Response to Reviewers

Article Ref.: essd-2025-83 Earth System Science Data

CY-Bench: A comprehensive benchmark dataset for sub-national crop yield forecasting

Dear Editor and Reviewers,

We thank you for the insightful suggestions, feedback, and the time you dedicated to reviewing our manuscript. Below, we provide a point-by-point response addressing each comment and suggestion.

We would also like to notify the Editor of an adjustment to the authorship order in this revised manuscript. The first and second author positions have been swapped to reflect the primary contribution of the current first author in executing the additional analyses, coordinating the revision, and finalizing the manuscript. This adjustment accurately reflects the intellectual leadership of the current work and has been explicitly approved by the previous lead author and all co-authors.

Please contact us if there are further suggestions. Thank you for your kind consideration.

Sincerely, Authors

Reviewer #1 comments and suggestions:

I am satisfied with most responses and revision, while the following should be double checked: 1. In response to my previous comments 3 and 4, you mentioned "Figures 5 and 6..." but I did not find Figure 6. If you referred to Figures 4 and 5 then I am satisfied with all responses and revision.

Thank you for pointing this out. Figures 3 and 4 from the original manuscript were merged and figure numbers were regenerated in the revised manuscript which caused the figure numbers to change. We apologize for any confusion this may have caused and have verified that all figure references are consistent.

Reviewer #2 comments and suggestions:

Harmonization: *"We have put considerable effort into harmonizing the datasets, focusing on aligning them spatially and temporally, while also applying unit conversions across all sources. This ensures consistency across diverse data sources and provides an analysis ready resource, eliminating the need for extensive preprocessing."*: The response is not persuasive. Due to the spatial heterogeneity, there is no guarantee that the harmonized datasets provide the same evaluation capability across different regions. I think it makes more sense to design region-specific datasets and features.

We thank the reviewer for highlighting a key consideration regarding the interpretation of harmonized datasets. In this work, we harmonized multi-source satellite, meteorological, and soil datasets of varying spatial and temporal resolution by aligning them to a common spatial grid and temporal resolution, and applying consistent aggregation strategies and unit conversions to ensure uniformity. While this harmonization does not eliminate intrinsic regional agro-climatic heterogeneity nor does it guarantee identical predictive capacity across regions, it removes methodological inconsistencies that typically arise from disparate data sources and preprocessing pipelines. The resulting dataset provides a standardized, analysis-ready benchmark with a fixed set of predictors across regions. This design choice enables controlled, reproducible comparison of crop yield prediction methods across countries, isolating the impact of modeling approaches from that of region-specific data engineering. Rather than replacing region-specific datasets or tailored feature engineering, our benchmark serves as a common reference framework upon which localized extensions can be built. We would also like to reiterate that the selected predictors represent a well-established set of agro-climatic and environmental factors known to strongly influence crop yields.

"Our study spans multiple continents, including regions with diverse agricultural practices and infrastructure" I do not think the scope is a critical character to a good benchmark dataset.

We agree that geographic scope alone does not define the quality of a benchmark dataset. However, in the context of crop yield forecasting, scope becomes scientifically and societally relevant because many downstream analyses and policy decisions rely on models operating at continental to global scales (e.g., in global food security assessments and IPCC-related analyses). In this setting, the absence of a standardized and curated global benchmark makes it difficult to assess whether modeling approaches perform consistently across regions or whether apparent global conclusions are driven by uneven regional performance. The inclusion of multiple continents is a deliberate design choice aimed at (1) providing data over several countries including ones where data-driven yield prediction is highly lacking due to inability to accessing crop statistics, (2) enabling systematic evaluation of model generalization and robustness under diverse agro-climatic regimes, and data availability conditions (3) supporting transfer learning research. This diversity introduces controlled domain shifts that are difficult to capture with country-specific benchmarks, opening opportunities for researchers to assess how well crop yield models transfer beyond the conditions under which they are trained.

"Panel of experts" it is good to have a panel of experts. Given that, I think some of the experts may be able to propose region-specific features, instead of using NDVI for all regions.

We thank the reviewer for highlighting the interdisciplinary expertise of our team. We agree that region-specific, expert-designed features can improve local predictive performance. However, the goal of this work is to construct a global standardized benchmark dataset rather than expert-tuned features. To ensure fair and reproducible comparison across regions, we deliberately adopt a common set of predictors that are physically meaningful, widely available,

and commonly used in the crop yield prediction literature. Accordingly, we rely on globally consistent indicators such as NDVI, not because they are optimal in every local context, but because their availability, interpretability, and widespread use make them suitable for standardized benchmarking across diverse agro-climatic regions. Introducing region-specific features would lead to heterogeneous input spaces across regions and introduce subjective design choices that complicate cross-region comparison and transfer learning. The interdisciplinary expertise of the team was instead leveraged to curate a minimal yet sufficient, defensible set of predictors that generalize across agro-climatic contexts. We view region-specific feature engineering as a complementary direction that can be built on top of this benchmark, rather than as part of the benchmark itself. A recent work carried out by the EU Joint Research Center (JRC) which operates a large-scale yield forecasting system, relied on a core suite of predictors rather than bespoke ones to forecast yield at the national level in about 77 countries. The use of a core set of standardized predictors across regions is a well-established practice in large-scale operational yield forecasting systems (Sabo et al., 2024).

The authors did not address the static crop mask issues. It is a very critical issue, and it is not scientifically sound to use the same crop mask for year 2003 and year 2023. Crop rotation and fallow can happen frequently. The prediction results for corn can be totally wrong if NDVI signals are extracted from wheat or rice fields. A benchmark dataset should have a high quality.

We agree with the reviewer that, in principle, temporally explicit crop masks would be preferable to capture field-level dynamics such as crop rotation in countries where such practices are common. This limitation has been acknowledged in the manuscript. However, at present there is no globally consistent, annually updated, crop-specific mask covering the full 2003–2023 period across all regions included in this benchmark. We therefore adopt a single global static crop mask, rather than combining heterogeneous dynamic products with differing definitions, accuracies, and temporal coverage.

We considered the alternative suggested by the reviewer: using dynamic masks (e.g., MIRCA-OS). However, these products have a much coarser spatial resolution (≈10km) compared to our static mask (≈500m). At sub-national aggregation scales, this introduces a systematic mixed-pixel bias, whereby crop signals are diluted by non-agricultural land cover. For benchmarking purposes, we argue that the stochastic noise introduced by a static mask—which largely averages out during spatial aggregation—is preferable to this systematic bias.

To directly assess the practical relevance of the reviewer's concern, we conducted a dedicated sensitivity analysis using MIRCA-OS to isolate the contribution of crop mask variability relative to the natural interannual variability of the indicators themselves. This analysis shows that the median relative variability introduced by mask changes ($\sigma_{mask}/\sigma_{signal}$) is approximately 2% across more than 12,000 regions, indicating that mask-induced variability is small compared to the intrinsic year-to-year signal variability.

Technical Note: Sensitivity Analysis

We formalized the regional aggregation $A_{y,\tau}$ as the weighted mean of the pixel-level indicator $I_{i,y,\tau}$ using the crop fraction w:

$$A_{y,\tau} = \frac{\sum_{i=1}^{N} w_i \cdot I_{i,y,\tau}}{\sum_{i=1}^{N} w_i} \tag{1}$$

We isolated the "mask effect" by calculating the indicator in two distinct ways:

1. **Signal Variance:** The crop mask is held constant, while the vegetation signal varies according to its observed interannual dynamics.

2. **Mask Variance:** The vegetation signal is held constant at its climatological mean, isolating the variability driven solely by annual changes in the crop mask (using MIRCA-OS 2000–2015).

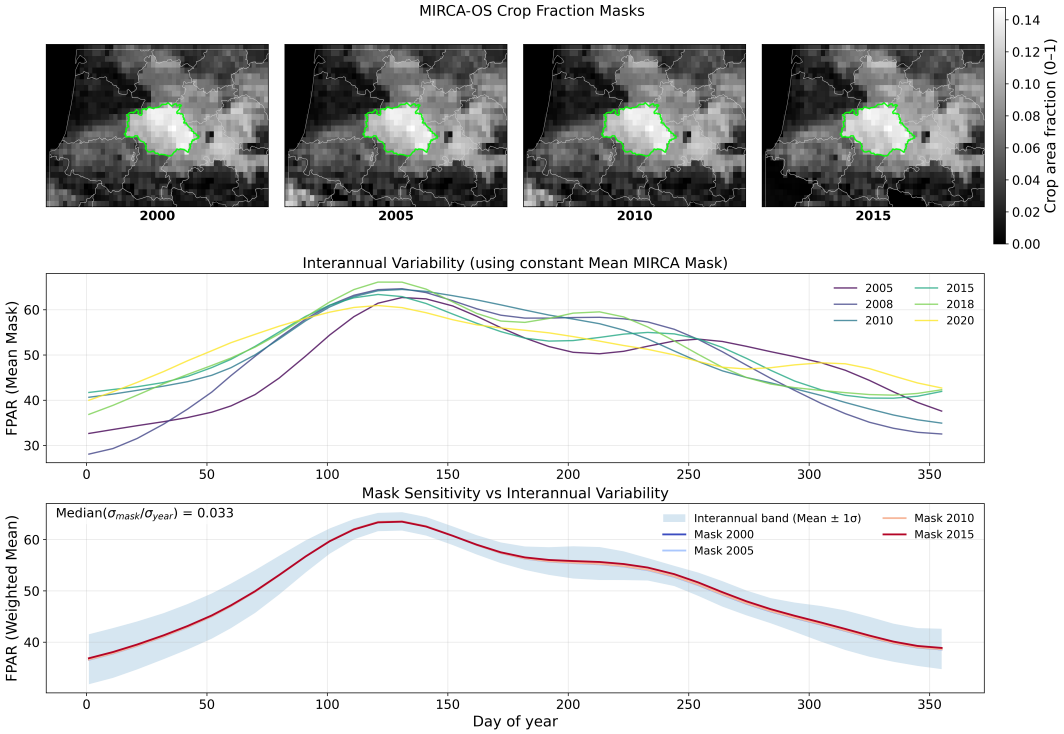

Figure 1: The plot compares the variability of the regional indicator driven by year-to-year fluctuations versus the variability driven purely by crop mask evolution for a representative region. The vertical spread of the mask-induced variability (colored lines) is minor compared to the natural interannual variability of the indicator signal (blue envelope).

Figure 1 illustrates the results for a typical region. The variance introduced by mask changes is small relative to the interannual variability of the indicator (blue envelope).

We found that the median relative variability—defined as the median over time of the ratio $\sigma_{mask}/\sigma_{signal}$, where both quantities are evaluated at each time point—is approximately 2%. This result is robust across regions, based on an analysis of over 12,000 regions. This quantitative evidence confirms that the error introduced by a static mask is sufficiently low relative to the signal of interest.

In addition, our preprocessing pipeline restricts predictor time series to crop-specific SOS–EOS windows and aggregates signals at the administrative-unit level. These steps further reduce the influence of off-season vegetation and isolated field-level crop substitutions, such that the extracted signals represent the dominant cropping system within each region rather than individual rotated fields.

Finally, it is worth noting that many dynamic crop-type products are generated retrospectively, relying on observations from the full growing season. Because CY-Bench is explicitly designed to support in-season, pre-harvest yield forecasting, the use of a static mask is also consistent with realistic operational settings where future crop-type information is not yet available.

Notwithstanding these considerations, we recognize that the lack of temporally explicit crop masks remains a limitation. We emphasize that the use of a static crop mask reflects current global data availability and benchmarking priorities rather than a conceptual limitation. The WorldCereal project is actively developing temporally explicit, global crop type products, and we are committed to incorporating such dynamic crop masks into future releases of CY-Bench once they become consistently available across regions and years.

We have added the sensitivity analysis to the Discussion (Limitations) section of the manuscript.

"It is important to note that lower performance in some regions should not necessarily be viewed as a flaw, but rather as an indication that the current set of predictors may be insufficient to fully capture the yield variability in those areas" I fully agree with the authors on this point. If the current set of predictors is not sufficient to fully capture the yield variability in certain areas, it means that it cannot be used as a benchmark. No one would be able to further improve the accuracy based on the data.

We appreciate the reviewer's agreement that the current predictor set may not fully explain yield variability in some regions. However, we respectfully disagree with the inference that this disqualifies the dataset as a benchmark. A benchmark dataset is not required to contain all information necessary to achieve near-optimal or region-specific accuracy. Instead, a benchmark provides a fixed, well-defined input space against which methodological advances can be evaluated in a controlled and reproducible manner. Importantly, this fixed input space also enables systematic evaluation of which predictors and feature sets are most informative across regions, rather than conflating feature availability with modeling choices.

Many widely used benchmarks in machine learning and Earth observation (e.g., Xia et al. (2023), Yeh et al. (2021), Lacoste et al. (2023)), include challenging regions or regimes where performance is limited, precisely to expose methodological differences and research gaps.

Even with an identical set of predictors, substantial performance differences can arise from e.g. modeling choices or number of data points. Regions with lower predictive performance are therefore informative rather than problematic: they highlight where commonly used predictors are insufficient, reveal the interaction between data availability and model performance, and

identify settings where transfer learning or additional data sources are most needed.

We have revised the Discussion (Impact) section of the manuscript to clarify that the purpose of the benchmark is to support comparative evaluation of modeling approaches under realistic and widely available data constraints, rather than to define an upper bound on achievable yield prediction accuracy. An indepth look into performance disparity and potential causes will be presented in a follow-up analysis paper that provides a thorough evaluation of data-driven and statistical models across all countries.

References

Lacoste, A., Lehmann, N., Rodriguez, P., Sherwin, E., Kerner, H., Lütjens, B., Irvin, J., Dao, D., Alemohammad, H., Drouin, A., et al.: Geo-bench: Toward foundation models for earth monitoring, Advances in Neural Information Processing Systems, 36, 51 080–51 093, 2023.

Sabo, F., Meroni, M., Kerdiles, H., Vojnovic, P., Piles, M., Munoz-Mari, J., Mateo Sanchis, A., and Rembold, F.: Technical note on Large Scale Yield Forecasting v 1.0, 2024.

Xia, J., Yokoya, N., Adriano, B., and Broni-Bediako, C.: Openearthmap: A benchmark dataset for global high-resolution land cover mapping, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 6254–6264, 2023.

Yeh, C., Meng, C., Wang, S., Driscoll, A., Rozi, E., Liu, P., Lee, J., Burke, M., Lobell, D., and Ermon, S.: SustainBench: Benchmarks for Monitoring the Sustainable Development Goals with Machine Learning, in: Thirty-fifth Conference on Neural Information Processing Systems, Datasets and Benchmarks Track (Round 2), `https://openreview.net/forum?id=5HR3vCylqD`, 2021.