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,

In response to the reviewer's request for a clear reconciliation of our previous commitments with the final manuscript, we have carefully reviewed all comments and responses from the first round. The reviewer's instructions are quoted below:

"In the first round of review, the authors promised a few things to refine in the revised manuscript. For example, in response to my comment 3, the authors wrote "[...] In the revised manuscript, we will also expand the discussion on quality control and uncertainty associated with the yield statistics". Please review all of my previous comments and your corresponding responses, highlight all your promised revisions with tracked changes, and indicate their line numbers in the manuscript so that I can assess the revisions. If the changes were made on the zenodo or github repos, please indicate where and what has been changed."

We present our responses in a structured format, organized point by point as follows:

- **Reviewer Comment:** The original feedback from the previous round (Blue).
- **Previous Author Response:** Our initial justification or clarification (Black).
- **Action Promised:** The specific commitment we made to improve the paper (Red).
- **Action Taken:** A description of the implemented changes, including specific Section and Line numbers in the revised manuscript and, where applicable, links to updated repository files (Green).

This submission includes: (1) an annotated version of the manuscript with tracked changes; and (2) this document.

Sincerely,

Authors

Reviewer comments and suggestions from previous rounds:

Reviewer Comment:

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:

Previous Author Response:

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.

Reviewer Comment:

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.

Previous Author Response:

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, and of the journal itself. 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.

Action Promised:

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.

Action Taken:

We added the following to *Section 4.1 (L324-L332)*: "To illustrate the potential of CY-Bench for crop yield prediction, we provide initial benchmarking results using several machine learning models in the accompanying code repository. These results offer a preliminary sense of model performance across regions and model choices and show that the predictive value of features in CY-Bench varies. This highlights the complexity of yield prediction across diverse landscapes and underscores the importance of CY-Bench as a standardized benchmark to study these differences systematically for data-rich and -sparse regions. While these results are only illustrative, a full comparison of methods, including systematic benchmarking and analysis of findings relative to existing datasets will be addressed in a dedicated follow-up study. We emphasize that the purpose of CY-Bench 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."

Reviewer Comment:

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.

Previous Author Response:

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.

Action Promised:

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.

Action Taken:

We added the following to *Section 4.1 (L324-L332)*: "To illustrate the potential of CY-Bench for crop yield prediction, we provide initial benchmarking results using several machine learning models in the accompanying code repository. These results offer a preliminary sense of model performance across regions and model choices and show that the predictive value of features in CY-Bench varies. This highlights the complexity of yield prediction across diverse landscapes and underscores the importance of CY-Bench as a standardized benchmark to study these differences systematically for data-rich and -sparse regions. While these results are only illustrative, a full comparison of methods, including systematic benchmarking and analysis of findings relative to existing datasets will be addressed in a dedicated follow-up study. We emphasize that the purpose of CY-Bench 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."

Reviewer Comment:

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).

Previous Author Response:

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.

Action Promised:

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.

Action Taken:

We added the following to *Section 4.2 (L365-L373)*: "Government-reported yield statistics can vary in quality across countries due to differences in data collection, aggregation, and reporting accuracy. For a large fraction of the data, we rely on reprocessed and curated statistics from the relevant national agencies. For quality control of these data, we refer to the relevant studies and papers that detail the validation methods applied. Where available, the data card for each country provides a link to these references. In addition, we apply basic quality checks, such as filtering out zero or missing yields and verifying internal consistency ($yield = production/area$). Nevertheless, there is currently no universally accepted protocol for quality control or uncertainty assessment of yield data (Davis et al., 2025). As such, CY-Bench does not include formal uncertainty estimates for each observation, though future iterations could incorporate quality indicators based on the source, cropping area, or other metadata."

Beyond the manuscript revisions, we have also updated the CY-Bench repository on Zenodo (v1.10) to provide additional (meta)data that supports more robust modeling and quality assessment. Specifically, we added "region_area", "crop_area", and "crop_area_percentage" to allow for weighting models by actual cultivation intensity or filtering out low-confidence regions where the target crop is sparse. These additions allow users to implement their own uncertainty weighting or quality-filtering protocols, thereby addressing the lack of a universal protocol mentioned in the text.

Furthermore, we are preparing a follow-up analysis paper to systematically evaluate how data quantity, unobserved determinants, and regional characteristics influence modeling outcomes, building on the limitations and variability highlighted in this study.

Reviewer Comment:

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

Previous Author Response:

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.

Action Taken:

No changes were required as this was already addressed in the original Figs 5 and 6 (which in the revised version is now Figure 4 and 5)

Reviewer Comment:

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?

Previous Author Response:

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, aka data leaking.

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.

Action Promised:

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

Action Taken:

We added the following to *Section 3.3 (L305-L309)*: Yields from neighboring regions in the same year are typically correlated, so random sampling can cause data leakage: information from the same year appears in both training and testing sets. This violates the independent and identically distributed (IID) assumption and produces overly optimistic performance estimates. LOYO avoids this by holding out an entire year, ensuring that correlations within that year are only encountered during evaluation. It also guarantees that extreme yield years are fully represented, rather than being diluted across random splits.

In addition we added the following to *Section 4.2 (L405-L408)*: "While LOYO is a prac-

tical 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."

Reviewer Comment:

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 ...").

Previous Author Response:

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.

Action Promised:

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

Action Taken:

We added the following to *Section 2.1.5 (L206-L207)*: "(...) standardizing crop definitions, administrative boundaries, and reporting practices to produce comparable annual yield time series. (...)"

In addition we added the following to *Section 2.1.5 (L211-L212)*: "(...) and harmonized by (Lee et al., 2025) to account for changing administrative boundaries and reporting inconsistencies over time (...)"

Reviewer Comment:

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.

Previous Author Response:

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.

Action Promised:

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.

Action Taken:

We added the following to *Section 2.1.5 (L127-L128)*: "(...) Each selected dataset is further described in accompanying data cards, which provide links to sources, reports, and related publications. (...)"

We also revised our data cards on *Github* to include more details on provenance, collection, processing, and curation. Revisions can be seen via `https://github.com/WUR-AI/Ag ML-CY-Bench/commits/main/data_preparation?since=2025-03-12`

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.