Review Comments

Manuscript: essd-2026-347
Title: Mapping 20-years winter wheat dynamics in global primary planting areas using Gaussian mixture models with adaptive thresholds
Journal: Earth System Science Data (ESSD)
Recommendation: Major Revision

General Comments

This manuscript presents a framework for generating fractional winter wheat maps at 1 km resolution across 53 countries for the period 2001–2020, using MODIS surface reflectance data combined with Random Forest regression and Gaussian Mixture Models (GMM). The topic is relevant to ESSD and addresses a genuine need for long-term, large-scale crop distribution data. The overall ambition—producing the first global winter wheat fraction time series at 1 km—is commendable.

However, the manuscript in its current form suffers from serious deficiencies in conceptual clarity, methodological rigor, and depth of analysis. The approach is almost entirely empirical/statistical, with insufficient physical or mechanistic grounding. Many core concepts are introduced without adequate definition or justification. The validation strategy has potential circularity issues. I recommend **major revision**, with the specific concerns detailed below.

Major Concerns

1. Lack of Physical/Mechanistic Foundation

The entire framework is built on a cascade of statistical models: Random Forest regression → Gaussian Mixture Model fitting → Random Forest threshold prediction. At no point does the paper ground any of this in crop phenology, radiative transfer physics, or the spectral properties of wheat canopies. This is a fundamental weakness for a paper that claims to provide a data product for the Earth system science community.

Specific issues:

- **Why do Gaussian parameters carry physical meaning?** The paper states (L245–246) that the Gaussian model shows clear physical meaning regarding the dominant crop types. This claim is never substantiated. What is the biophysical interpretation of the mean (μ), standard deviation (σ), and amplitude (A) of a Gaussian fitted to the *histogram of predicted wheat fractions* within a 100 km grid? How do these parameters relate to planting intensity, field size, crop variety, or phenological stage? Without this link, the GMM is merely a curve-fitting exercise.
- **Why should a bimodal fraction distribution imply two crop types?** The paper uses bi-Gaussian fitting to discriminate wheat from non-wheat signals. But a bimodal distribution of RF-predicted fractions could arise from many factors: systematic RF bias, landscape fragmentation, topographic effects, atmospheric correction residuals, or artifacts of the MODIS compositing. The authors need to demonstrate—not assume—that the modes correspond to distinct crop types.
- **Spectral confusion is mentioned but never quantified.** Section 1 mentions spectral confusion within crop regions (L83–84), yet the method section provides no analysis of which crop types are most frequently confused with winter wheat, in which regions, or at which phenological stages. This is essential for users to understand the product's limitations.
- **The paper never explains *why* MODIS bands can separate winter wheat fractions.** Which spectral bands are most important in the Random Forest model? What phenological signals does MODIS capture for winter wheat specifically (e.g., the winter-spring green-up that distinguishes winter wheat from spring wheat)? The authors should report variable importance from the RF models and link these to known spectral-temporal signatures of winter wheat.

2. Validation Circularity and Independence

The validation strategy raises serious concerns about independence:

- **Training and validation data come from the same sources.** The fractional samples used for RF training are derived from the same CDL, EUCROPMAP, China winter wheat maps, and GlobalWheatYield4km products that are later used as reference data for validation. For instance, Section 4.3.2 compares the authors' results for China against aggregated maps from Dong et al. (2020)—the very dataset used to generate training samples. This is not independent validation.
- **The 70/30 train-test split at the grid level (Section 3.2) is insufficient** to guarantee independence when grids are spatially autocorrelated. The authors must demonstrate that the training and validation grids are spatially independent, e.g., through a spatial block cross-validation scheme.
- **No comparison with independent field survey data.** Given the availability of in-situ crop type data (e.g., LUCAS for Europe, field surveys by agricultural agencies), the authors should include validation against ground-truth points that were not used in any form during model development.
- **The FAO comparison (Section 4.3.1) is the only truly independent validation**, but it is only at the national level, which masks pixel-scale errors. The reported R^2 of 0.81 at the country level does not justify the claim that the product is robust at the pixel scale.

3. Temporal Extrapolation Without Validation

The paper describes a temporal extension strategy (L176–178) where winter wheat samples from 2020 are used to train models for earlier years (2001–2019). This is a critical step that receives almost no attention:

- **The assumption of temporal stationarity in wheat spectral signatures is never discussed**, let alone tested. Crop varieties, planting practices, and climate conditions change over two decades. A sample of winter wheat in Kansas in 2020 may not be spectrally representative of winter wheat in 2010, let alone 2001.
- **No temporal validation is performed for years before the reference data exist.** For China, the reference product begins in 2016, yet results are mapped back to 2001. How do the authors know the model works for 2001–2015?
- The authors should at minimum: (a) use a subset of years with available reference data to test temporal transferability, (b) report separate accuracy metrics for early vs. late periods, and (c) discuss how changes in MODIS sensor performance (e.g., Terra degradation over 20 years) may affect results.

4. The Adaptive Threshold Concept Is Poorly Defined

The title promises adaptive thresholds, and the GMM-derived optimal threshold is the claimed novelty. Yet:

- **The threshold is never formally defined in mathematical terms.** What exactly is being thresholded? The RF-predicted wheat fraction? After thresholding, is the output binary or still fractional?
- **What does locally optimal mean?** Optimal with respect to what objective function? The ROC balanced point is mentioned (L241), but its relationship to the GMM parameters is unclear. The paper states that the GMM parameters are used as *inputs* to a second Random Forest model that *predicts* the optimal threshold. This creates a convoluted chain: RF1 (MODIS → fractions) → GMM (fractions → parameters) → RF2 (parameters → threshold). The propagation of errors through this chain is never analyzed.
- **Figure 4 is referenced but the underlying mathematics are absent.** The paper needs equations defining (1) the single and bi-Gaussian models, (2) the ROC threshold derivation, (3) the RF model for threshold prediction, and (4) how the final map is produced from the threshold.
- **The choice of 100 km grid is arbitrary.** Why 100 km? A sensitivity analysis comparing results at 50 km, 100 km, and 200 km grids is needed.

5. Insufficient Error and Uncertainty Analysis

Section 4.4 on uncertainty is only two paragraphs and is far too superficial for a data paper in

ESSD:

- **No pixel-level uncertainty map is provided.** Users need to know *where* the product is reliable and where it is not. This is standard practice for gridded data products (see e.g., MODIS quality assurance layers, ESA CCI uncertainty maps).
- **No breakdown of error sources by magnitude.** The authors mention several error sources qualitatively (cropland mask errors, reference data errors, MODIS limitations, systematic RF bias), but never quantify their relative contributions.
- **The systematic regression bias** (underestimation of high fractions, overestimation of low fractions) is acknowledged in Section 4.1 but never quantitatively corrected or characterized.
- **How do errors propagate through the two-stage RF → GMM → RF pipeline?** This is never addressed.

6. Inadequate Comparison with Existing Products

- **No comparison with other global crop maps** such as GFSAD (30 m), GCROP (1 km), SPAM (10 km), or the ESA WorldCereal products. These are the products users would compare against when deciding whether to use this dataset.
- **No comparison with the MODIS Land Cover product (MCD12Q1) cropland classes**, which would be the most natural benchmark given the shared sensor.
- **The comparison with GlobalWheatYield4km (Section 4.3.2) is ambiguous.** The paper claims improved performance (L393) but provides no quantitative basis for this statement. If GlobalWheatYield4km was used for training, comparing against it and claiming improvement is circular.

7. Conceptual Gap: Fraction vs. Area

The paper repeatedly conflates winter wheat fraction with winter wheat area. A pixel with 50% wheat fraction means 50% of the 1 km² pixel is planted with winter wheat. When the authors aggregate these fractions to estimate national wheat area (and compare to FAO statistics), they implicitly assume that the fraction values are unbiased and additive. This assumption is never tested. Systematic biases in fraction estimation (acknowledged in Section 4.1) will propagate linearly to area estimates.

Specific Technical Issues

Methodology (Section 3)

1. **Section 3.1, Generating regressed fraction maps:**

- The adaptive strategies approach (Wen et al., 2022) is referenced at L221 but the method is not described. A reader should not need to consult an external paper to understand the core training procedure.
- How many training samples were collected per country? Per fraction stratum? The stratified sampling is described qualitatively but no sample counts are provided.
- The MODIS bands used and the temporal compositing window are not specified with sufficient precision. Which 8-day or 16-day composites are used? How are cloudy observations handled?

2. **Section 3.2, Gaussian mixture parameters:**
   - No equations are provided for the GMM. At minimum, Eq. (1) should define the single-Gaussian model and Eq. (2) the bi-Gaussian model.
   - How is model selection performed (single vs. bi-Gaussian)? AIC? BIC? Visual inspection?
   - The ROC mask threshold of 0.206 (Fig. 4a) is cited but its derivation is not explained. Is this value applied globally or per grid?
3. **Section 3.3, Evaluation:**
   - The validation is performed at 5 km aggregation (L422). Why 5 km? How sensitive are the reported metrics to the aggregation scale?

Results (Section 4)

4. **Section 4.1:** The histogram comparison in Fig. 5 uses only 5,000 samples. How representative are these of the full dataset? What is the K-S statistic or Wasserstein distance between true and predicted fraction distributions?

5. **Section 4.3.1:** The R^2 values in Fig. 7 pool all countries and years. This masks interannual and inter-country variability. The paper should report per-country and per-year metrics.

6. **The two distinct clusters in China's scatter plots (L428–429)** are attributed to higher planting density in China. This is a post-hoc explanation; the authors should test this hypothesis by comparing the high-fraction and low-fraction clusters against independent data on planting density or field size.

Figures

7. Several figures are described only in captions without sufficient standalone interpretability (e.g., Fig. 6, Fig. 10).

8. Fig. 10 claims six sites but the text (L402) references them with symbols that do not render clearly in the manuscript.

Minor Issues

1. **L150:** 500 km should be 500 m — this typo could confuse readers.

2. **L239:** The GLAD cropland dataset reference (Potapov et al., 2022) should specify the version and year used.

3. **Abstract:** States $R^2 = 0.81$ with FAO statistics, but Section 4.3.1 does not list this exact number in context. The abstract should specify whether this is the mean, median, or overall R^2.

4. **L82:** coarse-resolution satellite observations (e.g., 250 m, 500 m, and 1000 m) — 250 m and 500 m are not coarse for crop mapping; this statement is misleading. The issue is mixed pixels, not coarseness per se.

5. **English language:** The manuscript would benefit from careful proofreading. There are several awkward constructions (e.g., robust and harmonious random forest regression model at L206—what does harmonious mean in this context?; demonstrated significant performance at L505).

6. **Reference formatting:** Some references lack DOIs or have incomplete information (e.g., L539, AAFC reference).