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

D Paudel[1] **H Baja**[1] **R van Bree**[1] **M Kallenberg**[1] **S Ofori-Ampofo**[2] **A Potze**[1] **P Poudel**[3] **A Saleh**[4] **W Anderson**[5] **M von Bloh**[2] **A Castellano**[6] **O Ennaji**[7] **R Hamed**[8] **R Laudien**[9] **D Lee**[10] **I Luna** [11] **D Masiliunas**[1] **M Meroni**[12] **S Mkuhlani** [13] **J Mutuku**[14] **J Richetti**[15] **A Ruane**[6] **R Sahajpal** [5] **G Shuai**[5] **V Sitokonstantino**[11] **R de S. Nóia Jr**[2] **A Srivastava**[16] **R Strong** [17] **L Sweet**[18] **P Vojnović**[12] **A de Wit** [1] **M Zachow**[2] **I Athanasiadis**[1]

[1]Wageningen Uni. & Research [2]Technical Uni. Munich [3]Purdue Uni. [4]Ankara Uni. [5]Uni. Maryland [6]NASA [7]Univ. Mohammed VI [8]VU Amsterdam [9]PIK [10]Uni. Manitoba [11]Uni. València [12]JRC [13]IITA [14]ICRISAT [15]CSIRO [16]ZALF [17]Texas Uni. [18]UFZ

Abstract

In-season or pre-harvest crop yield forecasts are essential for enhancing transparency in commodity markets and for planning towards achieving the United Nations' Sustainable Development Goal 2 of zero hunger, especially in the context of climate change and extreme events leading to crop failures. Pre-harvest crop yield forecasting is a difficult problem, as several interacting factors contribute to yield formation, including in-season weather variability, extreme events, long-term climate change, pests, diseases and farm management decisions. Machine learning methods provide ways to capture complex interactions among such predictors and crop yields. Prior research in agricultural applications, including crop yield forecasting, has primarily been case-study based, which makes it difficult to compare modeling approaches and measure progress. To address this gap, we introduce CY-Bench (Crop Yield Benchmark), a comprehensive dataset and benchmark to forecast crop yields. We standardized data source selection, preprocessing and spatio-temporal harmonization of public sub-national yield statistics with relevant predictors such as weather, soil, and remote sensing indicators, in collaboration with domain experts such as agronomists, climate scientists, and machine learning researchers. With CY-Bench we aim to: (i) standardize machine learning model evaluation in a framework that covers multiple farming systems in more than twenty-five countries across the globe, (ii) facilitate robust and reproducible model comparison through a benchmark addressing real-world operational needs, (iii) share a dataset with the machine learning community to facilitate research efforts related to time series forecasting, domain adaptation and online learning. The dataset and code used will be openly available, supporting the further development of advanced machine learning models for crop yield forecasting that can be used to aid decision-makers in improving global and regional food security.

Keywords: benchmark dataset; crop yield forecasts; agriculture; food security.

1 Introduction

Despite steady improvements in the efficiency of agricultural production over the last decades, the global food system is still rife with inequalities (60; 1), such as disproportionate access to resources

Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.

between developed and developing countries. The interconnectedness of countries and international trade can help to smooth swings in commodity prices, but can also bring intra-annual price volatility to import-dependent countries (81; 12; 80). Experts have emphasized the need for improved data, maps, and predictions (20; 37; 17). In particular, pre-harvest yield forecasts are vital for improving global market transparency and enabling decision-makers to plan response actions to mitigate anticipated shortages (65; 63; 7).

National and sub-national crop yield forecasts are produced by both private sector and governmental institutes using a combination of statistical modeling approaches and process-based crop models (5; 58; 23). Due to the multiplicity of systems and hazards involved, and the importance of compounding effects which are not yet well-understood, data-driven methods provide less explored ways to capture the complex and nonlinear relationships driving crop growth and development(59; 31). Additionally, the availability of high-quality agricultural data varies significantly by region and by crop; recent developments in transfer learning and domain adaptation may be useful for serving data-scarce regions or neglected and under-utilized crops. Over the recent years, several review articles (14; 25; 32; 73; 8; 42) and publications have highlighted excellent performance of machine learning for pre-harvest yield forecasting (79; 19; 28; 36; 44; 45; 76; 34). However, the data and code used in these studies are often unavailable, meaning that the results cannot be reproduced, and the diverse range of evaluation procedures, metrics, and datasets used in these studies means that synthesizing their results is difficult.

In order to better understand the specific strengths and weaknesses of existing machine learning methods for pre-harvest yield forecasting, and to drive further research progress, well-specified benchmark datasets compiled by domain experts are vital (53; 67; 16)(Sweet et al. in review). These benchmark datasets must reflect the needs of the worldwide community (41; 71). Recently, researchers have emphasized the need for machine learning benchmark datasets that include data from more regions and countries (50). Additionally, while forecast accuracy is crucial, machine learning models must also be reliable in settings comparable to real-world use in order to be adopted by stakeholders (72). The evaluation metrics used should closely represent the needs of stakeholders and allow a more granular breakdown of model performance (66; 11) - for example, the model's ability to capture yield variability in years with climate extremes must be reported (77). Finally, to avoid overestimation of model skill, the evaluation procedure must take into account the specific challenges arising from the use of non-i.i.d spatiotemporal data (40; 64; 26).

We present CY-Bench, a comprehensive dataset and benchmark for sub-national crop yield forecasting, with coverage of major crop-growing countries across the world for maize and wheat. Here, sub-national refers to the administrative levels for which official crop statistics are published; crop yield refers to the end-of-season yield reported in the statistics; and forecasting refers to the production of end-of-season yield estimates with a certain lead time before harvest (e.g. mid-season or 30 days before harvest) or before the publication of official statistics. Thus, the dataset combines sub-national yield statistics with relevant predictors, such as growing-season weather indicators, remote sensing indicators, evapotranspiration, soil moisture indicators, and static soil properties. CY-Bench has been designed and curated by agricultural experts, climate scientists, and machine learning researchers from the AgML community (`https://www.agml.org/`), with the aim of facilitating model intercomparison across the diverse agricultural systems around the globe in conditions as close as possible to real-world operationalization. Ultimately, by lowering the barrier to entry for ML researchers in this crucial application area, CY-Bench will facilitate the development of improved crop forecasting tools that can be used to support decision-makers in food security planning worldwide.

2 Related work

Crop yields are commonly forecast using weather, soil, moisture and crop productivity or remote-sensing-derived vegetation health indicators as predictors. Methods used include field surveys, process-based crop models, statistical regression and machine learning (5; 58). Data-driven approaches are appealing as they can capture processes not yet well-covered by biophysical crop models, but typically require access to predictor data and yield data over large areas and spanning multiple years. The availability of these datasets determines the type of yield forecasting setup, which can range from national and sub-national level to field level. For example, the European Commission's Joint Research Centre (EC-JRC) regularly produces national crop yield forecasts for the EU and surrounding countries using crop models, agro-meteorological analyses and expertise of analysts (72). Sub-national yield forecasting utilizes data for a large number of sub-national administrative units (e.g. regions, provinces) typically collected by national statistical offices and captures spatial yield variability within a country (38; 44), which is crucial for targeted food security planning.

An increasing number of publications have demonstrated excellent performance of a diverse range of machine learning approaches for crop yield forecasting (34; 35; 49; 76; 45). Unfortunately, while results suggest that machine learning methods have great potential for providing accurate and timely crop yield forecasts, the datasets used by previous studies are, in most cases, unpublished. This has prevented the community from reproducing their results or comparing the strengths and weaknesses of different methods across different crops and regions. To our knowledge, SustainBench (78), which curates multi-source data for various tasks spanning the United Nations' seven sustainable development goals, includes a benchmark dataset designed to measure the performance of machine learning models for crop yield prediction. However, it targets end-of-season prediction for only one crop (soybean) in three countries (United States, Brazil and Argentina) and uses a relatively small set of predictors. Another public dataset is CropNet (33), which only includes the United States. Similarly, there are ongoing efforts (82) to produce a multi-task benchmark dataset which includes yield prediction in the USA as a sub-task. Apart from these, other available data contributions include yield statistics only (15; 48; 47; 54; 3; 4; 10; 13; 24; 39) or have been made available in combination with predictor data published with existing studies (28; 21; 43; 45) but are not explicitly tailored for yield forecasting benchmarking studies.

In comparison, CY-Bench data covers forty-two countries across six continents. This enables a comprehensive evaluation of model performance across regions with heterogeneous agricultural practices and infrastructure, including developing countries which are generally under-represented in machine learning benchmarks. Furthermore, we closely mimic real-world operationalization settings in the predictor data used, data pre-processing steps and evaluation set-up, including the use of temporal Leave-One-Year-Out validation (as opposed to the random sampling methods used in SustainBench and multiple previous studies). This means that novel machine learning methods which achieve excellent performance on CY-Bench could be used to improve yield forecasting systems in practice, providing accurate and timely information urgently needed by decision-makers.

Although we have identified a distinct lack of benchmark datasets for agricultural yield forecasting, there have been many recent developments in the related field of crop type mapping using satellite data (55; 69; 78; 29), leading to exciting progress in the development of methods for extracting meaningful patterns from time series of earth observation data (56; 55; 46; 57). Other related work (70; 68; 27) has been able to exploit meta-learning and multitask learning to improve model performance for land cover classification, crop mapping and agricultural yield forecasting. While CY-Bench is focused on pre-harvest yield forecasting, the dataset includes time series of crop productivity or vegetation health indicators from earth observation as predictors, and can therefore be easily combined with existing crop mapping benchmark datasets to explore such approaches.

3 CY-Bench task and datasets

3.1 Task

CY-Bench is designed to evaluate model performance for in-season crop yield forecasting at sub-national level. Forecasts are generated for selected crops (maize and wheat) at different time points, based on stakeholder needs (e.g. mid-season, a quarter of the season, or a certain number of days before harvest). For this exercise, we only report forecasts generated mid-season, the timing of which can differ by location. Mid-season was selected because peak model performance is typically reached around the mid-point of the growing season. This mid-point is also when the transition from vegetative to reproductive growth stage happens for most crops (30; 5). Season length and mid-season information is derived from crop calendars. As in the operational setting, models must forecast the end-of-season crop yield outcomes based on the available time series data only up until the designed lead time.

3.2 Dataset overview

Agricultural yield data. The CY-Bench dataset includes crop statistics from twenty-nine countries for wheat and forty-two countries for maize (see Figure 1, 2). Models are trained to predict official crop yield statistics for sub-national administrative levels, which are obtained from national statistics offices (e.g. National Agricultural Statistics Service of the United States Department of Agriculture) or regional agencies (e.g. Eurostat and FEWSNET). Details of each source are indicated in the data preparation section in GitHub. Depending on the country, the term 'sub-national' can refer to administrative division 1 (province, state, region), division 2 (district), or division 3 (county, municipality, commune). When statistics for multiple administrative levels are available, we select the highest resolution.

Predictor data. CY-Bench predictor data includes static soil properties and time series of weather variables, soil moisture indicators and vegetation indicators (Table 1). Soil data comes from the WISE Soil database (6), weather variables from AgERA5 (9), potential or reference evapotranspiration (ET0) from FAO-AQUASTAT (2), soil moisture indicators from GLDAS (52), vegetation indicators (fraction of absorbed photosynthetically active radiation (FPAR) and normalized difference vegetation index (NDVI)) from EC-JRC and MODIS MOD09CMG respectively (62; 75). Predictor data and yield statistics often differ in spatial and temporal resolution, requiring further processing to align them effectively. Weather, ET0 and soil moisture data come in daily time steps. FPAR comes in dekadal time step, with three values per month (days 1-10, 11-20, 21-31). NDVI data is available approximately every week, but the dates are not regular. Predictor data is filtered using crop-type maps (or crop masks) from EC-JRC (18), which are derived from the WorldCereal project (74). This step restricts predictor data to pixels in harvested crop areas only. When crop masks and predictor data differ in resolution, the crop mask is resampled to the resolution of the predictor data. After masking, predictor data is aggregated to match the boundaries and spatial level of the yield data according to the administrative level (Figure 3). Additionally, as the sensitivity of crops vary throughout the phenological cycle, time series predictor data must correspond to the growing season. As this depends on the specific crop, management practices, and location, crop calendar information such as the start and end of season is required. In CY-Bench, these crop calendars are obtained from the WorldCereal project (22).

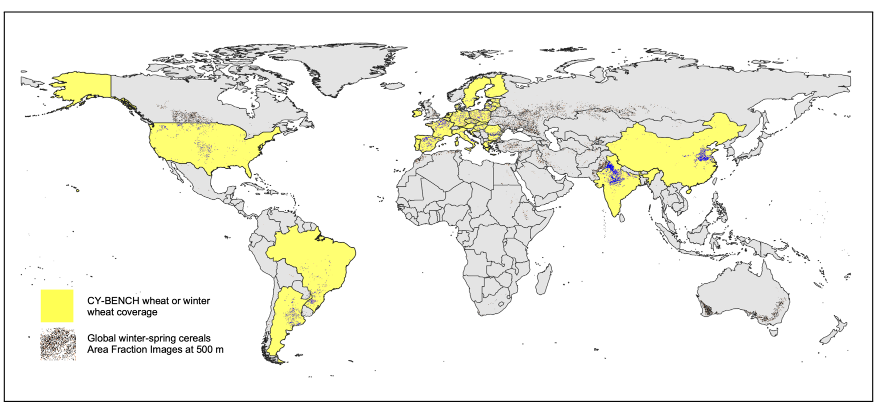

Figure 1: A map of the countries covered by CY-Bench for wheat yield forecasting. CY-Bench has coverage in 31 countries in total.

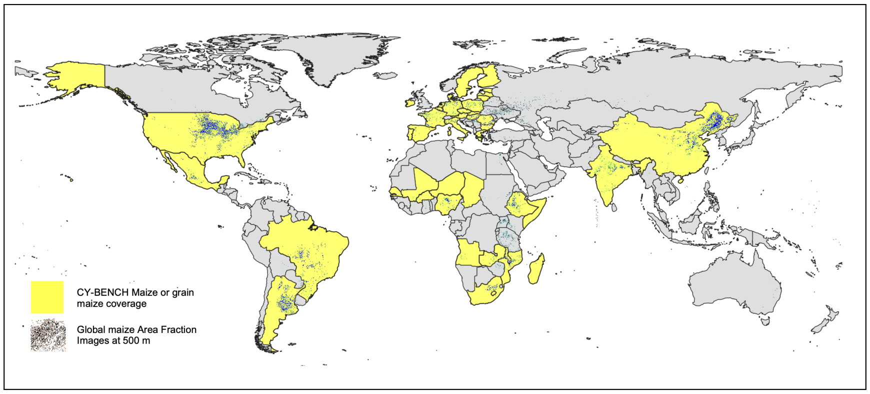

Figure 2: A map of countries covered by CY-Bench for maize yield forecasting. CY-Bench has coverage in 42 countries in total.

Table 1: Overview of the predictor data, crop mask and crop calendar. NDVI refers to the normalized difference vegetation index, FPAR is the fraction of absorbed photosynthetically active radiation and AWC is the available water capacity.

Category	**Data**		**Spatial resolution**	**Temporal resolution**	**Source**
	Name	**Unit**			
Meteorological	temperature precipitation solar radiation	°C mm Jm^{-2}	0.1°	daily	AgERA5 (9)
	evapotranspiration	mm	0.1°	daily	AQUASTAT-FAO (2)
Vegetation	FPAR NDVI	% -	500m 5000m	10-days 8-days	JRC (62) MOD09CMG (75)
Soil	AWC bulk density drainage class	cm m^{-1} kg dm^{-3} -	30”	static	WISE (6)
	moisture content	kg m^{-2}	0.25°	daily	NASA GLDAS (52)
Crop	crop mask crop calendar	-	0.5°	-	Crop masks (74; 18) Crop calendars (22)

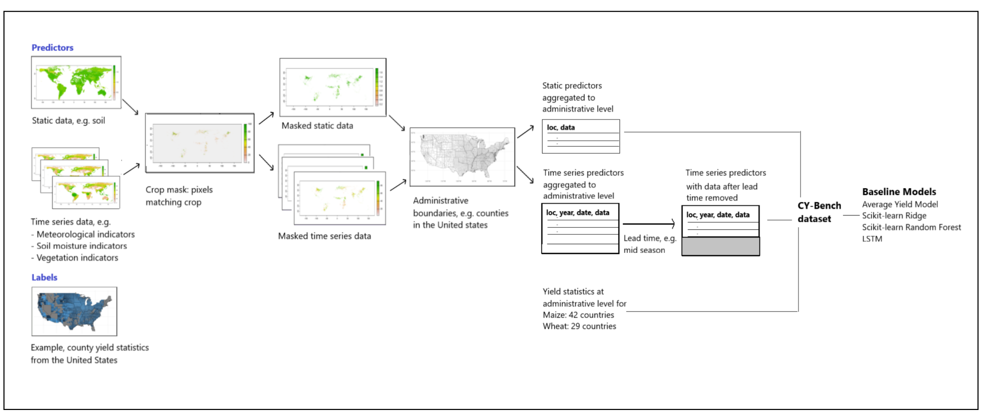

Figure 3: Overview of the CY-Bench data preparation process.

For deep learning models, such as Long Short Term Memory networks (LSTM), time series data is aggregated to dekadal time steps (days 1-10, 11-20, 21-31, and so on), which allows all datapoints to have the same number of time steps and therefore fixed input dimension.

For tree-based models and other machine learning models which are designed for tabular data, the time series data is aggregated in the temporal dimension to create domain-relevant features. These include monthly averages of minimum daily temperature ($tmin$), maximum daily temperature ($tmax$), average daily temperature, daily precipitation ($prec$), cumulative climatic water balance (prec - ET0) and surface soil moisture. Similarly, monthly maximum values were calculated for cumulative growing degree days (GDD), cumulative precipitation, cumulative FPAR and cumulative NDVI. Furthermore, we calculated the number of days in which $tmin$ was less than 0 degree Celsius ('cold days'), days in which $tmax$ was greater than 35 degrees Celsius ('hot days') and days where $prec$ was less than 1 mm ('dry days').

Dataset access. The dataset is available in Google Drive. A python library has been developed to load the datasets and run CY-Bench.

Table 2: Maize NRMSE per model for Argentina (AR), Brazil (BR), China (CN), Germany (DE), France (FR) and the United States (US).

Model	AR	BR	CN	DE	FR	US
LSTM	87.206	42.352	20.584	13.778	21.967	23.962
Naive	**33.514**	**33.284**	**9.384**	14.838	**16.860**	**18.101**
RF	49.544	45.538	13.767	**14.227**	18.549	19.391
Ridge	152.41	64.363	48.245	48.798	23.043	22.443

Table 3: Wheat NRMSE per model for Argentina (AR), Brazil (BR), China (CN), Germany (DE), France (FR) and the United States (US).

Model	AR	BR	CN	DE	FR	US
LSTM	36.440	33.137	99.883	14.782	17.540	35.700
Naive	**24.349**	**28.008**	**10.808**	**9.941**	**9.546**	**19.410**
RF	32.941	31.059	45.804	15.490	17.323	39.305
Ridge	31.061	31.737	351.01	60.968	65.382	29.093

4 Model evaluation and baselines

In CY-Bench, models are trained per country and per crop, and evaluated using Leave One Year Out (LOYO) evaluation. The motivation for LOYO is to obtain a robust estimate of the performance of algorithms on both average and extreme harvest years. As each season can vary substantially from previous years, measurement of predictive performance on only the current season or the most recent year may under- or over-estimate the forecasting ability of a model. For more information regarding model evaluation strategies in the context of agriculture see ((51)).

We evaluate the performance of four baseline models. First, the Average Yield model (*Naive*) predicts the average of the training set by administrative region (if present in training data) or country (if absent in training data). Second, the Ridge model (implemented in Scikit-Learn) builds a linear model using features designed as described in the previous sub-section. Third, Random Forest is used (also implemented in Scikit-Learn), which is frequently used for agricultural machine learning studies. Finally, we include LSTM as a baseline for representation learning from time series data.

As our evaluation metrics, we use the normalized root mean squared error (NRMSE; i.e., the root mean squared error normalized by the average yield of the test set), and mean absolute percentage error (MAPE). NRMSE and MAPE are reported by averaging over all cross-validation test folds (which covers the complete dataset for LOYO) and all admin regions with a country. Additionally, metrics and box plots describing model performance for each year individually are included in the Supplement.

We report results of the baseline model benchmarks in figures 4 and 5 for maize and wheat, respectively, to show NRMSE of different countries and baseline models. Moreover, we report median NRMSE of select countries for each model in tables 2 and 3 for maize and wheat, respectively.

The results show that the *Naive* model outperforms all the other baseline models, except for Random Forests. The *Naive* model is a test of prediction skill. The performance of most machine learning models shows the difficulty of generalizing from the training set.

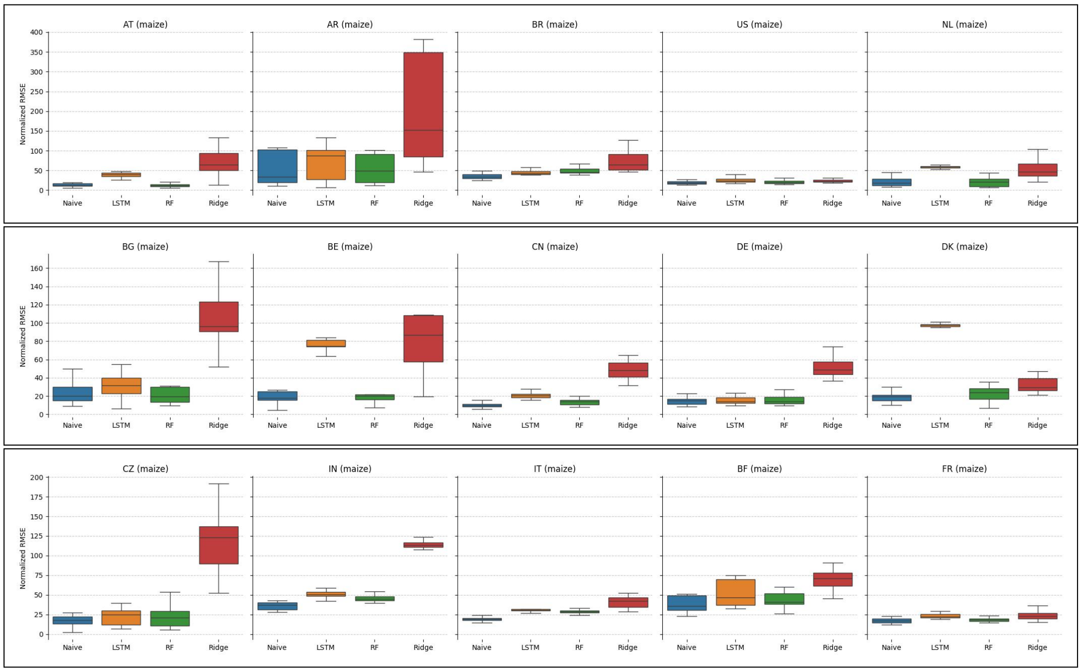

Figure 4: NRMSE for maize, predicted at mid-season lead time.

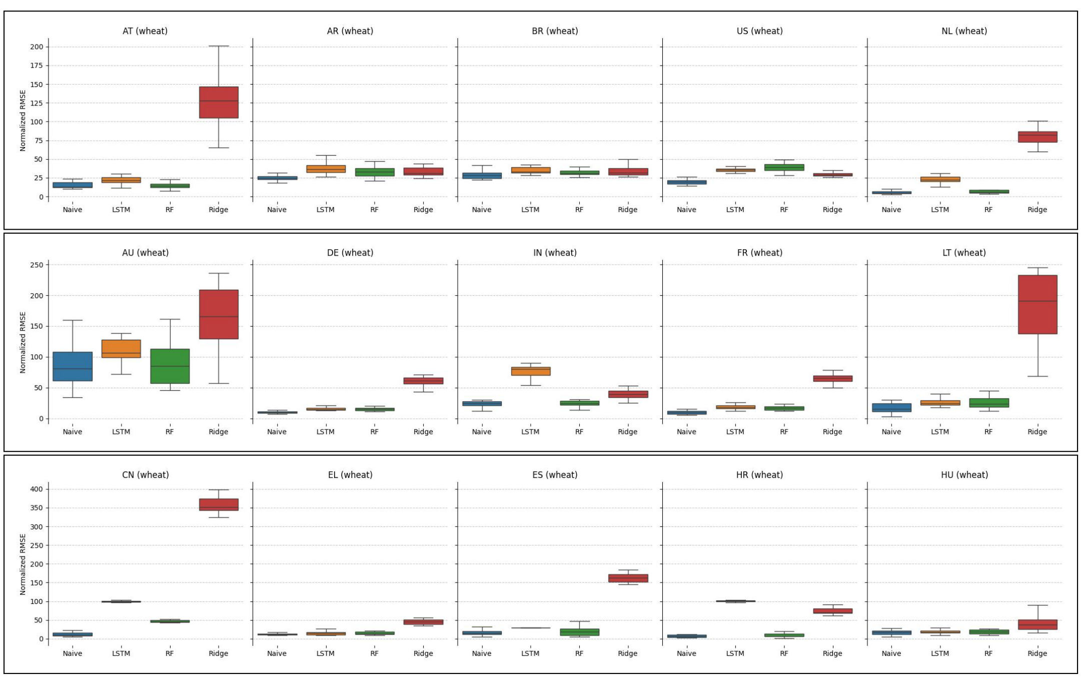

Figure 5: NRMSE for wheat, predicted at mid-season lead time.

5 Contributions, limitations and future work

In addition to the relevance for climate change, food security and United Nations' sustainable development goals, CY-Bench dataset is relevant to the ML community due to its comprehensive geographic coverage, capturing diverse agricultural practices and conditions. The inclusion of high-resolution satellite imagery, weather data, and soil properties provides a rich, heterogeneous dataset that presents numerous opportunities for the development of innovative machine learning methods. An inherent challenge of agricultural data, and crop-yield forecasting specifically, is the difficult and high level of domain knowledge required in collecting and processing the various data types and defining the task. This analysis-ready dataset is accessible to ML modelers who do not necessarily have to be experts in yield forecasting, lowering the barrier to entry for advanced yield forecasting research and fostering broader participation and innovation in the field. Beyond academic research, this dataset can significantly impact policy-making, agricultural planning, and disaster response by enabling the robust evaluation and development of operationalizable models. Researchers, policymakers, farmers, and agribusinesses can all benefit from the insights derived from this dataset, leading to better-informed decisions and improved agricultural outcomes.

Apart from the downstream task of in-season yield forecasting, CY-Bench enables explorations in transfer learning, domain adaptation, and representation learning. Researchers can leverage this dataset to assess if models are able to generalize well across diverse geographic and climatic conditions. While in this paper we focus on forecasting crop yields by training individual models for each crop and country, the dataset allows for a more integrated approach. We envision at least four directions for future research. First, transfer learning methods can be explored to improve model generalisation ability when training models on data from a data-rich region and deploying the forecasting model to data-sparse regions. Second, self-supervised learning could be used to harness the vast amounts of unlabeled agricultural data available. By training models to recognize patterns and structures within this data, we can build robust representations that capture essential features of the agricultural system. These representations can then be fine-tuned using the labeled datasets in CY-Bench specific to each country or crop. For instance, a self-supervised model trained on satellite images and environmental data can later be fine-tuned to predict specific crop yields in various regions, making it a powerful tool for global agricultural analysis. Third, another important area is to explore the stability of model predictions against natural and human interventions. This involves understanding how factors like extreme weather events, policy changes, or management practices impact yield forecasts. Causal invariant learning focuses on identifying and utilizing stable variables across different environments to ensure robustness and generalization. For example, soil quality and basic climatic factors like temperature and precipitation may have stable relationships with crop yields. By recognizing variables that consistently impact crop yields regardless of geographic or climatic differences, it may be possible to build models that are resilient to distributional shifts and perform reliably across diverse conditions. Fourth, deep learning techniques, such as autoencoders, can be employed to learn compact and informative representations of the input data, potentially uncovering latent variables that are more directly related to crop yields. This could improve the model's ability to generalize and perform well across different regions and conditions, while possibly giving scientific insight into the underlying drivers of agricultural crop yields.

We would like to also highlight several limitations and areas for improvement in future iterations of CY-Bench. First, some limitations stem from the data sources selected. The predictors do not capture certain factors that influence end-of-season yields, such as pests, diseases and farm management choices. Similarly, CY-Bench does not include socioeconomic factors such as market prices, labor availability, and policy changes. Including these variables could provide a more holistic understanding of yield fluctuations and help in developing more robust models. Additionally, our modeling setup does not differentiate between irrigated and non-irrigated systems. These systems can exhibit different responses to predictors due to varying water availability, leading to potential inaccuracies in yield forecasts. Our choice was driven by the fact that crop statistics in most countries are rarely reported separately for irrigated and non-irrigated areas. Second, CY-Bench does not supply process-based crop model outputs, which could be used as inputs to machine learning models, and features are aggregated in fixed time steps, rather than being designed according to the stage of crop growth and development. Access to crop model outputs could provide information on key phenological state changes, which can be useful to design more predictive features. Third, crop yield forecasting models could benefit from incorporating weather forecasts. In our setup, models cannot access data after the lead time and, therefore, cannot capture conditions that might affect the end-of-season yields after that point. In the real-life setting, forecasters would have access to weather forecasts that may provide useful information. Finally, the LOYO method of evaluation is used due to small data sizes in many countries. This approach assumes that all years are independent, which may be too strong of an assumption if consecutive years have correlated environmental and climatic conditions.

6 Conclusion

Innovative data-driven solutions will be crucial to achieve the United Nations' Sustainable Development Goal 2 of Zero Hunger (61). By providing consistent evaluation of large-scale crop yield forecasts, CY-Bench is a step forward in bridging the gap between agricultural modeling and machine learning community. Curated by an interdisciplinary group of experts in agronomy, food security, climate science and agriculture, this dataset can facilitate increased collaboration between fields, and ultimately help to produce reliable crop yield forecasts to support decisions of farmers, policymakers and commodity traders worldwide.

Acknowledgements

CY-Bench benefited from many helpful discussions with the participants of AgML, the Machine Learning team of the Agricultural Model Intercomparison and Improvement Project (AgMIP). We would also like to acknowledge the contributions of Marc Russwurm, Afef Marzougui, Hendrik Boogaard, Marijn van der Velde, Steven Hoek, Filip Szabo, Francesco Collivignarelli, Xiaomao Lin, Toshi Iizumi, Peng Fu, Paresh Shirsath, Soora Naresh Kumar, Sibiri Traore and Javier Garcia Navarro in the design and implementation of the benchmark and preparation of the manuscript.

References

[1] Ramya Ambikapathi, Kate R. Schneider, Benjamin Davis, Mario Herrero, Paul Winters, and Jessica C. Fanzo. Global food systems transitions have enabled affordable diets but had less favourable outcomes for nutrition, environmental health, inclusion and equity. *Nature Food*, 3 (9):764–779, 2022. ISSN 2662-1355. doi: 10.1038/s43016-022-00588-7.

[2] FAO AQUASTAT. Reference evapotranspiration - AgERA5 derived (Global - Daily - 10km). `https://data.apps.fao.org/catalog//iso/f22813e9-679e-4864-bd92-d48f5dfc436c`, 2021. Accessed: 2024-06-05.

[3] Argentina. Ministerio de agrícultura, ganaderia y pesca. "estimaciones agrícolas". `https://datosestimaciones.magyp.gob.ar/reportes.php?reporte=Estimaciones`, 2016. Accessed: 2016-04-29.

[4] Australia. Abares. australian bureau of agricultural and resource economics and sciences farm data portal. `https://www.agriculture.gov.au/abares/data/farm-data-portal#data-download`, 2024. Accessed: 2024-03-05.

[5] Bruno Basso and Lin Liu. Seasonal crop yield forecast: Methods, applications, and accuracies. In *Advances in Agronomy*, volume 154, pages 201–255. Elsevier, 2019. doi: 10.1016/bs.agron.2018.11.002.

[6] Niels H Batjes. Harmonized soil property values for broad-scale modelling (WISE30sec) with estimates of global soil carbon stocks. *Geoderma*, 269:61–68, 2016. doi: 10.1016/j.geoderma.2016.01.034.

[7] Inbal Becker-Reshef, Christina Justice, Brian Barker, Michael Humber, Felix Rembold, Rogerio Bonifacio, Mario Zappacosta, Mike Budde, Tamuka Magadzire, Chris Shitote, Jonathan Pound, Alessandro Constantino, Catherine Nakalembe, Kenneth Mwangi, Shinichi Sobue, Terence Newby, Alyssa Whitcraft, Ian Jarvis, and James Verdin. Strengthening agricultural decisions in countries at risk of food insecurity: The geoglam crop monitor for early warning. *Remote Sensing of Environment*, 237:111553, 2020. doi: 10.1016/j.rse.2019.111553.

[8] Lefteris Benos, Aristotelis C Tagarakis, Georgios Dolias, Remigio Berruto, Dimitrios Kateris, and Dionysis Bochtis. Machine learning in agriculture: A comprehensive updated review. *Sensors*, 21(11):3758, 2021. doi: 10.3390/s21113758.

[9] H. Boogaard, J. Schubert, A. De Wit, J. Lazebnik, R. Hutjes, and G. Van der Grijn. Agrometeorological indicators from 1979 to present derived from reanalysis. Climate Data Store - Copernicus Climate Change Service, `https://doi.org/10.24381/cds.6c68c9bb`, 2022.

[10] Brazil. Ibge sidra. "tabela 1612: Área plantada, área colhida, quantidade produzida, rendimento médio e valor da produção das lavouras temporárias". `https://sidra.ibge.gov.br/tabela/1612`, 2022. Accessed: 2024-02-06.

[11] Ryan Burnell, Wout Schellaert, John Burden, Tomer D. Ullman, Fernando Martinez-Plumed, Joshua B. Tenenbaum, Danaja Rutar, Lucy G. Cheke, Jascha Sohl-Dickstein, Melanie Mitchell, Douwe Kiela, Murray Shanahan, Ellen M. Voorhees, Anthony G. Cohn, Joel Z. Leibo, and Jose Hernandez-Orallo. Rethink reporting of evaluation results in AI. *Science*, 380(6641):136–138, 2023. doi: 10.1126/science.adf6369.

[12] Bowen Chen and Nelson B. Villoria. Foreign yield shocks and domestic price variability: the case of maize in developing countries. *Environmental Research Letters*, 17(12):124044, 2022. ISSN 1748-9326. doi: 10.1088/1748-9326/aca7d5.

[13] China. National bureau of statistics of china. national data portal. `https://data.stats.gov.cn`, 2024. Accessed: 2024-02-18.

[14] Anna Chlingaryan, Salah Sukkarieh, and Brett Whelan. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and Electronics in Agriculture*, 151:61–69, 2018. doi: 10.1016/j.compag.2018.05.012.

[15] Christoph Duden, Christina Nacke, and Frank Offermann. German yield and area data for 11 crops from 1979 to 2021 at a harmonized spatial resolution of 397 districts. *Scientific Data*, 11, 01 2024. doi: 10.1038/s41597-024-02951-8.

[16] Peter D. Dueben, Martin G. Schultz, Matthew Chantry, David John Gagne, David Matthew Hall, and Amy McGovern. Challenges and Benchmark Datasets for Machine Learning in the Atmospheric Sciences: Definition, Status, and Outlook. *Artificial Intelligence for the Earth Systems*, 1(3), 2022. ISSN 2769-7525. doi: 10.1175/AIES-D-21-0002.1.

[17] Oumnia Ennaji, Leonardus Vergutz, and Achraf El Allali. Machine learning in nutrient management: A review. *Artificial Intelligence in Agriculture*, 9:1–11, 2023. doi: 10.1016/j.aiia.2023.06.001.

[18] European Commission, Joint Research Centre, 2024. Elaboration of Van Tricht et al, 2023.

[19] Joshua Fan, Junwen Bai, Zhiyun Li, Ariel Ortiz-Bobea, and Carla Gomes. A gnn-rnn approach for harnessing geospatial and temporal information: Application to crop yield prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36:11873–11881, 06 2022. doi: 10.1609/aaai.v36i11.21444.

[20] Jessica Fanzo. Kicking around in the wreck. *PLOS Climate*, 3(4):e0000401, April 2024. ISSN 2767-3200. doi: 10.1371/journal.pclm.0000401.

[21] Ruben Fernandez-Beltran, Tina Baidar, Jian Kang, and Filiberto Pla. Rice-Yield Prediction with Multi-Temporal Sentinel-2 Data and 3D CNN: A Case Study in Nepal. *Remote Sensing*, 13(7):1391, 2021.

[22] Belén Franch, Juanma Cintas, Inbal Becker-Reshef, María José Sanchez-Torres, Javier Roger, Sergii Skakun, José Antonio Sobrino, Kristof Van Tricht, Jeroen Degerickx, Sven Gilliams, et al. Global crop calendars of maize and wheat in the framework of the worldcereal project. *GIScience & Remote Sensing*, 59(1):885–913, 2022. doi: 10.1080/15481603.2022.2079273.

[23] Yohanne Larissa Gavasso-Rita, Simon Michael Papalexiou, Yanping Li, Amin Elshorbagy, Zhenhua Li, and Corinne Schuster-Wallace. Crop models and their use in assessing crop production and food security: A review. *Food and Energy Security*, 13(1):e503, 2023. ISSN 2048-3694. doi: 10.1002/fes3.503.

[24] India. Icrisat. district level database. `http://data.icrisat.org/dld/src/crops.html`, 2024. Accessed: 2024-02-09.

[25] Andreas Kamilaris and Francesc X Prenafeta-Boldú. Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, 147:70–90, 2018. doi: 10.1016/j.compag.2018.02.016.

[26] Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023. ISSN 2666-3899. doi: 10.1016/j.patter.2023.100804.

[27] Hannah Kerner, Gabriel Tseng, Inbal Becker-Reshef, Catherine Nakalembe, Brian Barker, Blake Munshell, Madhava Paliyam, and Mehdi Hosseini. Rapid response crop maps in data sparse regions. In *ACM SIGKDD Conference on Data Mining and Knowledge Discovery Workshops*, 2020. doi: https://doi.org/10.48550/arXiv.2006.16866.

[28] Saeed Khaki, Lizhi Wang, and Sotirios V Archontoulis. A CNN-RNN framework for crop yield prediction. *Frontiers in Plant Science*, 10:1750, 2020. doi: 10.3389/fpls.2019.01750.

[29] Lukas Kondmann, Aysim Toker, Marc Ruß wurm, Andrés Camero, Devis Peressuti, Grega Milcinski, Pierre-Philippe Mathieu, Nicolas Longepe, Timothy Davis, Giovanni Marchisio, Laura Leal-Taixé, and Xiaoxiang Zhu. Denethor: The dynamicearthnet dataset for harmonized, inter-operable, analysis-ready, daily crop monitoring from space. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran, 2021. URL `https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/5b8add2a5d98b1a652ea7fd72d942dac-Paper-round2.pdf`.

[30] Donghoon Lee, Frank Davenport, Shraddhanand Shukla, Greg Husak, Chris Funk, Laura Harrison, Amy McNally, James Rowland, Michael Budde, and James Verdin. Maize yield forecasts for Sub-Saharan Africa using Earth Observation data and machine learning. *Global Food Security*, 33:100643, 2022. doi: 10.1016/j.gfs.2022.100643.

[31] Corey Lesk, Weston Anderson, Angela Rigden, Onoriode Coast, Jonas Jägermeyr, Sonali McDermid, Kyle F. Davis, and Megan Konar. Compound heat and moisture extreme impacts on global crop yields under climate change. *Nature Reviews Earth & Environment*, 3(12):872–889, 2022. ISSN 2662-138X. doi: 10.1038/s43017-022-00368-8.

[32] Konstantinos Liakos, Patrizia Busato, Dimitrios Moshou, Simon Pearson, and Dionysis Bochtis. Machine learning in agriculture: A review. *Sensors*, 18(8):2674, 2018. doi: 10.3390/s18082674.

[33] Fudong Lin, Kaleb Guillot, Summer Crawford, Yihe Zhang, Xu Yuan, and Nianfeng Tzeng. Cropnet: An open large-scale dataset with multiple modalities for climate change-aware crop yield predictions, 2024. URL `https://openreview.net/forum?id=lzpHNyhIbr`.

[34] Qinqing Liu, Meijian Yang, Koushan Mohammadi, Dongjin Song, Jinbo Bi, and Guiling Wang. Machine Learning Crop Yield Models Based on Meteorological Features and Comparison with a Process-Based Model. *Artificial Intelligence for the Earth Systems*, 1(4), 2022. ISSN 2769-7525. doi: 10.1175/AIES-D-22-0002.1.

[35] Y. Ma, Z. Yang, Q. Huang, and Z. Zhang. Improving the Transferability of Deep Learning Models for Crop Yield Prediction: A Partial Domain Adaptation Approach. *Remote Sensing*, 15(18), 2023. ISSN 2072-4292. doi: 10.3390/rs15184562.

[36] Anna Mateo-Sanchis, Maria Piles, Julia Amorós-López, Jordi Muñoz-Marí, Jose E Adsuara, Álvaro Moreno-Martínez, and Gustau Camps-Valls. Learning main drivers of crop progress and failure in Europe with interpretable machine learning. *International Journal of Applied Earth Observation and Geoinformation*, 104:102574, 2021. doi: 10.1016/j.jag.2021.102574.

[37] Zia Mehrabi, Ruth Delzeit, Adriana Ignaciuk, Christian Levers, Ginni Braich, Kushank Bajaj, Araba Amo-Aidoo, Weston Anderson, Roland A. Balgah, Tim G. Benton, Martin M. Chari, Erle C. Ellis, Narcisse Z. Gahi, Franziska Gaupp, Lucas A. Garibaldi, James S. Gerber, Cecile M. Godde, Ingo Grass, Tobias Heimann, Mark Hirons, Gerrit Hoogenboom, Meha Jain, Dana James, David Makowski, Blessing Masamha, Sisi Meng, Sathaporn Monprapussorn, Daniel Müller, Andrew Nelson, Nathaniel K. Newlands, Frederik Noack, MaryLucy Oronje, Colin Raymond, Markus Reichstein, Loren H. Rieseberg, Jose M. Rodriguez-Llanes, Todd Rosenstock, Pedram Rowhani, Ali Sarhadi, Ralf Seppelt, Balsher S. Sidhu, Sieglinde Snapp, Tammara Soma, Adam H. Sparks, Louise Teh, Michelle Tigchelaar, Martha M. Vogel, Paul C. West, Hannah Wittman, and Liangzhi You. Research priorities for global food security under extreme events. *One Earth*, 5(7):756–766, 2022. ISSN 2590-3322. doi: 10.1016/j.oneear.2022.06.008.

[38] Michele Meroni, François Waldner, Lorenzo Seguini, Hervé Kerdiles, and Felix Rembold. Yield forecasting with machine learning and small data: What gains for grains? *Agricultural and Forest Meteorology*, 308:108555, 2021. doi: 10.1016/j.agrformet.2021.108555.

[39] Mexico. Inegi. agricultural census and survey data. `https://www.inegi.org.mx`, 2019. Accessed: 2024-04-10.

[40] Hanna Meyer and Edzer Pebesma. Machine learning-based global maps of ecological variables and the challenge of assessing them. *Nature Communications*, 13(1):2208, 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-29838-9.

[41] Catherine Nakalembe and Hannah Kerner. Considerations for AI-EO for agriculture in Sub-Saharan Africa. *Environmental Research Letters*, 18(4):041002, 2023. ISSN 1748-9326. doi: 10.1088/1748-9326/acc476.

[42] Alexandros Oikonomidis, Cagatay Catal, and Ayalew Kassahun. Deep learning for crop yield prediction: a systematic literature review. *New Zealand Journal of Crop and Horticultural Science*, pages 1–26, 2022. doi: 10.1080/01140671.2022.2032213.

[43] Dilli Paudel, Hendrik Boogaard, Allard de Wit, Sander Janssen, Sjoukje Osinga, Christos Pylianidis, and Ioannis N Athanasiadis. Machine learning for large-scale crop yield forecasting. *Agricultural Systems*, 187:103016, 2021. doi: 10.1016/j.agsy.2020.103016.

[44] Dilli Paudel, Hendrik Boogaard, Allard de Wit, Marijn van der Velde, Martin Claverie, Luigi Nisini, Sander Janssen, Sjoukje Osinga, and Ioannis N Athanasiadis. Machine learning for regional crop yield forecasting in Europe. *Field Crops Research*, 276:108377, 2022. doi: 10.1016/j.fcr.2021.108377.

[45] Dilli R. Paudel, Diego Marcos, Allard de Wit, Hendrik Boogaard, and Ioannis N. Athanasiadis. A weakly supervised framework for high resolution crop yield forecasts. *Environmental Research Letters*, 18(9):094062, 2023. doi: 10.1088/1748-9326/acf50e.

[46] Charlotte Pelletier, Geoffrey I. Webb, and François Petitjean. Temporal convolutional neural network for the classification of satellite image time series. *Remote Sensing*, 11(5), 2019. ISSN 2072-4292. doi: 10.3390/rs11050523. URL `https://www.mdpi.com/2072-4292/11/5/523`.

[47] Nicholas Potter. *rnassqs: Access the NASS 'Quick Stats' API*, 2022. URL `https://CRAN.R-project.org/package=rnassqs`. R package version 0.6.1.

[48] Nicholas A Potter. rnassqs: An ‘r‘ package to access agricultural data via the usda national agricultural statistics service (usda-nass) 'quick stats' api. *The Journal of Open Source Software*, aug 2019.

[49] Rhorom Priyatikanto, Yang Lu, Jadu Dash, and Justin Sheffield. Improving generalisability and transferability of machine-learning-based maize yield prediction model through domain adaptation. *Agricultural and Forest Meteorology*, 341:109652, October 2023. ISSN 0168-1923. doi: 10.1016/j.agrformet.2023.109652.

[50] Megan Richards, Polina Kirichenko, Diane Bouchacourt, and Mark Ibrahim. Does Progress On Object Recognition Benchmarks Improve Real-World Generalization?, 2023.

[51] Jonathan Richetti, Foivos I Diakogianis, Asher Bender, André F Colaço, and Roger A Lawes. A methods guideline for deep learning for tabular data in agriculture with a case study to forecast cereal yield. *Computers and Electronics in Agriculture*, 205:107642, 2023. doi: 10.1016/j.compag.2023.107642.

[52] Matthew Rodell, PR Houser, UEA Jambor, J Gottschalck, Kieran Mitchell, C-J Meng, Kristi Arsenault, B Cosgrove, J Radakovich, M Bosilovich, et al. The global land data assimilation system. *Bulletin of the American Meteorological society*, 85(3):381–394, 2004. doi: 10.1175/BAMS-85-3-381.

[53] David Rolnick, Alan Aspuru-Guzik, Sara Beery, Bistra Dilkina, Priya L. Donti, Marzyeh Ghassemi, Hannah Kerner, Claire Monteleoni, Esther Rolf, Milind Tambe, and Adam White. Application-Driven Innovation in Machine Learning, 2024.

[54] Giulia Ronchetti, Luigi Nisini Scacchiafichi, Lorenzo Seguini, Iacopo Cerrani, and Marijn van der Velde. Harmonized european union subnational crop statistics can reveal climate impacts and crop cultivation shifts. *Earth System Science Data*, 16(3):1623–1649, 2024. doi: 10.5194/essd-16-1623-2024.

[55] Marc Rußwurm, Sébastien Lefèvre, and Marco Körner. Breizhcrops: A satellite time series dataset for crop type identification. *ArXiv*, 2019. doi: 10.48550/arXiv.1905.11893.

[56] Marc Rußwurm and Marco Körner. Multi-temporal land cover classification with sequential recurrent encoders. *ISPRS International Journal of Geo-Information*, 7(4), 2018. ISSN 2220-9964. doi: 10.3390/ijgi7040129. URL `https://www.mdpi.com/2220-9964/7/4/129`.

[57] Vivien Sainte Fare Garnot, Loic Landrieu, Sebastien Giordano, and Nesrine Chehata. Satellite image time series classification with pixel-set encoders and temporal self-attention. *CVPR*, 2020.

[58] Bernhard Schauberger, Jonas Jägermeyr, and Christoph Gornott. A systematic review of local to regional yield forecasting approaches and frequently used data resources. *European Journal of Agronomy*, 120:126153, 2020. doi: 10.1016/j.eja.2020.126153.

[59] Wolfram Schlenker and Michael J. Roberts. Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. *Proceedings of the National Academy of Sciences*, 106(37):15594–15598, 2009. doi: 10.1073/pnas.0906865106.

[60] Kate R. Schneider, Jessica Fanzo, Lawrence Haddad, Mario Herrero, Jose Rosero Moncayo, Anna Herforth, Roseline Remans, Alejandro Guarin, Danielle Resnick, Namukolo Covic, Christophe Béné, Andrea Cattaneo, Nancy Aburto, Ramya Ambikapathi, Destan Aytekin, Simon Barquera, Jane Battersby, Ty Beal, Paulina Bizzoto Molina, Carlo Cafiero, Christine Campeau, Patrick Caron, Piero Conforti, Kerstin Damerau, Michael Di Girolamo, Fabrice DeClerck, Deviana Dewi, Ismahane Elouafi, Carola Fabi, Pat Foley, Tyler J. Frazier, Jessica Gephart, Christopher Golden, Carlos Gonzalez Fischer, Sheryl Hendriks, Maddalena Honorati, Jikun Huang, Gina Kennedy, Amos Laar, Rattan Lal, Preetmoninder Lidder, Brent Loken, Quinn Marshall, Yuta J. Masuda, Rebecca McLaren, Lais Miachon, Hernán Muñoz, Stella Nordhagen, Naina Qayyum, Michaela Saisana, Diana Suhardiman, U. Rashid Sumaila, Maximo Torero Cullen, Francesco N. Tubiello, Jose-Luis Vivero-Pol, Patrick Webb, and Keith Wiebe. The state of food systems worldwide in the countdown to 2030. *Nature Food*, 4(12):1090–1110, 2023. ISSN 2662-1355. doi: 10.1038/s43016-023-00885-9.

[61] M. Schneider, T. Schelte, F. Schmitz, et al. Eurocrops: The largest harmonized open crop dataset across the european union. *Scientific Data*, 10:612, 2023. doi: 10.1038/s41597-023-02517-0.

[62] L. Seguini, A. Klish, M. Meroni, et al. Global near real-time filtered 500 m 10-day fraction of photosynthetically active radiation absorbed by vegetation (FPAR) from MODIS and VIIRS instruments suited for operational agriculture monitoring and crop yield forecasting systems. `https://agricultural-production-hotspots.ec.europa.eu/data/indicators_fpar/`, 2024. Under preparation.

[63] Lauren Stuart, Mike Hobbins, Emily Niebuhr, Alex C Ruane, Roger Pulwarty, Andrew Hoell, Wassila Thiaw, Cynthia Rosenzweig, Francisco Muñoz-Arriola, Molly Jahn, et al. Enhancing Global Food Security: Opportunities for the American Meteorological Society. *Bulletin of the American Meteorological Society*, 104(4):E760—-E777, 2024. doi: 10.1175/BAMS-D-22-0106.1.

[64] Lily-belle Sweet, Christoph Müller, Mohit Anand, and Jakob Zscheischler. Cross-Validation Strategy Impacts the Performance and Interpretation of Machine Learning Models. *Artificial Intelligence for the Earth Systems*, 2(4), 2023. ISSN 2769-7525. doi: 10.1175/AIES-D-23-0026.1.

[65] Tetsuji Tanaka, Laixiang Sun, Inbal Becker-Reshef, Xiao-Peng Song, and Estefania Puricelli. Satellite forecasting of crop harvest can trigger a cross-hemispheric production response and improve global food security. *Communications Earth & Environment*, 4(1):1–9, 2023. ISSN 2662-4435. doi: 10.1038/s43247-023-00992-2.

[66] Rachel L. Thomas and David Uminsky. Reliance on metrics is a fundamental challenge for AI. *Patterns*, 3(5):100476, 2022. ISSN 2666-3899. doi: 10.1016/j.patter.2022.100476.

[67] Sotirios A. Tsaftaris and Hanno Scharr. Sharing the Right Data Right: A Symbiosis with Machine Learning. *Trends in Plant Science*, 24(2):99–102, 2019. ISSN 1360-1385. doi: 10.1016/j.tplants.2018.10.016.

[68] Gabriel Tseng, Hannah Kerner, Catherine Nakalembe, and Inbal Becker-Reshef. Learning to predict crop type from heterogeneous sparse labels using meta-learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1111–1120, 2021. doi: 10.1109/CVPRW53098.2021.00122.

[69] Gabriel Tseng, Ivan Zvonkov, Catherine Nakalembe, and Hannah Kerner. CropHarvest: A global dataset for crop-type classification. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran, 2021. URL `https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/54229abfcfa5649e7003b83dd4755294-Paper-round2.pdf`.

[70] Gabriel Tseng, Hannah Kerner, and David Rolnick. TIML: Task-Informed Meta-Learning for Agriculture, 2022.

[71] Asaf Tzachor, Medha Devare, Brian King, Shahar Avin, and Seán Ó hÉigeartaigh. Responsible artificial intelligence in agriculture requires systemic understanding of risks and externalities. *Nature Machine Intelligence*, 4(2):104–109, 2022. ISSN 2522-5839. doi: 10.1038/s42256-022-00440-4.

[72] M. van der Velde and L. Nisini. Performance of the MARS-crop yield forecasting system for the European Union: Assessing accuracy, in-season, and year-to-year improvements from 1993 to 2015. *Agricultural Systems*, 168:203–212, 2019. ISSN 0308-521X. doi: 10.1016/j.agsy.2018.06.009.

[73] Thomas Van Klompenburg, Ayalew Kassahun, and Cagatay Catal. Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, 177:105709, 2020. doi: 10.1016/j.compag.2020.105709.

[74] Kristof Van Tricht, Jeroen Degerickx, Sven Gilliams, Daniele Zanaga, Marjorie Battude, Alex Grosu, Joost Brombacher, Myroslava Lesiv, Juan Carlos Laso Bayas, Santosh Karanam, et al. WorldCereal: a dynamic open-source system for global-scale, seasonal, and reproducible crop and irrigation mapping. *Earth System Science Data*, 15(12):5491–5515, 2023. doi: 10.5194/essd-15-5491-2023.

[75] E. Vermote. MOD09CMG MODIS/Terra Surface Reflectance Daily L3 Global 0.05Deg CMG V006., 2015.

[76] Sem Vijverberg, Raed Hamed, and Dim Coumou. Skillful U.S. Soy Yield Forecasts at Presowing Lead Times. *Artificial Intelligence for the Earth Systems*, 2(3), 2023. ISSN 2769-7525. doi: 10.1175/AIES-D-21-0009.1.

[77] Peter A. G. Watson. Machine learning applications for weather and climate need greater focus on extremes. *Environmental Research Letters*, 17(11):111004, 2022. ISSN 1748-9326. doi: 10.1088/1748-9326/ac9d4e.

[78] Christopher Yeh, Chenlin Meng, Sherrie Wang, Anne Driscoll, Erik Rozi, Patrick Liu, Jihyeon Lee, Marshall Burke, David Lobell, and Stefano Ermon. 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)*, 12 2021. URL `https://openreview.net/forum?id=5HR3vCylqD`.

[79] Jiaxuan You, Xiaocheng Li, Melvin Low, David Lobell, and Stefano Ermon. Deep gaussian process for crop yield prediction based on remote sensing data. In *AAAI*, pages 4559–4566, 2017.

[80] Rotem Zelingher and David Makowski. Investigating and forecasting the impact of crop production shocks on global commodity prices. *Environmental Research Letters*, 19(1):014026, 2023. ISSN 1748-9326. doi: 10.1088/1748-9326/ad0dda.

[81] Tianyi Zhang, Karin van der Wiel, Taoyuan Wei, James Screen, Xu Yue, Bangyou Zheng, Frank Selten, Richard Bintanja, Weston Anderson, Russell Blackport, Solveig Glomsrød, Yu Liu, Xuefeng Cui, and Xiaoguang Yang. Increased wheat price spikes and larger economic inequality with 2°C global warming. *One Earth*, 5(8):907–916, 2022. ISSN 2590-3322. doi: 10.1016/j.oneear.2022.07.004.

[82] A. Öhl, S. Ofori-Ampofo, I. Obadic, M.-Á. Fernández-Torres, R. Salih Kuzu, and X. Zhu. Uscc: A benchmark dataset for crop yield prediction under climate extremes. In *EGU General Assembly 2023*, 2023. doi: 10.5194/egusphere-egu23-15540. URL `https://doi.org/10.5194/egusphere-egu23-15540`. EGU23-15540.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] We introduce CY-Bench, a comprehensive dataset and benchmark to forecast crop yields at sub-national level. CY-Bench standardizes selection, processing and spatio-temporal harmonization of public sub-national yield statistics with relevant predictors. Our goal is to engage the machine learning community in advancing the development of sophisticated machine learning models for crop yield forecasting.
   (b) Did you describe the limitations of your work? [Yes] Limitations are discussed in section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] A description of the code and data is given in section 3.1. The reader is referred to our Github, which also contains scripts to reproduce our results.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We developed a benchmark to compare different algorithms under consistent evaluation conditions. We chose leave-one-year-out cross-validation, as justified in Section 4. We also provided a selection of models and algorithms as baselines. For model specific details, such as hyperparameter settings, the reader is referred to our Github. Hyperparameter were not optimized; some default values were used.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] While the primary focus of the benchmark is on the dataset, it does provide baseline models. The reader is referred to our Github for details on the total amount of compute resources used and the specific resource type
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We included citations to the main data sources we harvested in our work. Additionally, for a comprehensive list of all data sources, including specific citation information, the supplementary information refers the reader to a dedicated document on our Github
   (b) Did you mention the license of the assets? [Yes] We stated the use of EUPL license (version 1.2) in section 3.1. Data sources used in the benchmark may have their own license requirements. They are linked from the README in Github.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] This paper introduces a benchmark on crop yields at sub-national level. The dataset we created is accessible from Google Drive.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] The data we used is open and freely available and does not contain information about people.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] The data that we use is open, freely available, and free from personally identifiable information or offensive content. We do not curate or modify the data in a way that would introduce such concerns.

5. If you used crowdsourcing or conducted research with human subjects... (Not applicable)
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]