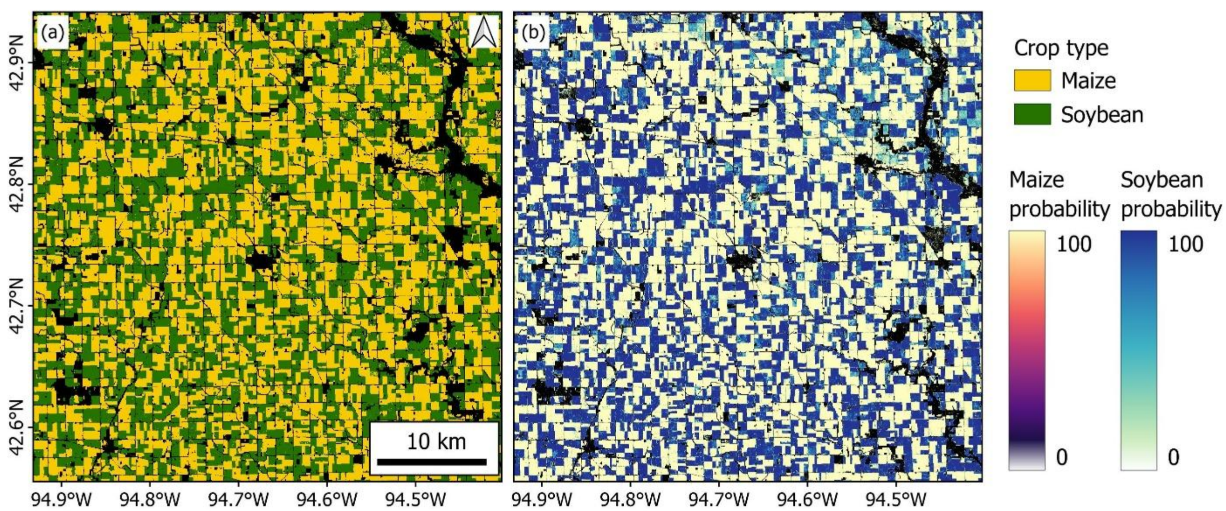

Fig R.5 Per-pixel classification aggregation from probability layers. (a) aggregated maize and soybean map; (b) probability layers derived from RF-Maize and RF-Soybean.

We have revised the manuscript in Line 255-260 to clarify: "We aggregated the per-pixel class probability layers from RF-Maize and RF-Soybean by selecting the highest probability (maize vs. soybean) and derived the aggregated probability layer and corresponding crop mask layer (see Figure S3 for an example). We then applied a 5 × 5 pixel kernel opening followed by a 10 × 10 pixel kernel closing, to eliminate scattered pixels and fill holes within large homogeneous fields. The kernel sizes were selected based on tests and visual assessments to balance noise removal while preserving fine details. We generated the final maize and soybean map using the aggregated probability layer following the area-matching approach reported by Song et al (2017), Song et al. (2021b) and Li et al. (2023)."

L305: Sect. 4.1 focuses on the difference between 10m and 30m based on the same original 10m map. This is insightful to the community. Focusing on the map outputs themselves, it would be helpful though if a comparison was made between the 10m map, the aggregated 30m map and the corresponding 30m CDL map. This demonstrates the resolution effect, but also compares map quality between the map produced here and the CDL map.

Thanks for the comments. The goal for this aggregation is to quantify the numbers of mixed pixels that 10-m maps can reduce, and analyze the spatial patterns at the national and regional scales. This part has been done.

We have conducted additional analysis by comparing our aggregated 30 m map with the 30 m CDL for the soybean and maize/corn class. We applied a 50% threshold to convert our aggregated 30 m map to binary soybean and maize maps and conducted a per-pixel comparison. In addition, we overlaid the CDL map on our 30 m fractional cover map, and computed the number and proportion of CDL soybean and corn pixels that are mixed pixels.

We have added these analyses in Section 4.1 in Line 347-372:

We conducted additional analysis by comparing our aggregated 30-m map with the 30-m CDL for the maize and soybean classes. We applied a 50% threshold to convert our aggregated 30-m map to binary maize and soybean maps and conducted a per-pixel comparison. We then aggregated the per-pixel results to the county level based on the ratio of the difference of maize or soybean pixels between our map and CDL divided by maize or soybean pixels in CDL. Positive values indicate more maize or soybean pixels in our map, whereas negative values indicate more pixels in CDL (Figure 15). In several regions, our map presented more maize pixels, such as in northern Texas and southern Louisiana (Figure 15a), and more soybean pixels, such as in western North Dakota and Georgia (Figure 15c). In general, CDL reported more maize and soybean pixels in most counties compared to our aggregated 30-m map (Figure 15b, d), which is likely due to the exclusion of 10-m maize and soybean pixels below the 50% threshold during aggregation.

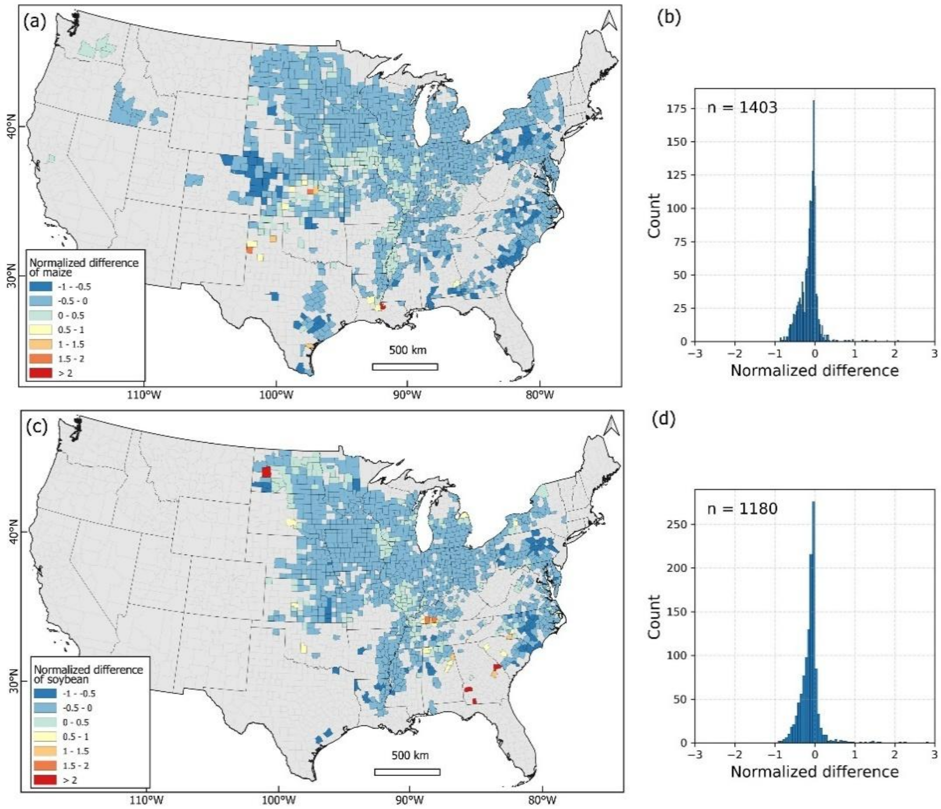

Figure 15: Comparison between our aggregated 30-m map with the 2022 30-m CDL at the county level. (a) normalized difference of maize pixel; (b) histogram of the normalized difference of maize pixel; (c-d) the same as (a-b) but for soybean. Counties accounting for 99.9% coverage of the national maize and soybean cultivation derived from the 2022 NASS statistics are shown.

To estimate the number of mixed pixels that might be reduced by improving the CDL's spatial resolution from 30 m to 10 m, we overlaid the 2022 30-m CDL on our 30-m fractional cover map, and computed the number and proportion of CDL maize and soybean pixels that were mixed pixels at the county level (Figure 16). The mixed pixel derived from CDL showed similar spatial distribution patterns to results derived from the aggregated 30-m map (see Figure 14 above). The median percentage of mixed pixels across all counties for both maize and soybean was 8% (Figure 16b, d), which is close to the values of 8% (9% for soybean) derived from our aggregated 30-m map (see Figure 14 above). These consistent mixed pixel estimates from our map and the CDL indicate substantial benefits of 10-m crop maps in reducing mixed pixels over existing 30-m products, especially for regions outside of Midwest, as illustrated in our analyses.

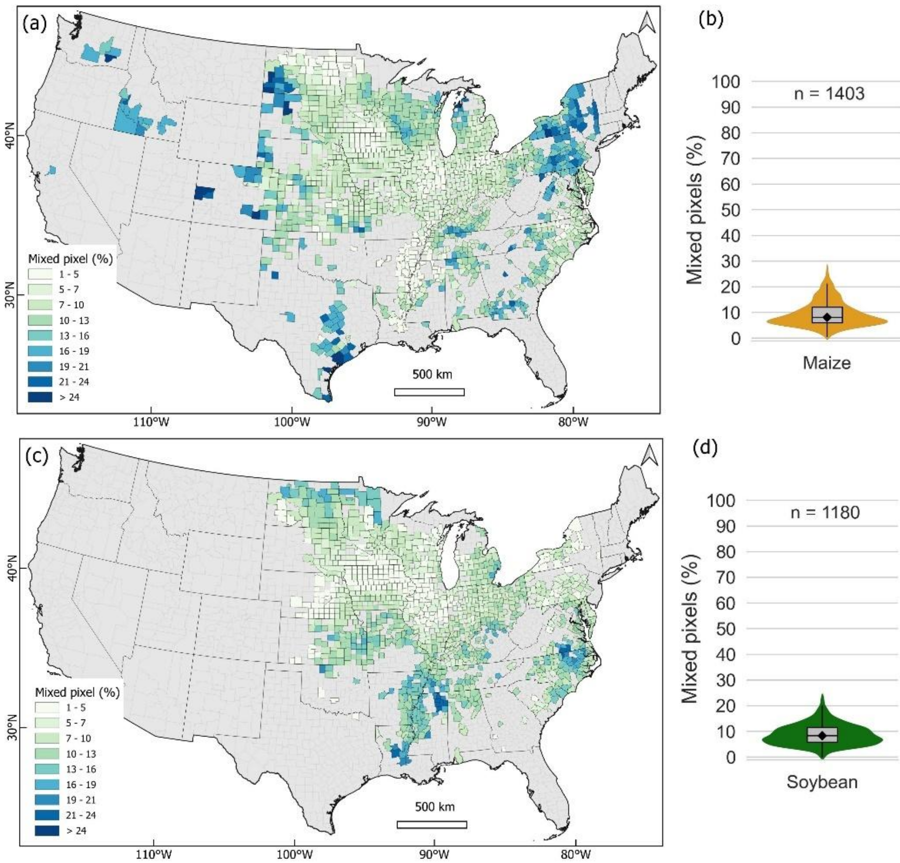

Figure 16: The percentages of 30-m mixed maize and soybean pixels at the county level by comparing our aggregated 30-m map and the 2022 30-m CDL. (a) the spatial distribution of mixed maize pixels; (b) the statistical distribution of mixed maize pixels; (c-d) the same as (a-b) but for soybean. Counties accounting for 99.9% coverage of the national maize and soybean cultivation derived from the 2022 NASS statistics are shown.

L393: add the usage of the 0 and 255 values in the data as well

Thank you for the suggestion. We have revised in Line 424: “The values 0, 1, 2, 255 represent other, maize and soybean, and no data, respectively.”

L394: as is always the case for this group, the field data is unfortunately not made public. The claim "Data used in this study are openly accessible online" is therefore not fully correct. Please clarify that only the external data being used is openly accessible.

Thanks for the suggestion. We have revised in Line 425: “External data used in this study are openly accessible online”

We understand your concern, but sharing raw field data with precise geolocation info and crop info is not as easy as one might hope. It is especially challenging in the context of the United States. In our many conversations with US farmers, they are often very concerned when data with their crop info are made public while people from geopolitically sensitive regions could use that data. For now, we try our best to make high quality maps and release those map products. These maps are evaluated with the highest standard of accuracy assessment. Other people may make better maps with access to our field data, but the additional utility to data users may be limited.

L405-L406: given the already published 10m CDL (see source in major comments section), the authors may reconsider the use of the word "first" here.

We have removed the “first” from the text.

Data

- The TIFF files provided are not cloud-optimized. I would recommend offering these kind of huge files as COG because they greatly simplify the visualization in GIS applications.

Thank you for the suggestion. We have updated the files as COG associated with metadata on the GLAD website (https://glad.umd.edu/dataset/mapping-crops-10-m-resolution-united-states) and in the Figshare repository (https://doi.org/10.6084/m9.figshare.28934993.v2).

- Some extra metadata in the TIFF files such as the legend would be useful for users.

Thank you for the suggestion. We have updated the files as COG associated with metadata on the GLAD website (https://glad.umd.edu/dataset/mapping-crops-10-m-resolution-united-states) and in the Figshare repository (https://doi.org/10.6084/m9.figshare.28934993.v2).

Supplementary material

- Table S1: does this table miss a horizontal line?

Thank you for your careful review. We have revised the Table S1.

Response to referee #3:

This is an excellent paper that introduces an extremely valuable data product: the 2019-2022 annual 10-m resolution maps of maize and soybean for the Contiguous United States (CONUS). Given that the journal *Earth System Science Data* (ESSD) aims to publish high-quality, thoroughly validated, and well-described datasets, this paper excels in its methodological rigor, validation thoroughness, and data accessibility, making it an excellent fit for the journal.

Overall, this is impressive work, and the data product is of great significance for agricultural monitoring, food security assessment, and ecological modeling. The paper is clearly written and well-structured. While I hold this work in high regard, I also propose the following modification suggestions, primarily focused on clarifying key technical details and enhancing the analytical depth, which I believe will further improve the quality of the paper and the usability of the data product.

We thank the reviewer for the positive comments. Our response to other comments is as below.

1. **Major Technical Question**

The authors trained two binary classifiers separately for maize (RF-Maize) and soybean (RF-Soybean). In practice, a pixel might be predicted as "present" by both models (with high probability), or predicted as "absent" by both. The authors have not explained how this "maize vs. soybean" conflict or overlap is handled.

The "area-matching approach" is typically used to guide post-classification processing (e.g., adjusting classification thresholds) to match the total area with a reference value (like official statistics). In this study, was this method applied *before* or *after* the morphological operations (opening and closing)? How was it combined with the two independent probability layers (maize, soybean)?

We aggregated the per-pixel class probability layers from RF-Maize and RF-Soybean by selecting the highest probability. In Random Forest, the final output probability is the mean value across all the trees in the forest. A higher final probability for a crop class (maize vs. soybean) for a pixel indicated higher confidence across all the trees. And we measured the prediction uncertainty by calculated the standard error of the probability values across all the trees. In addition to the probability layer, we also calculated the standard deviation of the probability outputs of all the trees. When the probability values are the same for RF-Maize and RF-Soybean (accounting for ~2.7% of total pixels), the value with smaller standard deviation is selected.

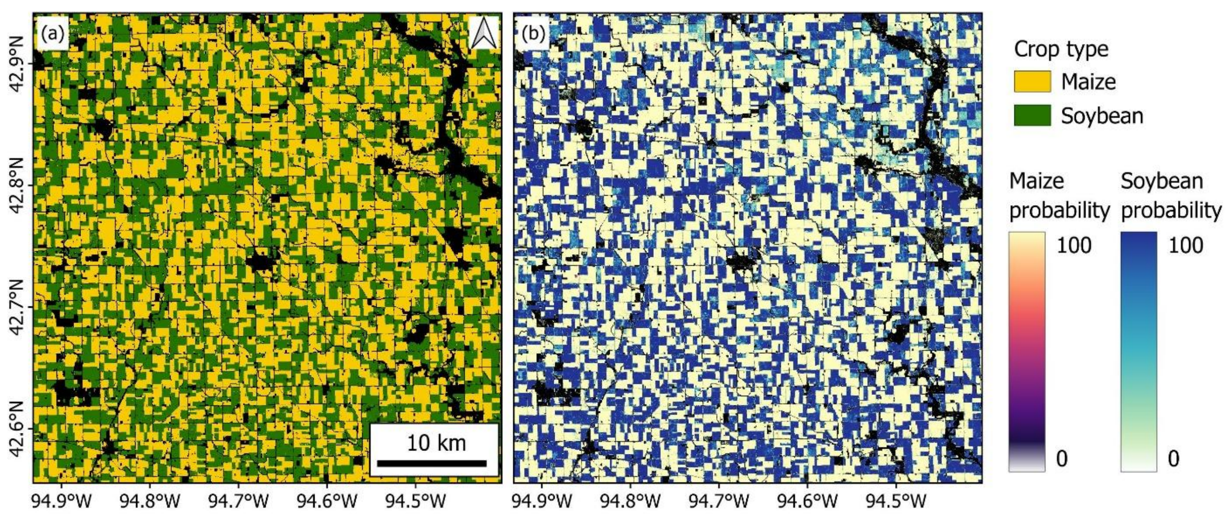

Fig R.6 Per-pixel classification aggregation from probability layers. (a) aggregated maize and soybean map; (b) probability layers derived from RF-Maize and RF-Soybean.

We have revised the manuscript in Line 255-260 to clarify: “We aggregated the per-pixel class probability layers from RF-Maize and RF-Soybean by selecting the highest probability (maize vs. soybean) and derived the aggregated probability layer and corresponding crop mask layer (see Figure S3 for an example). We then applied a 5 × 5 pixel kernel opening followed by a 10 × 10 pixel kernel closing, to eliminate scattered pixels and fill holes within large homogeneous fields.

The kernel sizes were selected based on tests and visual assessments to balance noise removal while preserving fine details. We generated the final maize and soybean map using the aggregated probability layer following the area-matching approach reported by Song et al (2017), Song et al. (2021b) and Li et al. (2023)."

The area-matching is the final step after the morphological operations. Please see also our response to the similar comment from Reviewer #2 above.

1. **Technical Details**

The authors state that the 12-m TanDEM-X elevation data was resampled to 10-m using "nearest neighbor" resampling. For continuous surface data like elevation, "bilinear interpolation" or "cubic convolution" is generally recommended to obtain a smoother, more realistic topographic transition.

While this may have a negligible impact on the final classification results, the authors could briefly explain the rationale for choosing "nearest neighbor," or, if possible, switch to bilinear interpolation. If "nearest neighbor" is retained, it is recommended to confirm that this choice does not introduce significant topographic artifacts at the 10-m scale.

Thank you for the thoughts. The selection of resampling methods to derive the 10-m elevation data from the original 12-m TanDEM-X data has a negligible impact on the final classification. As illustrated below, the demo TanDEM tile N43W103 is located in the Badlands National Park in southwestern South Dakota, which is characterized by eroded landscapes, sharp buttes and pinnacles (https://tandemx-science.dlr.de/cgi-bin/wcm.pl?page=DEM_Promotion_Page). There are also a few crop fields in the southeastern of this tile. Using different resampling strategies of nearest, bilinear or cubic did not result in significantly different 10-m elevation (Fig R.7). The per-pixel comparison between these resampling strategies shows that the slight elevation difference centered ~1 m, when all 10-m pixels including the crop and non-crop landscapes were used (Fig R.8). The elevation differences among different resampling strategies become smaller when only focusing on maize and soybean pixels within the 1-degree tile (Fig R.9).

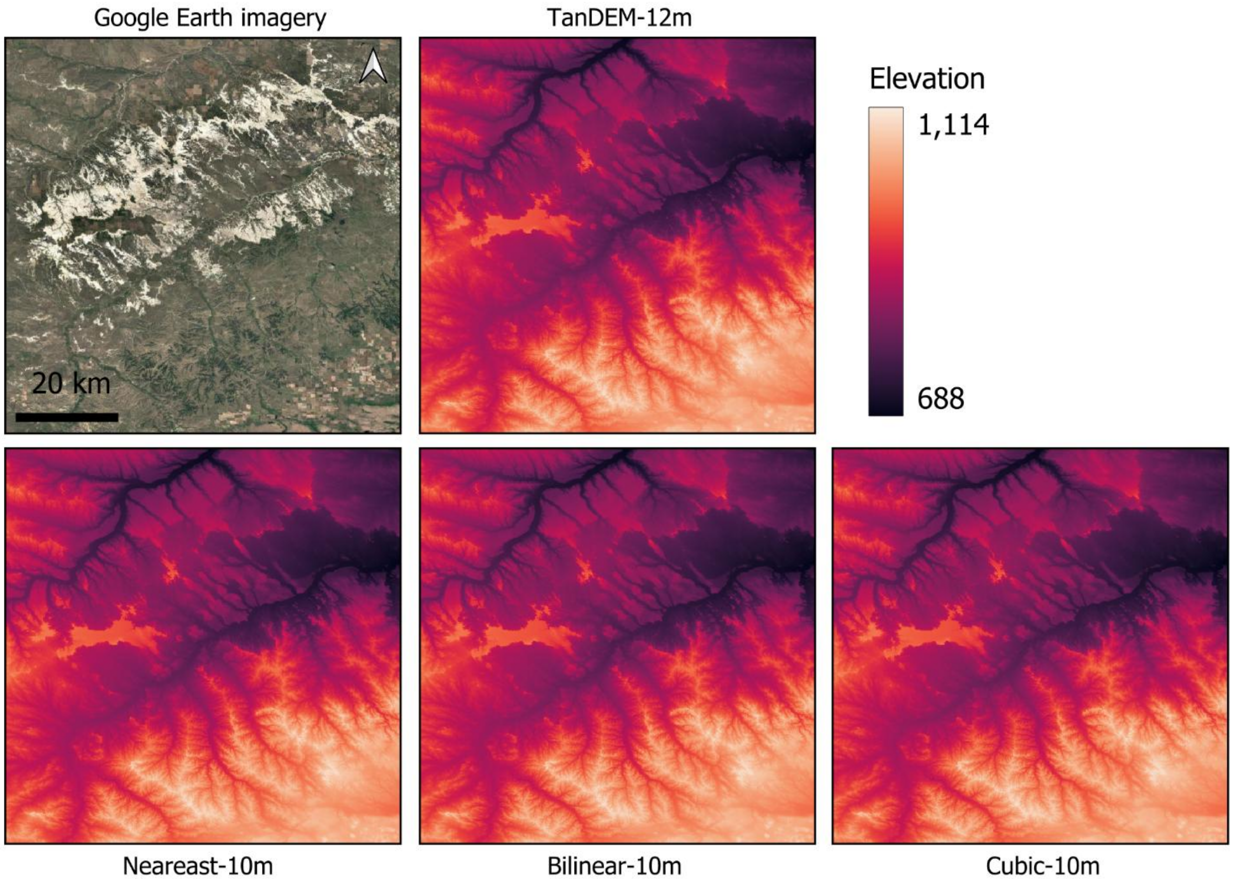

Fig R.7 The original 12-m TanDEM elevation and resampled 10-m elevation over the demo tile N43W103 in Badlands National Park in South Dakota.

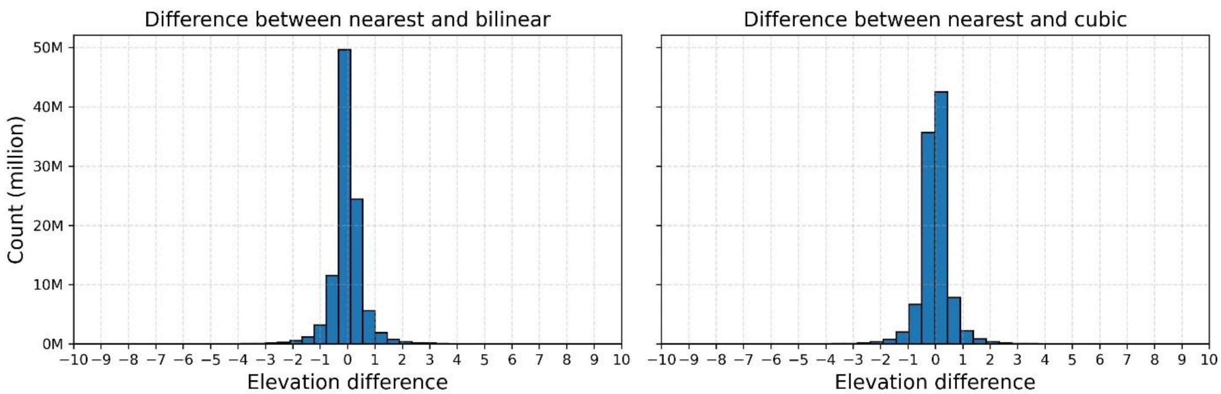

Fig R.8 The per-pixel elevation difference at 10-m resolution between nearest, bilinear and cubic resampling strategies. All pixels within the N43W103 are used to generate the statistics.

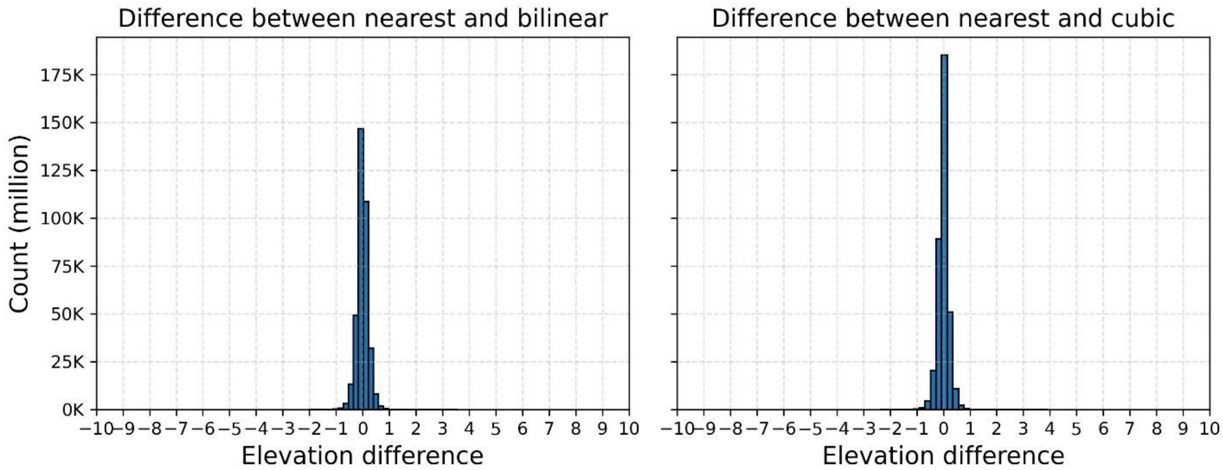

Fig R.9 The per-pixel elevation difference at 10-m resolution between nearest, bilinear and cubic resampling strategies. Only maize and soybean pixels within the N43W103 are used to generate the statistics.

In Section 2.3.2, the authors mention applying a "5x5 pixel kernel opening" followed by a 10x10 pixel kernel closing" after combining the probability layers. This is a key post-processing step to eliminate scattered pixels and fill holes.

However, the chosen kernel sizes (especially the 10x10 closing, i.e., 100m x 100m) are relatively large. The 5x5 opening will remove isolated patches smaller than 50x50m, while the 10x10 closing will fill non-crop holes (like ponds or buildings) within large fields if they are smaller than 100x100m.

While this improves the visual "purity" of the map product, it may also (especially in highly fragmented agricultural landscapes) eliminate fine-scale features that the 10-m data was intended to capture.

Please add a sentence in Section 2.3.2 to briefly explain the justification for choosing these specific kernel sizes (5x5 and 10x10). For example, was this based on considerations of average field size in the US, or were tests conducted to balance noise removal with detail preservation?

Thank you for the careful observation and questions. The chosen kernel sizes were empirical values based on tests and visual assessments to balance the results of “salt-and-pepper” effects and preserve fine details. We tried different kernel sizes and selected the 5x5 pixel kernel opening followed by the 10x10 pixel kernel closing since it resulted in better results compared to other kernel sizes and combinations (see an example in Fig R.10 below)

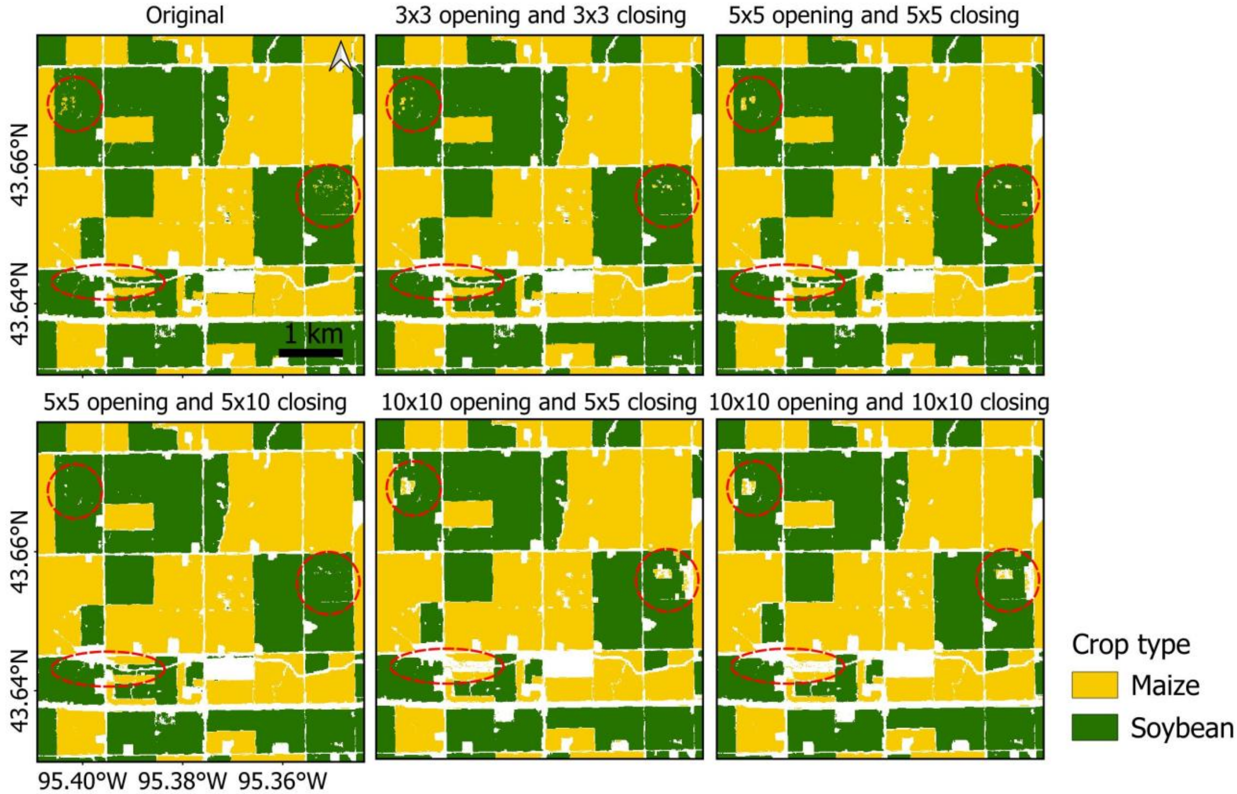

Fig R.10 Visual assessment of morphological results using different kernels.

We have added a sentence in Line 258: "The kernel sizes were selected based on tests and visual assessment to balance noise removal while preserving fine details."