Response to Reviewers

Article Ref.: essd-2025-83

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

Earth System Science Data

Dear Editor and Reviewers,

We thank you for the insightful suggestions, feedback, and the time you have dedicated to review our manuscript. This is our initial response to your comments. We have provided a point-by-point response below, addressing each specific comment and suggestion. Our actionable items are highlighted in *italics*.

We appreciate your consideration of the effort we made. Please contact us if there are other further suggestions.

Thank you for your kind consideration.

Sincerely,

Authors

Reviewer #2 comments and suggestions:

This study proposes CY-Bench, a global, sub-national benchmark for in-season, pre-harvest crop-yield forecasting of maize (covering 38 countries) and wheat (covering 29 countries). The data are collected and curated from large-scale open-access sources. The data are potentially useful for developing and evaluating machine learning models for crop yield forecasting and other Earth system related tasks. In general, CY-Bench is novel and useful for the machine learning and agricultural communities. The following comments should be considered before formally publishing it:

We appreciate your recognition of the novelty and potential value of CY-Bench for both the machine learning and agricultural research communities. We carefully considered your comments and a detailed, point-by-point response to each suggestion is provided below.

1. As CY-Bench is proposed for developing and evaluating data driven models, you should discuss and compare the performance of the models you have benchmarked in the main text, despite a table of model performance is provided in the code repo. You should also discuss

whether there are different findings (of benchmarking these models) compared to existing datasets.

The primary objective of the current study is to introduce and document the CY-Bench dataset, including its scope, geographic coverage, and potential applications. While the performances of a variety of models are provided in the code repository for reference, a comprehensive analysis and discussion of model benchmarking will be addressed in a forthcoming follow-up study. This separation allows us to focus this manuscript on dataset creation, accessibility, and usability, which we believe are the critical contributions of CY-Bench.

A full evaluation of model performance is therefore beyond the scope of this paper. Nevertheless, as detailed in our introduction, in the literature, machine learning methods often outperform classical statistical baselines and this trend does not fully emerge in our experiments. Direct comparisons remain challenging, however, due to non-standardized datasets and evaluation protocols.

We will revise the manuscript by adding a short note on the initial benchmarking experiments included in the repository. This will briefly acknowledge the availability of baseline results while clarifying that a full and systematic evaluation of model performance will be presented in a separate follow-up study.

2. Your benchmarking results (as shown in the code repo) displays large variations across different regions. For example, Maize (CN) achieves low NRMSE (8.78) and close-to-one R2 (0.81), while Maize (DK) reaches much higher NRMSE and all very negative R2. This problem can be caused by the models you train or the dataset itself (e.g., quality of the data, predictors you choose), and should be sufficiently discussed in the paper.

The variation in benchmarking results across regions is not unexpected and reflects several underlying factors. In the case of China, yields exhibit a relatively stable spatial pattern, where differences between regions are fairly consistent over time. This makes it relatively easy to achieve high accuracy, especially with location-specific models. For Denmark, the lower performance is likely related to the significantly smaller size of the available dataset, which reduces the predictive signal for the models.

We agree that variability in model performance may also reflect the relevance and completeness of the input features and/or quality of the labels. The fact that the same predictors perform reasonably well in one region while underperforming in others strongly suggests that the explanatory power of these features can vary significantly by location. In our view this finding is not a flaw; it warrants a revisit on the conventional modeling and data selection approaches and offers valuable guidance for assessing the quality of the yield statistics.

We would like to stress that the modeling component and feature importance analysis will be covered in our follow-up analysis paper that systematically compares different modeling approaches and explores their performance across regions and conditions.

We are preparing a separate analysis paper to systematically explore and compare different

modeling approaches, including an examination of the completeness and suitability of input features, building directly on the insights revealed by CY-Bench. Nevertheless, in response to the reviewer's suggestion we will add brief comments on our initial modeling results. In particular, we will note that the benchmarking experiments suggest that the predictive value of the available features differs across regions, leading to variation in model performance. This highlights the complexity of yield prediction across diverse contexts and underscores the importance of CY-Bench as a standardized benchmark to study these differences systematically.

3. Following the previous comment, I recommend adding a note for each region (or a group of regions) to guide potential users on the specific precautions needed when working with its data (e.g. quality concerns, noise or any other risks).

We agree that region-specific guidance can help users better utilize and interpret the dataset. Our current version already addresses some aspects of this guidance: Figures 5 and 6 show the number of observations per country and the number of administrative regions, highlighting data availability across regions. We also address quality issues with yield labels sourced from government-reported statistics. In the discussion, we highlight risks such as the lack of a temporal crop mask—which may impact feature quality in regions where crop rotation is common—and the lack of separation between irrigated and non-irrigated yields, which could otherwise provide additional insights for data-driven models.

To meet the reviewer's suggestion, we will include notes and provide discussions summarizing key considerations for the use of CY-Bench. These notes will address three key aspects:

First, data quantity: The number of observations varies across regions, which can affect the performance of data-driven models.

Second, input predictor relevance: While our selected predictors generally capture major determinants of yield, their explanatory power arguably varies by region. In some areas, other factors—such as management practices or resource constraints—may more strongly influence yields, which can limit predictive accuracy.

Third, label quality: With respect to yield data, we rely on reprocessed and curated government-reported statistics and apply quality checks to filter implausible values and ensure internal consistency. We note that the community has limited consensus on how to implement quality control and uncertainty analysis for yield data (Davis et al., 2025). Possible approaches include assigning a quality tag based on the data source or on cropping area.

In the next iteration of the dataset, we plan to provide region-specific notes addressing these aspects wherever relevant. In the revised manuscript, we will also expand the discussion on quality control and uncertainty associated with the yield statistics.

4. The size of data in each region should be indicated.

The dataset size for each country is presented in Figures 5 and 6, which report both the number of regions and the number of yield observations. These figures also lists the average administrative size of a region within each country.

5. Lines 275 - 280, you should discuss why LOYO is better than random sampling, especially if random sampling was previously more commonly used as you mentioned. As you have noted in limitations, LOYO is a compromise for small datasets; is there a better way for regions with more data?

We used leave-one-year-out (LOYO) cross-validation rather than random sampling to address spatial correlations within the same year. Yields from neighboring regions in the same year are correlated, so random sampling can include similar data in both training and testing sets, violating the independent and identically distributed (IID) assumption and leading to overly optimistic performance estimates.

LOYO tests on an entire year excluded from training, avoiding this issue. It also preserves extreme years in evaluation: the impact of unusually low or high yields is fully represented, rather than being diluted or blended across random splits.

While LOYO is a practical compromise for smaller datasets—allowing maximum use of available training data while ensuring each year is evaluated—regions with larger datasets could benefit from forward sliding (rolling-window) validation, which better mimics operational forecasting. Overall, LOYO provides a balance between maintaining spatial independence, including extreme years in the evaluation, and efficiently using the available data.

We will elaborate on our choice of LOYO cross-validation compared to random sampling in the revised manuscript, highlighting how it addresses spatial correlations within the same year and preserves the impact of extreme yield years without blending them across splits

6. In Section 2, how the data are harmonized should be described with more details, instead of simply mentioning it (e.g., Lines 180 - 185 "as they follow a harmonization procedure developed by ..." and Lines 185 - 190 "The data was compiled and harmonized to account for ...").

We will clarify in the revised manuscript that yield datasets used in our study are harmonized to ensure spatial and temporal consistency. For the EU, Ronchetti et al. (2024) harmonized data from multiple sources, standardizing crop definitions, administrative boundaries, and reporting practices, producing comparable annual yield time series. For African countries, the FEWS NET HarvestStat Africa dataset, as compiled and harmonized by Lee et al. (2025), adjusts for changes in administrative boundaries and reporting inconsistencies over time. These harmonization procedures ensure that the annual yield time series are suitable for trend analysis and model evaluation.

We will elaborate further on the harmonization procedures described above in the revised manuscript

7. Following the previous comment, you may consider describing more details on how you collect and curate data from a large number of different sources, as this is a major methodological contribution of your study and will be useful for other researchers.

For each data source, both targets (yield), predictors (weather, vegetation, soil) and auxiliary

data (crop mask and crop calendar), we provide detailed data cards in our GitHub repository. These data cards describe provenance, collection, processing, and curation in a transparent and reusable way. Our workflow, which shows how the yield and predictor data is processed, is outlined in Section 2.2 and graphically summarized in Figure 1.

In the revised manuscript we will expand Section 2.2 to provide a clearer description of how the multi-source data are harmonized, including the steps for temporal alignment to crop seasons and the aggregation to administrative units.

References

Davis, K. F., Anderson, W., Ehrmann, S., Flach, R., Meyer, C., Proctor, J., Ray, D. K., You, L., Foley, M., Kerdiles, H., et al.: HarvestStat: A global effort towards open and standardized sub-national agricultural data, Environmental Research Letters, 2025.

Lee, D., Anderson, W., Chen, X., Davenport, F., Shukla, S., Sahajpal, R., Budde, M., Rowland, J., Verdin, J., You, L., et al.: HarvestStat Africa–harmonized subnational crop statistics for sub-Saharan Africa, Scientific Data, 12, 690, 2025.

Ronchetti, G., Nisini Scacchiafichi, L., Seguini, L., Cerrani, I., and van der Velde, M.: Harmonized European Union subnational crop statistics can reveal climate impacts and crop cultivation shifts, Earth System Science Data, 16, 1623–1649, https://doi.org/10.5194/essd-16-1623-2024, 2024.