

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

Dilli Paudel<sup>1,\*</sup>, Michiel Kallenberg<sup>1,\*</sup>, Stella Ofori-Ampofo<sup>2</sup>, Hilmy Baja<sup>1</sup>, Ron van Bree<sup>1</sup>, Aike Potze<sup>1</sup>, Pratihtha Poudel<sup>3</sup>, Abdelrahman Saleh<sup>4</sup>, Weston Anderson<sup>5</sup>, Malte von Bloh<sup>2</sup>, Andres Castellano<sup>6</sup>, Oumnia Ennaji<sup>7</sup>, Raed Hamed<sup>8</sup>, Rahel Laudien<sup>9</sup>, Donghoon Lee<sup>10</sup>, Inti Luna<sup>11</sup>, Dainius Masiliūnas<sup>12</sup>, Michele Meroni<sup>13, 21</sup>, Janet Mumo Mutuku<sup>14</sup>, Siyabusa Mkuhlani<sup>15</sup>, Jonathan Richetti<sup>16</sup>, Alex C. Ruane<sup>6</sup>, Ritvik Sahajpal<sup>5</sup>, Guanyuan Shai<sup>5</sup>, Vasileios Sitokonstantinou<sup>11</sup>, Rogério de Souza Nória Júnior<sup>17</sup>, Amit Kumar Srivastava<sup>18</sup>, Robert Strong<sup>19</sup>, Lily-belle Sweet<sup>20</sup>, Petar Vojnović<sup>21</sup>, Allard de Wit<sup>22</sup>, Maximilian Zachow<sup>23</sup>, and Ioannis N. Athanasiadis<sup>1</sup>

<sup>1</sup>Wageningen University and Research, Artificial Intelligence, PO Box 16, Wageningen, 6700 AA, the Netherlands.

<sup>2</sup>Technical University of Munich, Chair of Data Science in Earth Observation, Arcisstraße 21, Munich, 80333, Germany

<sup>3</sup>Purdue University, Department of Agronomy, 915 Mitch Daniels Blvd, West Lafayette, IN 47907, United States

<sup>4</sup>Ankara University, Faculty Of Agriculture Engineering, Döğol Caddesi 06100 Tandoğan, Ankara, 6110, Turkey

<sup>5</sup>University of Maryland, Department of Geographical Sciences, 7251 Preinkert Drive, Collega Park, MD 20742, United States

<sup>6</sup>NASA Goddard Institute for Space Studies, GISS Climate Impacts Group, Mail Code 611, New York, NY 10025, United States

<sup>7</sup>Mohammed VI Polytechnic University, College of Computing, Lot 660, Benguerir, 43150, Morocco

<sup>8</sup>Vrije Universiteit Amsterdam, Institute for Environmental Studies, De Boelelaan 1105, Amsterdam, 1081 HV, the Netherlands

<sup>9</sup>Potsdam Institute for Climate Impact Research, Department of Climate Resilience, PO Box 60 12 03, Potsdam, 4412, Germany

<sup>10</sup>University of Manitoba, Department of Civil Engineering, 15 Gillson Street, Winnipeg, MB R3T 5V6, Canada

<sup>11</sup>Universitat de València, Image Processing Laboratory, C/ Catedràtic Agustín Escardino Benlloch, 9, València, 46980, Spain

<sup>12</sup>Wageningen University and Research, Laboratory of Geo-Information Science and Remote Sensing, PO Box 47, Wageningen, 6700 AA, the Netherlands.

<sup>13</sup>Seidor Consulting, C/Provençals 44, Barcelona, 08019, Spain

<sup>14</sup>International Crops Research Institute for the Semi-Arid Tropics, West and Central Africa Region Hub, PO Box 320, Bamako, Mali

<sup>15</sup>International Institute of Tropical Agriculture, Natural Resources Management, PO Box 30677, Nairobi, 00100, Kenya

<sup>16</sup>Commonwealth Scientific and Industrial Research Organisation (CSIRO), Agriculture and Food, 147 Underwood Av, Perth, WA 6014, Australia

<sup>17</sup>National Research Institute for Agriculture, Food and Environment (INRAE), UMR LEPSE, 2 Pl. Pierre Viala, Montpellier, 34000, France

<sup>18</sup>Leibniz Centre for Agricultural Landscape Research, Simulation and Data Science, Eberswalder Straße 84, Müncheberg, 15374, Germany

<sup>19</sup>Texas A&M University, Agriculture Education, 600 John Kimbrough Blvd, College Station, TX 77843-2116, United States

<sup>20</sup>Helmholtz Centre for Environmental Research, Department of Computational Hydrosystems, Permoserstraße 15, Leipzig, 04318, Germany

<sup>21</sup>European Commission's Joint Research Centre, Food Security Unit, Via E. Fermi 2749, Ispra, VA I-21027, Italy

<sup>22</sup>Wageningen University and Research, Earth Observation and Environmental Informatics, PO Box 47, Wageningen, 6700 AA, the Netherlands.

<sup>23</sup>Technical University of Munich, Chair of Digital Agriculture, Liesel-Beckmann-Straße 2, Freising, 85354, Germany

\*These authors contributed equally to this work.

**Correspondence:** Michiel Kallenberg (michiel.kallenberg@wur.nl)

**Abstract.** In-season, pre-harvest crop yield forecasts are essential for enhancing transparency in commodity markets and improving food security. They play a key role in increasing resilience to climate change and extreme events and thus contribute to the United Nations' Sustainable Development Goal 2 of zero hunger. Pre-harvest crop yield forecasting is a complex task, as several interacting factors contribute to yield formation, including in-season weather variability, extreme events, long-term climate change, soil, pests, diseases and farm management decisions. Several modeling approaches have been employed to capture complex interactions among such predictors and crop yields. Prior research for in-season, pre-harvest crop yield forecasting has primarily been case-study based, which makes it difficult to compare modeling approaches and measure progress systematically. To address this gap, we introduce CY-Bench (Crop Yield Benchmark), a comprehensive dataset and benchmark to forecast maize and wheat yields at a global scale. CY-Bench was conceptualized and developed within the Machine Learning team of the Agricultural Model Intercomparison and Improvement Project (AgML) in collaboration with agronomists, climate scientists, and machine learning researchers. It features publicly available sub-national yield statistics and relevant predictors, such as weather data, soil characteristics, and remote sensing indicators, that have been pre-processed, standardized, and harmonized across spatio-temporal scales. With CY-Bench, we aim to: (i) establish a standardized framework for developing and evaluating data-driven models across diverse farming systems in more than 25 countries across six continents; (ii) enable robust and reproducible model comparisons that address real-world operational challenges; (iii) provide an openly accessible dataset to the earth system science and machine learning communities, facilitating research on time series forecasting, domain adaptation, and online learning. The dataset (<https://doi.org/10.5281/zenodo.11502142>, (Paudel et al., 2025a)) and accompanying code (<https://github.com/WUR-AI/AgML-CY-Bench>, (Paudel et al., 2025b))) are openly available to support the continuous development of advanced data driven models for crop yield forecasting to enhance decision-making on food security.

*Copyright statement.* The CY-Bench dataset and the python library are licensed under EUPL-1.2, which is compatible with all of the licenses for the datasets included.

## 1 Introduction

The global food system faces significant challenges, including unequal access to resources and volatile markets, despite advancements in agricultural production (Ambikapathi et al., 2022; Zhang et al., 2022; Chen and Villoria, 2022; Zelingher and Makowski, 2023; Schneider et al., 2023a) To enhance food security policies, experts have emphasized the need for improved data, maps, and predictions (Mehrabi et al., 2022; Ennaji et al., 2023; Fanzo, 2024). Pre-harvest yield forecasts, in particular, play a critical role in enhancing global market transparency and enabling decision-makers to plan and respond effectively to potential food shortages, especially in the face of a changing climate (Becker-Reshef et al., 2020; Tanaka et al., 2023; Stuart et al., 2024).

Crop yield forecasts are produced by both private entities and government institutes using field surveys, process-based crop models (Gavasso-Rita et al., 2023), and data-driven methods. The latter include statistical regression (e.g., linear models with predefined assumptions) and machine learning (e.g., random forests, neural networks) (Basso and Liu, 2019; Schauburger et al., 2020; Paudel et al., 2021). Commonly used predictors are weather, soil moisture, crop productivity, and remotely-sensed  
35 vegetation health indicators. Data availability determines the yield forecasting modeling setup and the selected spatial scale, which can range from national to sub-national and field levels. 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 the expertise of analysts (van der Velde and Nisini, 2019). In India, the Ministry of Agriculture generates multi-season crop production forecasts for major crops from remote sensing and agro-meteorological data using  
40 statistical and crop models (Singh et al., 2022). In the United States, the USDA’s National Agricultural Statistics Service (NASS) publishes monthly crop forecasts on a multi-stage administrative scale, primarily based on in-field surveys (Good and Irwin, 2003). Data for sub-national crop yield forecasting, which focuses on higher resolution administrative units (e.g., regions, provinces) and captures spatial yield variability within a country (Meroni et al., 2021; Paudel et al., 2022), is crucial for targeted food security planning. Such data is usually publicly available, but compiling them for many countries is challenging due to  
45 differences in collection and reporting protocols, including language and data format.

Traditionally, crop yield prediction has been based on biophysical process based crop models, grounded in decades of agricultural knowledge. Their challenges in parametrization (He et al., 2017; Wallach et al., 2021; Seidel et al., 2018) however limits scalability and their reliance on first principles hinders significant improvements in forecasting accuracy (van der Velde and Nisini, 2019). Machine learning methods offer promising alternatives, capturing processes not fully covered by biophysical  
50 models. However, they typically require high-quality datasets covering large areas and multiple years. Several review articles (Chlingaryan et al., 2018; Kamilaris and Prenafeta-Boldú, 2018; Liakos et al., 2018; Van Klompenburg et al., 2020; Benos et al., 2021; Oikonomidis et al., 2022) have highlighted promising performance of machine learning methods, including deep learning, for pre-harvest yield forecasting (Schlenker and Roberts, 2009; You et al., 2017; Khaki et al., 2020; Mateo-Sanchis et al., 2021; Paudel et al., 2022; Fan et al., 2022; Liu et al., 2022; Lesk et al., 2022; Paudel et al., 2023b; Vijverberg et al., 2023; Ma et al.,  
55 2023; Priyatikanto et al., 2023; Ennaji et al., 2024). However, the data and code used in such studies are not always available, and the diversity in evaluation procedures, metrics, and datasets makes intercomparison and synthesis of results difficult. As a result, the research community is unable to reproduce results and compare the strengths and weaknesses of different methods across crops and regions.

To better understand the specific strengths and weaknesses of data-driven methods for pre-harvest yield forecasting, and  
60 to drive future research progress, well-documented benchmark datasets compiled by domain experts are vital (Tsaftaris and Scharr, 2019; Dueben et al., 2022; Rolnick et al., 2024; Sweet et al., Under review). While benchmark datasets can focus on specific regions or tasks, geographic diversity is a particularly valuable property for agricultural machine learning. Crop yield is influenced by highly variable and region-specific factors such as climate, farming practices, soil properties, and socioeconomic conditions. Datasets that span multiple countries and agro-ecological contexts help ensure that models are  
65 robust, transferable, and better equipped to generalize beyond well-studied regions (Richards et al., 2023; Tzachor et al., 2022).

Including underrepresented or data-scarce regions also promotes research equity and allows the community to address the full spectrum of challenges in global food production (Nakalembe and Kerner, 2023). In addition to producing accurate forecasts, models must be reliable in real-world settings for adoption by stakeholders (van der Velde and Nisini, 2019). The evaluation metrics should closely represent the needs of stakeholders and allow a more granular breakdown of model performance (Thomas and Uminsky, 2022; Burnell et al., 2023) - for example, the model's ability to capture yield variability in years with climate extremes (Watson, 2022). To avoid overestimation of model skill, the evaluation procedure must take into account the specific challenges arising from the use of spatio-temporal data that does not satisfy independent and identically distributed assumptions (Meyer and Pebesma, 2022; Sweet et al., 2023; Kapoor and Narayanan, 2023; Richetti et al., 2023).

A few research works have compiled benchmark-like datasets that include components related to crop yield prediction. SustainBench (Yeh et al., 2021) includes a benchmark dataset for crop yield prediction, and targets end-of-season prediction for only one crop (soybean) in three countries (United States, Brazil and Argentina). Another public dataset is CropNet (Lin et al.), which only covers the United States. Similarly, there are ongoing efforts to produce a multi-task benchmark dataset which includes yield prediction in the USA as a sub-task (Höhl et al., 2023). Apart from these, other available data contributions include yield statistics only (Lee et al., 2025; Potter, 2019; Ronchetti et al., 2024; Duden et al., 2024; Argentina; Australia; Brazil; China; India; Mexico) or sample data published with articles (Khaki et al., 2020; Fernandez-Beltran et al., 2021; Paudel et al., 2021, 2023b) without releasing the full datasets.

We present CY-Bench, a comprehensive dataset and benchmark for sub-national crop yield forecasting, covering thirty-eight countries for maize and twenty-nine countries for wheat across six continents. 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). Thus, the dataset combines sub-national yield statistics with relevant predictors, such as growing-season weather, remote sensing indicators, and soil properties. The novelty of our contribution lies in the integration, harmonization, and standardization of multiple disparate sources into a single, analysis-ready benchmark. To our knowledge, this is the first publicly available dataset that enables consistent model comparison for sub-national yield forecasting across such a broad set of regions and production systems. Key novel aspects include: (i) facilitating benchmarking across countries rather than single-country datasets, filling a gap recently emphasized in the literature (Davis et al., 2025; Sweet et al., 2025); (ii) harmonization of yield statistics, weather, soil, and remote sensing predictors with consistent spatial and temporal alignment; (iii) broad geographical coverage, explicitly including low- and middle-income countries that are often underrepresented in benchmarks; and (iv) development in collaboration with agricultural experts, climate scientists, and machine learning researchers from the AgML community (<https://www.agml.org/>) who guided the selection and alignment of key variables and data sources and contributed to designing the dataset with benchmarking needs in mind. CY-Bench thus lowers the barrier to entry for machine learning researchers in this crucial application area, facilitates model intercomparison across diverse agricultural systems, and promotes reproducibility and transparency in agricultural forecasting. By providing an integrated, expert-curated benchmark, our work complements existing datasets that are either crop- or country-specific and enables the community to develop and evaluate forecasting tools that can support decision-makers in food security planning worldwide.

## 2 Dataset construction

CY-Bench is a benchmark dataset to train and evaluate crop yield forecasting models that produce in-season forecasts with a certain lead time ahead of harvest. The benchmark includes a comprehensive set of predictors that are known to be important drivers of crop yield. Crop yield is determined by the complex interaction of genetics (G), environmental conditions (E), and management decisions (M), commonly referred to as  $G \times E \times M$ . Genetics (G) includes factors such as genotype, phenotype, and cultivar; environmental conditions (E) encompass both abiotic factors (e.g., climate, soil) and biotic factors (e.g., pests, pollinators); and management decisions (M) involve farm practices like tillage, sowing dates, irrigation, and fertilization (Liliane and Charles, 2020). Technological advancements, including genetic improvements, better farm inputs, machinery, and enhanced management practices, have all contributed to increased yields over time (Liliane and Charles, 2020). These advancements are often reflected in the yield trend (Lecerf et al., 2019). At the sub-national spatial scale, data on genetic differences (e.g., cultivars) and farm management practices, with the exception of planting and harvest dates, are often not available. Therefore, previous studies of crop yield forecasting also commonly rely on environmental factors, crop calendar information (planting and harvest dates), and yield trend (You et al., 2017; Khaki et al., 2020; Paudel et al., 2021). Factors not explicitly captured in CY-Bench but known to influence end-of-season yields are biotic stressors (e.g., pests and diseases), farm management choices (e.g., irrigation, fertilization, cultivar selection), and socioeconomic factors (e.g., market prices, labor availability, and policy changes).

### 2.1 Data sources selection

In the absence of benchmark datasets like CY-Bench, modeling crop yield involves numerous uncertainties regarding data sources and quality. The data collection and pre-processing protocols of many predictor datasets, cropland or crop type maps, crop calendars and official statistics are important to select suitable data sources. Each data source has strengths and limitations related to spatial and temporal resolutions and coverage, sampling methods and gap-filling strategies. After data source selection, predictor data from diverse spatial resolutions needs to be aggregated (to the level of yield statistics), which can lead to information loss. In constructing CY-Bench, we engaged a diverse community of researchers to weigh the benefits and limitations of data sources for each type of data necessary to produce crop yield forecasts. The outcome is a curated set of sources, including alternative and our justifications for picking one data source over another. The selection process was guided by several principles: (i) global coverage, (ii) public accessibility, (iii) regular updates (except for inherently static datasets), (iv) near real-time availability, and (v) demonstrated relevance for crop growth and development. A summary of these decisions is provided in the dedicated data sections of this paper, while a more comprehensive overview, including specific benefits and limitations of each dataset considered, is available in our online documentation (Paudel et al., 2025b). Each selected dataset is further described in accompanying data cards, which provide links to sources, reports, and related publications.

#### 2.1.1 Weather and soil moisture data

The most relevant weather variables for crop yield forecasting are temperature, solar radiation, and precipitation (Frieler et al., 2017). Derived weather variables such as vapour pressure deficit and evapotranspiration provide additional information on

crop water demand and stress, which can improve yield prediction (López et al., 2021; Tadesse et al., 2015). Precipitation affects soil moisture availability, and evapotranspiration which is strongly influenced by vapour pressure deficit, governs water loss and stress risk. Although actual evapotranspiration would be preferred over reference evapotranspiration, the former is crop-dependent and not readily available at a global scale.

Temperature (`temp`), precipitation (`prec`), radiation (`rad`), reference evapotranspiration (`ETo`), and vapor pressure deficit (`VPD`) were selected from AgERA5 (Boogaard et al., 2022), which provides daily data at a  $0.1^\circ$  ( $\sim 11$  km) spatial resolution. AgERA5 offers agrometeorological indicators from 1979 to the present, derived from ERA5 reanalysis and is tailored for agricultural studies. Its key benefits include high-quality data with near real-time updates (i.e. lag of  $\sim 2$  weeks), comprehensive documentation, and free access via the Copernicus Climate Data Store (CDS). Other datasets, such as PRISM, GridMET, TerraClimate, MSWEP, and CPC, have limitations including restricted geographic coverage, coarser temporal or spatial resolution, and fewer variables. A detailed comparison of candidate datasets and trade-offs is available on our GitHub Repository (Paudel et al., 2025b)

For soil moisture, we selected surface soil moisture (`SSM`) and root-zone soil moisture (`RSM`) from the Global Land Data Assimilation System (GLDAS) dataset (Rodell et al., 2004). This dataset represents gridded and global soil moisture data developed by integrating satellite- and ground-based observational data products, using advanced land surface modeling and data assimilation techniques. The dataset is available from 2003 to present, and can be freely downloaded from Goddard Earth Sciences Data and Information Services Center (GES DISC). It has a temporal resolution of one day, and a spatial resolution of  $0.25^\circ$  ( $\sim 28$  km). An alternative dataset is the Global Land Evaporation Amsterdam Model (GLEAM) (Miralles et al., 2024), but it is typically updated only once a year and is currently available only until December 2023.

### 2.1.2 Remote sensing data

Remote sensing indicators of crop biomass and health include vegetation indices, such as the normalized difference vegetation index (`NDVI`) and enhanced vegetation index (`EVI`), as well as biophysical metrics like the fraction of absorbed photosynthetically active radiation (`fPAR`) and leaf area index (`LAI`). Sub-national yield forecasting requires long-term time series of these indicators, coupled with frequent satellite revisits to ensure cloud-free imagery. These requirements practically limit options to the coarse-resolution missions MODIS and its successor, VIIRS. While MODIS data can be directly downloaded from NASA Land Processes Distributed Active Archive Center (LPDACC), the raw vegetation indices and biophysical variables are often of low quality due to issues like cloud cover. These limitations require further screening, gap-filling, and corrections. Furthermore, additional processing, which includes the use of quality flags and the application of temporal smoothing procedures, is time-consuming and complex. These challenges become even more pronounced when processing near-real-time data, which is essential for operational yield forecasting.

In view of an operational deployment of sub-national yield forecasting, we selected two analysis-ready operational products representing crop biomass and health: `fPAR` and `NDVI`. `fPAR` is provided as dekadal (10 day) data with a spatial resolution of  $0.0045^\circ$  (500 m), utilizing gap-filled and smoothed MODIS and VIIRS datasets. This data is sourced from EC-JRC, and its quality is being evaluated in Seguini et al. (In preparation). `NDVI`, a key indicator of vegetation greenness, is derived from

MOD09CMG (Vermote, 2015), available from NASA LPDACC. The data is prepared as an eight-day composite, selecting the pixel with the highest quality for each composite period. This interval offers a practical compromise: shorter composites tend to be noisier due to clouds, while longer intervals risk missing short-term vegetation dynamics and phenological changes. The data is provided at a spatial resolution of  $0.05^\circ$  ( $\sim 5$  km). The quality of this NDVI product has been evaluated in Meroni et al. (2019).

While NDVI can saturate under dense canopies, alternatives such as EVI and GCVI overcome this limitation. We chose to focus on fPAR and NDVI because of their extensive use in crop yield studies, their direct link to canopy structure and light interception, and their availability as analysis-ready products. Moreover, EVI relies on empirical constants for canopy background adjustment, which can introduce calibration challenges across regions (Huete et al., 2002).

### 175 **2.1.3 Soil data**

We selected data from the World Inventory of Soil Emission Potentials (WISE) project (Batjes, 2016) for static soil properties. WISE data is constructed using the soil map unit delineations of the broad-scale Harmonized World Soil Database, overlaid by a climate zones map (Köppen-Geiger) as co-variate, and soil property estimates derived from analyses of the ISRIC-WISE soil profile database for the respective mapped 'soil/climate' combinations. The dataset has a spatial resolution of 30 arc-second (0.00833° ( $\sim 0.9$  km)).

While SoilGrids (Hengl et al., 2017; Poggio et al., 2021) is an alternative, WISE was selected due to its suitability for agricultural applications. Specifically, WISE data is considered to be more readily interpretable and provides essential parameters like soil rooting depth and water holding capacity, which are absent in SoilGrids.

### **2.1.4 Crop mask and crop calendar data**

185 Crop masks are selected from the European Space Agency WorldCereal (ESA WorldCereal) project (Van Tricht et al. (2023)), which provides an up-to-date and actively maintained source for cropland and crop type maps at a spatial resolution of  $0.0045^\circ$  (500 m). Alternative sources of crop masks include Anomaly Hotspots of Agricultural Production (ASAP) from EC-JRC and IIASA (JRC-IIASA) and Global Best Available Crop Specific Masks (GEOGLAM-BACS) from the Group on Earth Observations Global Agriculture Monitoring (GEOGLAM). We considered ESA WorldCereal to be a better choice than the generic cropland layer from JRC-IIASA because of the availability of crop type maps for maize and wheat (spring and winter cereals). Although GEOGLAM-BACS (Becker-Reshef et al., 2023) provides crop type maps for maize and wheat (spring and winter cereals), their spatial resolution ( $0.05^\circ$ ) is lower compared to ESA WorldCereal ( $0.00464^\circ$ ).

Crop calendars also come from the ESA WorldCereal project (Franch et al., 2022). ESA WorldCereal crop calendars combine information from existing global crop calendar products, such as GEOGLAM Crop Monitor, the United States Department of Agriculture Foreign Agricultural Service (USDA-FAS), FAO, and EC-JRC's ASAP, into a baseline map and sample them to train a Random Forest algorithm based on climatic and geographic data. They have global coverage and a spatial resolution of  $0.5^\circ$  ( $\sim 50$  km). We considered alternative sources, including Food and Agriculture Organization (FAO), GGCM (Waha et al., 2012; Minoli et al., 2019), MIRCA (Portmann et al., 2010), and SAGE (Sacks et al., 2010). However, we selected ESA

WorldCereal primarily due to its global coverage and alignment with our crop statistics data. A detailed comparison, based on  
200 crop types, country coverage, spatial resolution, and data sources, can be found in our GitHub Repository (Paudel et al., 2025b).

### 2.1.5 Crop statistics data

Crop yield statistics for sub-national administrative levels are obtained from national statistics offices or regional agencies, depending on their quality and timely availability. In most cases, they come from the national statistics offices. For example, in the United States, they are published by the National Agricultural Statistics Service (NASS) of the United States Department  
205 of Agriculture (USDA). For the European Union, member countries report statistics to Eurostat. However, we considered Ronchetti et al. (2024) a more reliable source than Eurostat, as they follow a harmonization procedure developed by EC-JRC, standardizing crop definitions, administrative boundaries, and reporting practices to produce comparable annual yield time series. For Germany, we selected data from Duden et al. (2024) instead of Ronchetti et al. (2024) because of better temporal coverage (1979-2021 vs 1999-2020), higher spatial resolution for maize (NUTS level 3 vs level 1) and better quality based on  
210 consistency checks (e.g.,  $yield = production/area$ ). For Africa, except for Mali, data comes from the USAID's Famine Early Warning Systems Network (FEWS NET) Data Warehouse via the HarvestStat Africa dataset. The data was compiled by FEWS NET and NASA Harvest and harmonized by Lee et al. (2025) to account for changing administrative boundaries and reporting inconsistencies over time. For Mali, we selected Compagnie Malienne pour le Developpement des Textiles (CMDT) dataset (ICRISAT Mali, 2018) that provides higher spatial resolution data at arrondissement-level (administrative level 3). Depending  
215 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) (Table 2). When statistics for multiple administrative levels are available, we select the highest resolution.

## 2.2 Data preparation

### 2.2.1 Crop yield data (targets)

220 CY-Bench dataset includes crop statistics from thirty-eight countries for maize and twenty-nine countries for wheat (Figure 3). Coverage maps show that CY-Bench has extensive coverage when layered on top of crop type maps from ESA WorldCereal, with notable omissions including Canada, Ukraine and Russia for wheat and Ukraine, Uganda and Tanzania for maize. Data preparation for yield data involved filtering out values that do not meet certain consistency checks, e.g.,  $yield \neq production/area$ , or zero values. The data sources or publications from which CY-Bench draws the data do additional consistency checks. We  
225 refer interested readers to respective data cards (Pushkarna et al., 2022) in our GitHub repository (Paudel et al., 2025b) which contains further links to data sources, related reports and publications.

### 2.2.2 Predictor data

CY-Bench predictor data includes static soil properties and time series of weather variables, soil moisture indicators and vegetation indicators (Table 1). Predictor data and yield statistics often differ in spatial and temporal resolution, requiring further

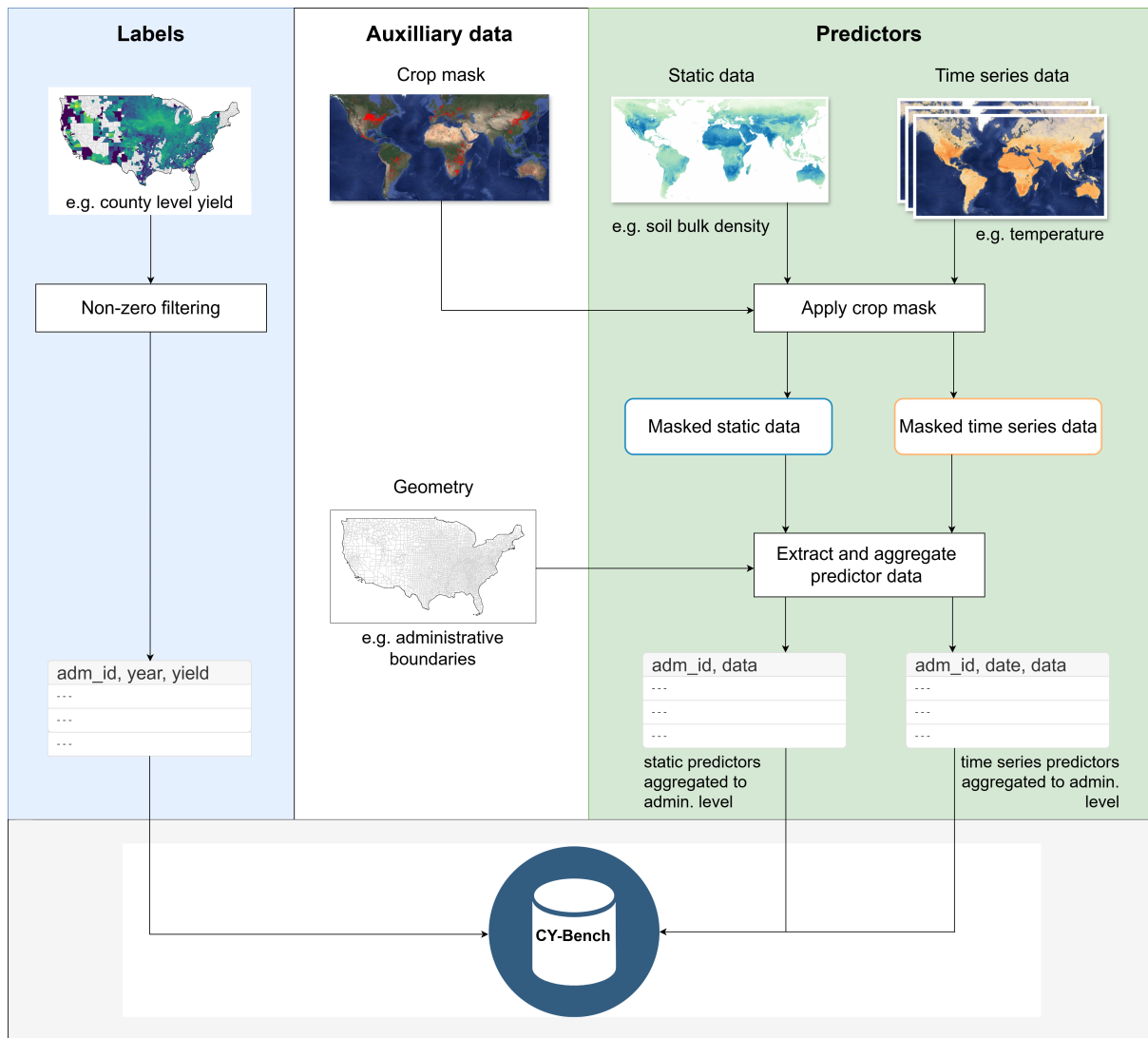
230 processing to align them effectively. While such aggregation can mask fine-scale variability or temporal dynamics, it can also  
smooth out noise and improve the stability of predictors (Hoffmann et al., 2016; Paudel et al., 2023). Weather,  $ET_0$  and soil  
moisture data come in daily time steps.  $fPAR$  comes in dekadal time step, with three values per month (days 1-10, 11-20, 21-31).  
NDVI data is available every eight days, with gaps due to cloud cover.

Predictor data is filtered using crop type maps (or crop masks) from EC-JRC (2024) which are derived from the ESA  
235 WorldCereal project (Van Tricht et al., 2023). This step restricts predictor data to pixels in harvested crop areas only. After  
masking, predictor data is aggregated to match the boundaries and spatial level of the yield data according to the administrative  
level (Figure 1). The data preparation workflow is implemented in a Python script in our GitHub repository (Paudel et al., 2025b).  
We note that all predictor data retain their temporal resolution from the original data source, creating a multi-modal dataset.

**Table 1.** Overview of the predictor data, crop mask and crop calendar

Abbreviations: Temperature ( $t_{temp}$ ), Precipitation flux ( $p_{prec}$ ), Solar radiation flux ( $rad$ ), Reference evapotranspiration ( $ET_0$ ), Vapor pressure  
deficit ( $VPD$ ), Climatic water balance ( $CWB$ ), Fraction of absorbed photosynthetically active radiation ( $fPAR$ ), Normalized difference  
vegetation index ( $NDVI$ ), and available water capacity ( $AWC$ ).

Category	Data		Spatial resolution	Temporal resolution	Source
	Name	Unit			
Weather (time series)	temp	$^{\circ}C$	$0.1^{\circ}(11\text{ km})$	daily	AgERA5 (Boogaard et al., 2022)
	prec	mm			
	rad	$J\ m^{-2}$			
	et0	mm			
	vpd	hPa			
	CWB	mm			
Vegetation (time series)	soil moisture	$kg\ m^{-2}$	$0.25^{\circ}(28\text{ km})$	daily	NASA GLDAS (Rodell et al., 2004)
	fPAR	%	$0.0045^{\circ}(0.5\text{ km})$	10 days	JRC (Seguini et al., In preparation)
Soil (static)	NDVI	-	$0.05^{\circ}\ (5.5\text{ km})$	8 days	MOD09CMG (Vermote, 2015)
	AWC	$cm\ m^{-1}$	$0.0083^{\circ}(0.9\text{ km})$	static	WISE (Batjes, 2016)
	bulk density	$kg\ dm^{-3}$			
drainage class	-				
Location (static)	latitude	degrees		static	
	longitude	degrees		static	
Crop (auxiliary)	crop mask	%	$0.0045^{\circ}(0.5\text{ km})$	static	WorldCereal (Van Tricht et al., 2023; EC-JRC, 2024)
	crop calendar	day	$0.5^{\circ}\ (55\text{ km})$	static	WorldCereal (Franch et al., 2022)

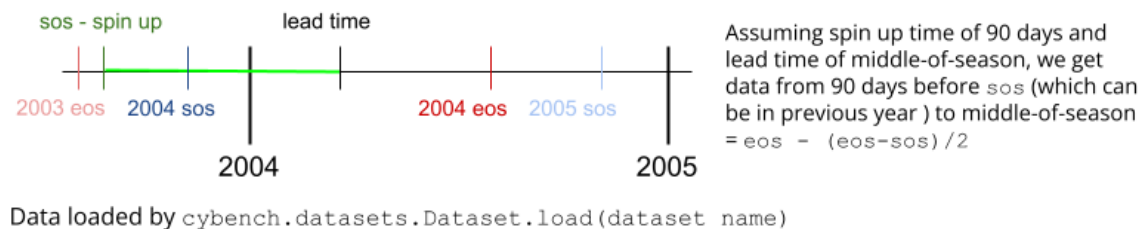


**Figure 1.** Overview of the CY-Bench data preparation process.

### 2.2.3 Additional pre-processing for yield forecasting models

240 Here we describe some additional pre-processing implemented in our `cybench` library that are relevant for building crop  
 yield forecasting models. Predictor data from different sources come with different temporal coverage. Similarly, they include  
 observations for the calendar year, which may not capture the crop season. First, we align time series inputs (weather variables,  
 remote sensing indicators and soil moisture indicators) to the crop season (see Figure 2). We define the boundaries of the crop  
 season as 90 days before the start of season (the spin-up time) to the end of season in a particular calendar year and filter out  
 245 data outside the boundaries. Therefore, data from the previous year can be included in the current calendar year's crop season

and data after the end-of-season date get pushed to the crop season for the next calendar year. Furthermore, data towards the end-of-season are filtered out based on the lead time relative to harvest or end-of-season. Second, we align the input data sources and label data to produce a set of data samples that are complete, i.e. each data sample includes all the relevant predictors for each time step (or static) and a label.



**Figure 2.** Alignment of time series predictors to the crop season.

250 The time series predictors need further pre-processing during modeling. Certain models require time series data to have the same number of time steps. Therefore, time series inputs are aggregated to dekadal time steps (days 1-10, 11-20, 21-30, and so on), taking the mean of most variables, minimum of minimum temperature, maximum of maximum temperature and the sum of precipitation flux, climatic water balance and solar radiation flux. Where the variable is categorical (such as soil drainage), we take the mode.

255 To further prepare features as tabular data, time series data are aggregated in the temporal dimension to create domain-relevant features. Following expert recommendations we create monthly averages of minimum daily temperature (`tmin`), maximum daily temperature (`tmax`), average daily temperature, daily precipitation (`prec`), cumulative climatic water balance (`prec - ETo`) and surface soil moisture. Similarly, monthly maximum values are calculated for cumulative growing degree days (`GDD`), cumulative precipitation, cumulative `fPAR` and cumulative `NDVI`. Furthermore, we calculate the number of days in which `tmin` is less than 0°C ('cold days'), days in which `tmax` is greater than 35°C ('hot days') and days where `prec` is less than 1 mm ('dry days').

#### 2.2.4 Future expansion and data integration

CY-Bench currently includes predictor data from 2003 through 2023. Availability of crop statistics varies by country (see Figures 4, 5). We share yield and predictor data preparation scripts and notebooks in our GitHub repository (Paudel et al., 2025b) to  
 265 make the inclusion of new data possible as it becomes available. For example, when crop statistics for new data years become available for specific countries, the data preparation pipeline for agricultural yield data can be run for the crop statistics, and predictor data preparation scripts can be run for predictor inputs. Expanding the database in the future primarily depends on onboarding crop statistics, as the global availability of input predictors ensures that integrating additional crop statistics is the only prerequisite for extending CY-Bench's coverage.

### 3.1 Dataset overview

CY-Bench covers two main crops, namely maize and wheat. Depending on the country, the crop names can refer to different varieties or seasons of maize and wheat as detailed in our GitHub repository (Paudel et al., 2025b).

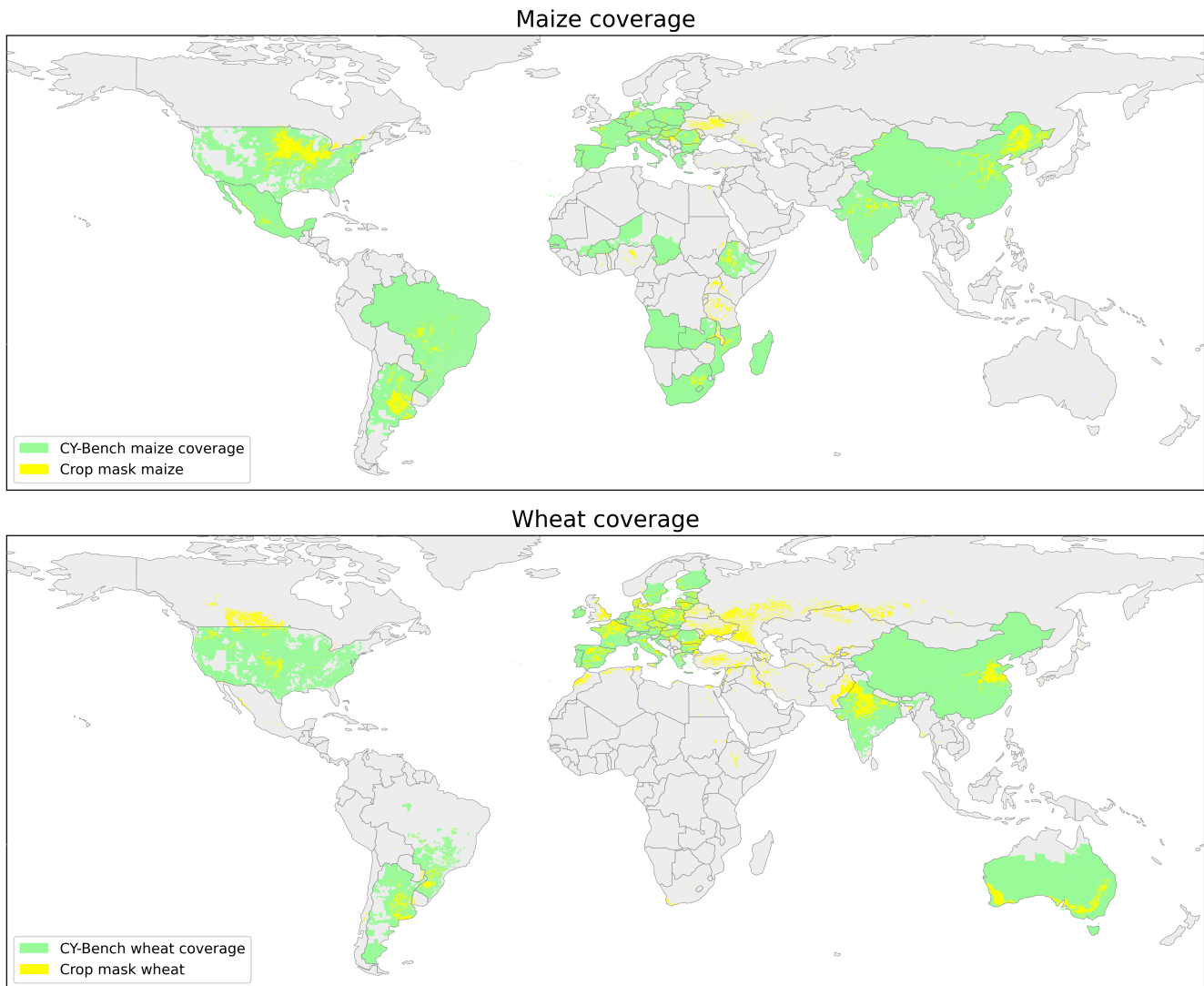
**Table 2.** Countries and the administrative levels for which CY-Bench data is provided

Group	Country name (country code) : Admin level or NUTS Level		
EU (n=23)	Austria (AT) : 2	Belgium (BE) : 2	Bulgaria (BG) : 2
	Czechia (CZ) : 3	Germany (DE) : 3	Denmark (DK) : 3
	Estonia (EE) : 3	Greece (EL) : 3	Spain (ES) : 3
	Finland (FI) : 3	France (FR) : 3	Croatia (HR) : 2
	Hungary (HU) : 3	Ireland (IE) : 2	Italy (IT) : 3
	Lithuania (LT) : 3	Latvia (LV) : 3	Netherlands (NL) : 2
	Poland (PL) : 2	Portugal (PT) : 2	Romania (RO) : 3
	Sweden (SE) : 3	Slovakia (SK) : 3	
FEWSNET (n=12)	Angola (AO) : 1	Burkina Faso (BF) : 2	Ethiopia (ET) : 2
	Lesotho (LS) : 1	Madagascar (MG) : 2	Malawi (MW) : 2
	Mozambique (MZ) : 1	Niger (NE) : 2	Senegal (SN) : 2
	Chad (TD) : 1	South Africa (ZA) : 1	Zambia (ZM) : 2
Other countries (n=8)	Argentina (AR) : department	Australia (AU) : ABARES region (sub-state)	
	Brazil (BR) : municipality	China (CN) : province	
	India (IN) : district	Mali (ML) : Municipality	
	Mexico (MX) : state	United States (US) : county	

### 3.2 Task

275 CY-Bench is designed to train and evaluate models for in-season crop yield forecasting of wheat and maize at the sub-national level, covering major and underrepresented crop-growing countries worldwide. Forecasts can be made at multiple time points from start of season ( $sos$ ) to end of season ( $eos$ ), based on a lead time relative to  $eos$ , e.g., middle-of-season ( $((eos - sos)/2)$ ), quarter-of-season ( $((eos - sos)/4)$ ) and  $n$ -days before harvest. The exact inference time depends on the crop calendar for the selected crop and region. The quarter-of-season mark often coincides with crops reaching physiological maturity, while the

280 middle-of-season typically represents the transition from vegetative to reproductive growth stages (Lee et al., 2022; Basso and Liu, 2019). The reproductive period, which includes flowering and grain filling, is generally the most climate-sensitive phase of the growing season (Hatfield and Prueger, 2015). By contrast, the vegetative phase before mid-season and the senescence phase after grain filling tend to be less sensitive to climate anomalies, although rainfall during the harvest period can significantly



**Figure 3.** Coverage of CY-Bench for maize (38 countries) and wheat (29 countries) overlaid on crop type maps from ESA WorldCereal. Coverage is extensive, with notable omissions including Ukraine, Uganda, and Tanzania for maize, and Canada, Ukraine, and Russia for wheat. Some administrative units are covered despite no visible crop mask; an apparent mismatch caused by thresholding of low-percentage crop mask pixels (e.g., Tasmania contains only a few hectares of wheat).

285 impact yield. While quarter-of-season forecasts may achieve higher accuracy, middle-of-season forecasts balance accuracy with actionable insights, allowing for management adjustments during the remainder of the growing season—something late forecasts cannot provide.

Country	ADMIN SIZE km²	Maize				Wheat			
		YEARS MIN-MAX	ADMIN REGIONS	LABELS	LABELS TIMELINE 2003-2024	YEARS MIN-MAX	ADMIN REGIONS	LABELS	LABELS TIMELINE 2003-2024
Austria	9541	2003-2020	9	159		2004-2020	9	133	
Belgium	3011	2011-2020	10	85		2004-2020	11	174	
Bulgaria	19435	2003-2020	6	106		2010-2020	6	45	
Czechia	5300	2005-2020	14	212		2004-2020	14	238	
Germany	798	2004-2021	215	3159		2004-2021	358	5888	
Denmark	3496	2011-2020	3	22		2006-2020	10	150	
Estonia	9065	-	0	0		2004-2020	5	85	
Greece	2345	2009-2019	40	440		2004-2019	42	640	
Spain	8039	2003-2020	50	853		2004-2020	45	555	
Finland	10626	-	0	0		2004-2020	18	176	
France	5976	2004-2020	92	1539		2004-2020	94	1581	
Croatia	28211	2005-2020	2	32		2008-2020	2	25	
Hungary	4394	2003-2020	20	352		2004-2020	20	278	
Ireland	25740	-	0	0		2010-2020	3	11	
Italy	2476	2003-2020	101	1603		2004-2020	86	1242	
Lithuania	6306	2003-2020	10	146		2004-2020	10	169	
Latvia	12188	-	0	0		2004-2018	5	75	
Netherlands	2960	2008-2020	12	126		2004-2020	12	195	
Poland	18177	2003-2020	17	291		2004-2020	17	284	
Portugal	4968	2003-2020	5	88		2004-2020	5	85	
Romania	5547	2003-2020	42	736		2004-2020	34	379	
Sweden	11671	2007-2020	1	10		2004-2020	17	271	
Slovakia	6538	2007-2018	8	94		2017-2018	5	10	

**Figure 4.** Data size summary per dataset for EU countries

Country	ADMIN SIZE km <sup>2</sup>	Maize				Wheat			
		YEARS MIN-MAX	ADMIN REGIONS	LABELS	LABELS TIMELINE 2003-2024	YEARS MIN-MAX	ADMIN REGIONS	LABELS	LABELS TIMELINE 2003-2024
Australia	123694	-	0	0		2003-2023	23	442	
China	166114	2003-2022	31	602		2004-2022	25	475	
India	3950	2003-2017	506	6597		2004-2017	476	6289	
Angola	59858	2004-2017	18	252		-	0	0	
Burkina Faso	5392	2003-2019	45	540		-	0	0	
Ethiopia	10350	2003-2020	65	778		-	0	0	
Lesotho	2864	2004-2021	10	163		-	0	0	
Madagascar	21902	2005-2010	22	132		-	0	0	
Mali	3183	2003-2017	24	360		-	0	0	
Malawi	3117	2018-2023	4	16		-	0	0	
Mozambique	75397	2004-2022	10	159		-	0	0	
Niger	4404	2003-2021	25	264		-	0	0	
Senegal	2853	2003-2015	40	401		-	0	0	
Chad	40330	2003-2017	17	231		-	0	0	
South Africa	123172	2004-2022	9	167		-	0	0	
Zambia	8892	2004-2017	71	994		-	0	0	
Mexico	58682	2014-2022	64	133		-	0	0	
United States	1614	2003-2023	1930	33193		2004-2023	1638	22834	
Argentina	3174	2004-2023	299	5599		2003-2023	240	4661	
Brazil	423	2003-2023	4567	88759		2003-2022	1015	18472	

Figure 5. Data size summary per dataset for other (non-EU) countries

### 3.2.1 Formal definition

The input data consists of time series inputs (weather, soil moisture, and vegetation indices) and static inputs (soil properties). Let  $\mathbf{x}_t$  represent the vector of time series inputs at time  $t$ , where  $t$  spans from  $sos$  up to the inference point  $T$ . Time series data up to the inference point is represented as  $\mathbf{X}_{sos:T} = (\mathbf{x}_{sos}, \mathbf{x}_{sos+1}, \dots, \mathbf{x}_T)$  and static inputs as  $\mathbf{z}$ . Each training or testing sample  $i$  corresponds to a specific region-season pair  $(r, s)$ . For each training sample  $i = (r, s)$ , the input consists of  $\mathbf{X}_{sos:T}^{(i)}$  and  $\mathbf{z}^{(i)}$ . The target is the end-of-season yield  $Y^{(i)}$  for the corresponding region  $r$  and season  $s$ . The objective is to learn a mapping function  $f$  such that  $Y^{(i)} = f(\mathbf{X}_{sos:T}^{(i)}, \mathbf{z}^{(i)}; \theta) + \epsilon^{(i)}$ , where  $\theta$  represents the model parameters, and  $\epsilon^{(i)}$  is the error term.

During testing, the model gets  $\mathbf{X}_{sos:T}^{(j)}$  from the start of the season ( $sos$ ) up to the inference point  $T$  and static inputs  $\mathbf{z}^{(j)}$  for a new sample  $j = (r', s')$ . The model then forecasts the end-of-season yield  $\hat{Y}^{(j)} = f(\mathbf{X}_{sos:T}^{(j)}, \mathbf{z}^{(j)}; \hat{\theta})$ , where  $\hat{\theta}$  are the model parameters learned during training. Model performance is evaluated by comparing yield forecasts  $\hat{Y}^{(j)}$  with reported yields  $Y^{(j)}$ .

Some details that are ignored in the above formulation:

- $t$  can actually start earlier than  $sos$ , based on spin-up time (e.g., 60 days or 90 days before  $sos$ ).
- The temporal resolution can be different for time series inputs, i.e. the  $+1$  in  $sos + 1$  represents different time units.

### 3.3 Training, validation and test splits

We closely mimic real-world operational settings in the predictor data used, data pre-processing steps and evaluation setup. Specifically, we adopt Leave-One-Year-Out validation (LOYO), which is more appropriate for this type of agricultural problem compared to the random sampling methods commonly used in prior studies (Richetti et al., 2023), including SustainBench (Yeh et al., 2021). 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.

Our library computes the following evaluation metrics: normalized root mean squared error (NRMSE; i.e., the root mean squared error normalized by the average yield of the test set), mean absolute percentage error (MAPE), and R-squared ( $R^2$ ). These metrics are computed by averaging over all cross-validation test folds (which covers the complete dataset for LOYO) and all admin regions with a country.

## 4 Discussion

### 315 4.1 Impact

In addition to the relevance for climate change, food security and the United Nations' sustainable development goals, CY-Bench dataset is relevant to the earth systems science and machine learning research communities due to its comprehensive geographic coverage, capturing diverse agricultural practices and conditions. The inclusion of (indicators derived from) 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 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 modelers who do not necessarily have expertise in yield forecasting, lowering the barrier to entry for advanced yield forecasting research and fostering broader participation and innovation in the field.

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.

Beyond academic research, this dataset can significantly impact policy-making, agricultural planning, and disaster response by enabling the robust evaluation and development of operational models. Researchers, policymakers, farmers, and agribusinesses can benefit from the insights derived from this dataset, leading to better-informed decisions and improved agricultural outcomes.

### 4.2 Limitations

335 We also would like to point out several limitations of CY-Bench that can also be areas for improvement in future iterations.

1. Some limitations stem from the data sources available in the public domain.

- The predictors do not capture certain factors that influence end-of-season yields, such as pests, diseases, and farm management choices. Similarly, CY-Bench excludes socioeconomic factors such as market prices, labor availability, and policy changes. Some of these omitted factors, like crop varieties and management practices, might exhibit spatial or temporal correlations. Consequently, these factors could be partially captured by incorporating spatial or temporal embeddings within a model. Other factors, such as pests, are less likely to be adequately captured in this way. Including these variables could provide a more holistic understanding of yield fluctuations and help develop more robust models. Nevertheless, the availability of such data on a global scale is typically difficult or often not feasible.

- 345 – The crop masks and crop calendars included in CY-Bench are static, i.e. they do not reflect yearly changes. While dynamic, up-to-date masks could improve data quality and subsequently model performance, the availability of consistent, high-quality, global-scale dynamic masks remains limited; a challenge shared by many yield prediction and benchmarking efforts (Kebede et al., 2025; Yeh et al., 2021). Regional, high-resolution products do exist (e.g., the USDA Cropland Data Layer for the U.S. (Boryan et al., 2011)), but their inclusion would break the harmonization across countries that is central to CY-Bench. As CY-Bench is designed as a living resource, we plan to evaluate and potentially incorporate dynamic masks in the future. One candidate would be the recently released MIRCA-OS dataset (Kebede et al., 2025), although its spatial resolution ( $0.0833^\circ$ , i.e., 9 km) is coarser than that of our current crop mask ( $0.0045^\circ$ , i.e., 0.5 km).
- 350
- 355 – The crop calendar information included in CY-Bench is limited to start-of-season and end-of-season dates. More detailed phenological stages, such as emergence and flowering date, are not currently represented. Incorporating these intermediate stages could improve expertise-based feature design and generate more predictive inputs for yield forecasting models.
- 360 – 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.
- 365
- 370 – Crop yield forecasting models could benefit from incorporating weather forecasts (Cunha et al., 2018). In our task definition, 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. We did not include weather forecasts in CY-Bench because the evidence supporting their usefulness is variable (Darbyshire et al., 2020) and there are multiple considerations that need to be discussed before including weather forecasts: a) Observed weather data and forecast data may originate from different sources. b) Some variables, e.g.,  $fPAR$  and  $NDVI$ , have no forecasts. Strategies for addressing these gaps are necessary. c) Weather forecasts from General Circulation Models (GCMs) typically have coarse spatial resolutions (50-400 km grid sizes) and often contain systematic errors or biases that must be adjusted. Bias correction and downscaling techniques must be applied using observed historical records of weather variables, such as precipitation and temperature, to adjust the climate data and better represent local conditions. d) Using weather
- 375

forecasts for yield prediction would also cause error/uncertainty propagation from the weather forecast models leading to an increase in overall uncertainty.

- 380 2. CY-Bench 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.
- 385 3. CY-Bench does not provide process-based crop model outputs, which could serve as valuable input features for machine learning models. Additionally, the current feature aggregation uses fixed time steps rather than adapting to crop growth stages. Access to crop model outputs, which contain information on key phenological state changes, could enable the development of more effective, stage-specific features.
- 390 4. CY-Bench does not provide raw surface reflectance but includes vegetation indices (i.e. NDVI – the most frequently-used index for crop yield forecasting (Schauberger et al., 2020), and  $fPAR$ ), that are strongly correlated with yields (Johnson, 2016). The utility of high-resolution (10-60 m) satellite images (considering spatial information without aggregating to administrative units as performed in CY-Bench) for crop yield forecasting at the sub-national level will result in a massive data size. In addition, the temporal availability of such data is not very long (e.g., Sentinel available since 2014/15). Meanwhile, moderate-resolution satellite images have been used to forecast crop yields in the United States (e.g., You et al. (2017)), demonstrating an advantage over time series data. Future work could investigate the value of satellite images for sub-national crop yield forecasting at a global level.
- 395 5. 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. Also, management strategies might affect multiple seasons; for example, there is evidence of the impact of crop rotation from a previous season on yields Lawes et al. (2022). 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.
- 400

### 4.3 Advancing earth system modeling

We noted a distinct lack of benchmark datasets for agricultural yield forecasting. Still, many recent developments in the related field of crop type mapping using satellite data (Rußwurm et al., 2019; Tseng et al., 2021b; Yeh et al., 2021; Kondmann et al., 405 2021) are leading to exciting progress in the development of methods for extracting meaningful patterns from time series of earth observation data (Rußwurm and Körner, 2018; Rußwurm et al., 2019; Pelletier et al., 2019; Sainte Fare Garnot et al., 2020). Similarly, some studies have shown improved model performance for land cover classification, crop mapping and agricultural yield forecasting using meta-learning and multitask learning (Tseng et al., 2022, 2021a; Kerner et al., 2020). CY-Bench includes

time series of crop productivity or vegetation health indicators from earth observation as predictors, and can therefore be  
410 combined with existing crop mapping benchmark datasets to explore such approaches.

Apart from the downstream task of pre-harvest yield forecasting, CY-Bench enables explorations in transfer learning, domain  
adaptation, and representation learning to assess whether models can generalize well across diverse geographic and climatic  
conditions. We envision at least four directions for future research. First, transfer learning methods can be explored to improve  
model generalization ability when training on data-rich regions and deploying to data-sparse regions (Koukos et al., 2024;  
415 Coulibaly et al., 2019; Nowakowski et al., 2021). Second, self-supervised learning could be used to harness the vast amounts of  
unlabeled agricultural data available (Wang et al., 2022; Xu et al., 2024). By training models to recognize patterns and structures  
within this data, we can build robust representations that capture essential features of agricultural systems. 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,  
420 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 invariance learning focuses on identifying and utilizing  
stable variables across different environments to ensure robustness and generalization (Mitrovic et al., 2020; Neophytides et al.,  
2024). For example, soil quality and basic climatic factors like temperature and precipitation may have stable relationships with  
425 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.

## 5 Conclusions

Innovative, data-driven approaches are crucial for enhancing the resilience of food systems to climate change and extreme events,  
which is essential for achieving the United Nations' Sustainable Development Goal 2 of Zero Hunger (Schneider et al., 2023b).  
430 By providing a well-curated dataset designed for the consistent development and evaluation of large-scale crop yield predictions,  
CY-Bench is a step forward in improving the accuracy of yield forecasting. 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 the decisions of farmers, policymakers and commodity  
traders worldwide. In summary, with CY-Bench we aim to contribute to advancing earth system sciences by providing a critical  
435 resource for modeling the effects of climate change, extreme events, and environmental variability on crop yields, while also  
supporting machine learning research in time series forecasting, transfer learning, and domain adaptation techniques.

## 6 Code and data availability

The complete code base encompassing data pre-processing, tools for model construction, training, evaluation, and data/  
metric visualization routines is available through our publicly accessible GitHub repository: <https://github.com/WUR-AI/>

440 AgML-CY-Bench/ (Paudel et al., 2025b). A summarizing overview can be found on <https://cybench.agml.org/>. We additionally  
provide a Python package `cybench` that can be installed via the repository to load the dataset and run CY-Bench. The dataset  
is available in Zenodo at <https://doi.org/10.5281/zenodo.11502142> (Paudel et al., 2025a) and is comprehensively documented  
using the framework of Data Cards. Each individual dataset subset is accompanied by a dedicated Data Card located within the  
data\_preparation directory of our repository (Paudel et al., 2025b). The CY-Bench dataset and the python library are licensed  
445 under EUPL-1.2, which is compatible with all of the licenses for the datasets included.

## Appendix A

*Author contributions.* DP: Conceptualization, Methodology, Project administration, Investigation, Software, Writing - original draft preparation,  
Writing - review & editing

MK: Project administration, Software, Validation, Writing - original draft preparation, Writing - review & editing

450 SOA: Data curation, Software, Validation, Writing - original draft preparation, Writing - review & editing

HB: Investigation, Software, Visualization, Writing - review & editing

RVB: Software, Validation, Writing - review & editing

AP: Software, Validation, Writing - review & editing

PP: Data curation, Validation, Writing - original draft preparation, Writing - review & editing

455 AS: Data curation, Validation, Writing - review & editing

WA: Data curation, Validation, Writing - original draft preparation, Writing - review & editing

MVB: Validation, Writing - original draft preparation, Writing - review & editing

AC: Validation, Writing - review & editing

OE: Data curation, Validation, Writing - original draft preparation, Writing - review & editing

460 RH: Data curation, Project administration, Validation, Writing - review & editing

RL: Data curation, Validation, Writing - review & editing

DL: Data curation, Validation, Writing - review & editing

IL: Data curation

DM: Software, Validation

465 MM: Data curation, Validation, Writing - original draft preparation, Writing - review & editing

JMM: Data curation, Writing - review & editing

SM: Data curation, Validation, Writing - original draft preparation, Writing - review & editing

JR: Data curation, Writing - original draft preparation, Writing - review & editing

ACR: Validation, Writing - original draft preparation, Writing - review & editing

470 RS: Data curation, Validation, Writing - original draft preparation, Writing - review & editing

GS: Data curation, Validation, Writing - review & editing

VS: Validation, Writing - original draft preparation, Writing - review & editing

RDSNJ: Data curation, Validation, Writing - review & editing

AKS: Data curation, Validation, Writing - original draft preparation, Writing - review & editing  
475 RS: Writing - original draft preparation, Manuscript writing and review  
LS: Validation, Writing - original draft preparation, Writing - review & editing  
PV: Data curation, Validation, Writing - review & editing  
ADW: Data curation, Validation, Writing - review & editing  
MZ: Data curation, Validation, Writing - review & editing  
480 INA: Funding acquisition, Supervision, Validation, Writing - review & editing

*Competing interests.* The authors declare that they have no conflict of interest.

*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  
485 Hoek, Filip Szabo, Francesco Collivignarelli, Xiaomao Lin, Toshi Iizumi, Peng Fu, Prakriti Bista, Paresh Shirsath, Soora Naresh Kumar, Sibiri  
Traore and Javier Garcia Navarro in the design and implementation of the benchmark and preparation of the manuscript.

This work was partially supported by the WUR Research Investment Theme on Data-Driven Discoveries in a Changing Climate, and the Digital Europe Programme under Grant agreement AgrifoodTEF - Test and Experiment Facilities for the Agri-Food Domain (ID 101100622).

## References

- 490 Ambikapathi, R., Schneider, K. R., Davis, B., Herrero, M., Winters, P., and Fanzo, J. C.: Global food systems transitions have enabled affordable diets but had less favourable outcomes for nutrition, environmental health, inclusion and equity, *Nature Food*, 3, 764–779, <https://doi.org/10.1038/s43016-022-00588-7>, 2022.
- Argentina: Ministerio de Agricultura, Ganadería y Pesca. "Estimaciones Agrícolas", <https://datosestimaciones.magyp.gob.ar/reportes.php?reporte=Estimaciones>, accessed: 2016-04-29, 2016.
- 495 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>, accessed: 2024-03-05, 2024.
- Basso, B. and Liu, L.: Seasonal crop yield forecast: Methods, applications, and accuracies, in: *Advances in Agronomy*, vol. 154, pp. 201–255, Elsevier, <https://doi.org/10.1016/bs.agron.2018.11.002>, 2019.
- Batjes, N. H.: Harmonized soil property values for broad-scale modelling (WISE30sec) with estimates of global soil carbon stocks, *Geoderma*, 500 269, 61–68, <https://doi.org/10.1016/j.geoderma.2016.01.034>, 2016.
- Becker-Reshef, I., Justice, C., Barker, B., Humber, M., Rembold, F., Bonifacio, R., Zappacosta, M., Budde, M., Magadzire, T., Shitote, C., Pound, J., Constantino, A., Nakalembe, C., Mwangi, K., Sobue, S., Newby, T., Whitcraft, A., Jarvis, I., and Verdin, J.: Strengthening agricultural decisions in countries at risk of food insecurity: The GEOGLAM Crop Monitor for Early Warning, *Remote Sensing of Environment*, 237, 111 553, <https://doi.org/10.1016/j.rse.2019.111553>, 2020.
- 505 Becker-Reshef, I., Barker, B., Whitcraft, A., Oliva, P., Mobley, K., Justice, C., and Sahajpal, R.: Crop type maps for operational global agricultural monitoring, *Scientific Data*, 10, 172, <https://doi.org/10.1038/s41597-023-02047-9>, 2023.
- Benos, L., Tagarakis, A. C., Dolias, G., Berruto, R., Kateris, D., and Bochtis, D.: Machine learning in agriculture: A comprehensive updated review, *Sensors*, 21, 3758, <https://doi.org/10.3390/s21113758>, 2021.
- Boogaard, H., Schubert, J., De Wit, A., Lazebnik, J., Hutjes, R., and Van der Grijn, G.: Agrometeorological indicators from 1979 to present 510 derived from reanalysis, *Climate Data Store - Copernicus Climate Change Service*, <https://doi.org/10.24381/cds.6c68c9bb>, 2022.
- Boryan, C., Yang, Z., Mueller, R., and Craig, M.: Monitoring US agriculture: the US department of agriculture, national agricultural statistics service, cropland data layer program, *Geocarto International*, 26, 341–358, 2011.
- 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>, accessed: 2024-02-06, 2022.
- 515 Burnell, R., Schellaert, W., Burden, J., Ullman, T. D., Martinez-Plumed, F., Tenenbaum, J. B., Rutar, D., Cheke, L. G., Sohl-Dickstein, J., Mitchell, M., Kiela, D., Shanahan, M., Voorhees, E. M., Cohn, A. G., Leibo, J. Z., and Hernandez-Orallo, J.: Rethink reporting of evaluation results in AI, *Science*, 380, 136–138, <https://doi.org/10.1126/science.adf6369>, 2023.
- Chen, B. and Villoria, N. B.: Foreign yield shocks and domestic price variability: the case of maize in developing countries, *Environmental Research Letters*, 17, 124 044, <https://doi.org/10.1088/1748-9326/aca7d5>, 2022.
- 520 China: National Bureau of Statistics of China. National Data Portal, <https://data.stats.gov.cn>, accessed: 2024-02-18, 2024.
- Chlingaryan, A., Sukkarieh, S., and Whelan, B.: Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review, *Computers and Electronics in Agriculture*, 151, 61–69, <https://doi.org/10.1016/j.compag.2018.05.012>, 2018.
- Coulibaly, S., Kamsu-Foguem, B., Kamissoko, D., and Traore, D.: Deep neural networks with transfer learning in millet crop images, 525 *Computers in industry*, 108, 115–120, 2019.

- Cunha, R. L., Silva, B., and Netto, M. A.: A scalable machine learning system for pre-season agriculture yield forecast, in: 2018 IEEE 14th international conference on e-science (e-Science), pp. 423–430, IEEE, 2018.
- Darbyshire, R., Crean, J., Cashen, M., Anwar, M. R., Broadfoot, K. M., Simpson, M., Cobon, D. H., Pudmenzky, C., Kouadio, L., and Kodur, S.: Insights into the value of seasonal climate forecasts to agriculture, *Australian Journal of Agricultural and Resource Economics*, 64, 1034–1058, <https://doi.org/10.1111/1467-8489.12389>, 2020.
- 530 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.
- Duden, C., Nacke, C., and Offermann, F.: German yield and area data for 11 crops from 1979 to 2021 at a harmonized spatial resolution of 397 districts, *Scientific Data*, 11, <https://doi.org/10.1038/s41597-024-02951-8>, 2024.
- 535 Dueben, P. D., Schultz, M. G., Chantry, M., Gagne, D. J., Hall, D. M., and McGovern, A.: Challenges and Benchmark Datasets for Machine Learning in the Atmospheric Sciences: Definition, Status, and Outlook, *Artificial Intelligence for the Earth Systems*, 1, <https://doi.org/10.1175/AIES-D-21-0002.1>, 2022.
- EC-JRC: Elaboration of Van Tricht et al, 2023, 2024.
- Ennaji, O., Vergutz, L., and El Allali, A.: Machine learning in nutrient management: A review, *Artificial Intelligence in Agriculture*, 9, 1–11, <https://doi.org/10.1016/j.aiia.2023.06.001>, 2023.
- 540 Ennaji, O., Baha, S., Vergutz, L., and El Allali, A.: Gradient boosting for yield prediction of elite maize hybrid ZhengDan 958, *PloS one*, 19, e0315493, <https://doi.org/10.1371/journal.pone.0315493>, 2024.
- Fan, J., Bai, J., Li, Z., Ortiz-Bobea, A., and Gomes, C.: 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, <https://doi.org/10.1609/aaai.v36i11.21444>, 2022.
- 545 Fanzo, J.: Kicking around in the wreck, *PLOS Climate*, 3, e0000401, <https://doi.org/10.1371/journal.pclm.0000401>, 2024.
- FAO: Crop Calendar Food and Agriculture Organization of the United Nations, <https://cropcalendar.apps.fao.org>, accessed: 2025-02-11, 2025.
- Fernandez-Beltran, R., Baidar, T., Kang, J., and Pla, F.: Rice-Yield Prediction with Multi-Temporal Sentinel-2 Data and 3D CNN: A Case Study in Nepal, *Remote Sensing*, 13, 1391, 2021.
- 550 Franch, B., Cintas, J., Becker-Reshef, I., Sanchez-Torres, M. J., Roger, J., Skakun, S., Sobrino, J. A., Van Tricht, K., Degerickx, J., Gilliams, S., et al.: Global crop calendars of maize and wheat in the framework of the WorldCereal project, *GIScience & Remote Sensing*, 59, 885–913, <https://doi.org/10.1080/15481603.2022.2079273>, 2022.
- Frieler, K., Schauburger, B., Arneth, A., Balkovič, J., Chryssanthacopoulos, J., Deryng, D., Elliott, J., Folberth, C., Khabarov, N., Müller, C., et al.: Understanding the weather signal in national crop-yield variability, *Earth's Future*, 5, 605–616, <https://doi.org/10.1002/2016EF000525>, 2017.
- 555 Gavasso-Rita, Y. L., Papalexiou, S. M., Li, Y., Elshorbagy, A., Li, Z., and Schuster-Wallace, C.: Crop models and their use in assessing crop production and food security: A review, *Food and Energy Security*, 13, e503, <https://doi.org/10.1002/fes3.503>, 2023.
- Good, D. L. and Irwin, S. H.: Understanding USDA corn and soybean production forecasts: An overview of methods, performance and market impacts, 2003.
- 560 Hatfield, J. L. and Prueger, J. H.: Temperature extremes: Effect on plant growth and development, *Weather and climate extremes*, 10, 4–10, 2015.
- He, D., Wang, E., Wang, J., and Robertson, M.: Data requirement for effective calibration of process-based crop models, *Agricultural and Forest Meteorology*, 234-235, 136–148, <https://doi.org/10.1016/j.agrformet.2016.12.015>, 2017.

- Hengl, T., Mendes de Jesus, J., Heuvelink, G. B., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., and Kempen, B.: SoilGrids250m: Global gridded soil information based on machine learning, *PLoS one*, 12, e0169748, <https://doi.org/10.1371/journal.pone.0169748>, 2017.
- Hoffmann, H., Zhao, G., Asseng, S., Bindi, M., Biernath, C., Constantin, J., Coucheney, E., Dechow, R., Doro, L., Eckersten, H., et al.: Impact of spatial soil and climate input data aggregation on regional yield simulations, *PloS one*, 11, e0151782, 2016.
- Höhl, A., Ofori-Ampofo, S., Obadic, I., Fernández-Torres, M.-Á., Salih Kuzu, R., and Zhu, X.: USCC: A Benchmark Dataset for Crop Yield Prediction under Climate Extremes, in: *EGU General Assembly 2023*, <https://doi.org/10.5194/egusphere-egu23-15540>, eGU23-15540, 2023.
- Huete, A., Didan, K., Miura, T., Rodriguez, E. P., Gao, X., and Ferreira, L. G.: Overview of the radiometric and biophysical performance of the MODIS vegetation indices, *Remote sensing of environment*, 83, 195–213, 2002.
- ICRISAT Mali: Compagnie Malienne pour le Developpement des Textiles (CMDT) Monitoring Operational Maize Yield Statistics, <https://dataverse.icrisat.org/dataverse/icrisat/>, accessed: 2025-02-11, 2018.
- India: ICRISAT. District Level Database, <http://data.icrisat.org/dld/src/crops.html>, accessed: 2024-02-09, 2024.
- Johnson, D. M.: A comprehensive assessment of the correlations between field crop yields and commonly used MODIS products, *International Journal of Applied Earth Observation and Geoinformation*, 52, 65–81, <https://doi.org/10.1016/j.jag.2016.05.010>, 2016.
- Kamilaris, A. and Prenafeta-Boldú, F. X.: Deep learning in agriculture: A survey, *Computers and Electronics in Agriculture*, 147, 70–90, <https://doi.org/10.1016/j.compag.2018.02.016>, 2018.
- Kapoor, S. and Narayanan, A.: Leakage and the reproducibility crisis in machine-learning-based science, *Patterns*, 4, 100804, <https://doi.org/10.1016/j.patter.2023.100804>, 2023.
- Kebede, E. A., Oluoch, K. O., Siebert, S., Mehta, P., Hartman, S., Jägermeyr, J., Ray, D., Ali, T., Brauman, K. A., Deng, Q., et al.: A global open-source dataset of monthly irrigated and rainfed cropped areas (MIRCA-OS) for the 21st century, *Scientific Data*, 12, 208, 2025.
- Kerner, H., Tseng, G., Becker-Reshef, I., Nakalembe, C., Barker, B., Munshell, B., Paliyam, M., and Hosseini, M.: Rapid Response Crop Maps in Data Sparse Regions, in: *ACM SIGKDD Conference on Data Mining and Knowledge Discovery Workshops*, <https://doi.org/10.48550/arXiv.2006.16866>, 2020.
- Khaki, S., Wang, L., and Archontoulis, S. V.: A CNN-RNN framework for crop yield prediction, *Frontiers in Plant Science*, 10, 1750, <https://doi.org/10.3389/fpls.2019.01750>, 2020.
- Kondmann, L., Toker, A., Ruß wurm, M., Camero, A., Peressuti, D., Milcinski, G., Mathieu, P.-P., Longepe, N., Davis, T., Marchisio, G., Leal-Taixé, L., and Zhu, X.: DENETHOR: The DynamicEarthNET dataset for Harmonized, inter-Operable, analysis-Ready, daily crop monitoring from space, in: *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, edited by Vanschoren, J. and Yeung, S., vol. 1, Curran, [https://datasets-benchmarks-proceedings.neurips.cc/paper\\_files/paper/2021/file/5b8add2a5d98b1a652ea7fd72d942dac-Paper-round2.pdf](https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/5b8add2a5d98b1a652ea7fd72d942dac-Paper-round2.pdf), 2021.
- Koukos, A., Jo, H.-W., Sitokonstantinou, V., Tsoumas, I., Kontoes, C., and Lee, W.-K.: Towards Global Crop Maps with Transfer Learning, in: *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, pp. 1540–1545, IEEE, 2024.
- Lawes, R., Mata, G., Richetti, J., Fletcher, A., and Herrmann, C.: Using remote sensing, process-based crop models, and machine learning to evaluate crop rotations across 20 million hectares in Western Australia, *Agronomy for Sustainable Development*, 42, 120, <https://doi.org/10.1007/s13593-022-00851-y>, 2022.

- Lecerf, R., Ceglar, A., López-Lozano, R., Van Der Velde, M., and Baruth, B.: Assessing the information in crop model and meteorological indicators to forecast crop yield over Europe, *Agricultural systems*, 168, 191–202, <https://doi.org/10.1016/j.agry.2018.03.002>, 2019.
- Lee, D., Davenport, F., Shukla, S., Husak, G., Funk, C., Harrison, L., McNally, A., Rowland, J., Budde, M., and Verdin, J.: Maize yield forecasts for Sub-Saharan Africa using Earth Observation data and machine learning, *Global Food Security*, 33, 100643, <https://doi.org/10.1016/j.gfs.2022.100643>, 2022.
- 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.
- Lesk, C., Anderson, W., Rigden, A., Coast, O., Jägermeyr, J., McDermid, S., Davis, K. F., and Konar, M.: Compound heat and moisture extreme impacts on global crop yields under climate change, *Nature Reviews Earth & Environment*, 3, 872–889, <https://doi.org/10.1038/s43017-022-00368-8>, 2022.
- Liakos, K., Busato, P., Moshou, D., Pearson, S., and Bochtis, D.: Machine learning in agriculture: A review, *Sensors*, 18, 2674, <https://doi.org/10.3390/s18082674>, 2018.
- Liliane, T. N. and Charles, M. S.: Factors Affecting Yield of Crops, in: *Agronomy*, edited by Amanullah, chap. 2, IntechOpen, Rijeka, <https://doi.org/10.5772/intechopen.90672>, 2020.
- Lin, F., Guillot, K., Crawford, S., Zhang, Y., Yuan, X., and Tzeng, N.: CropNet: An Open Large-Scale Dataset with Multiple Modalities for Climate Change-aware Crop Yield Predictions, <https://openreview.net/forum?id=lzpHNYhIbr>, submitted to ICLR 2024.
- Liu, Q., Yang, M., Mohammadi, K., Song, D., Bi, J., and Wang, G.: Machine Learning Crop Yield Models Based on Meteorological Features and Comparison with a Process-Based Model, *Artificial Intelligence for the Earth Systems*, 1, <https://doi.org/10.1175/AIES-D-22-0002.1>, 2022.
- López, J., Way, D. A., and Sadok, W.: Systemic effects of rising atmospheric vapor pressure deficit on plant physiology and productivity, *Global Change Biology*, 27, 1704–1720, 2021.
- Ma, Y., Yang, Z., Huang, Q., and Zhang, Z.: Improving the Transferability of Deep Learning Models for Crop Yield Prediction: A Partial Domain Adaptation Approach, *Remote Sensing*, 15, <https://doi.org/10.3390/rs15184562>, 2023.
- Mateo-Sanchis, A., Piles, M., Amorós-López, J., Muñoz-Marí, J., Adsuara, J. E., Moreno-Martínez, Á., and Camps-Valls, G.: Learning main drivers of crop progress and failure in Europe with interpretable machine learning, *International Journal of Applied Earth Observation and Geoinformation*, 104, 102574, <https://doi.org/10.1016/j.jag.2021.102574>, 2021.
- Mehrabi, Z., Delzeit, R., Ignaciuk, A., Levers, C., Braich, G., Bajaj, K., Amo-Aidoo, A., Anderson, W., Balgah, R. A., Benton, T. G., Chari, M. M., Ellis, E. C., Gahi, N. Z., Gaupp, F., Garibaldi, L. A., Gerber, J. S., Godde, C. M., Grass, I., Heimann, T., Hirons, M., Hoogenboom, G., Jain, M., James, D., Makowski, D., Masamha, B., Meng, S., Monprapussorn, S., Müller, D., Nelson, A., Newlands, N. K., Noack, F., Oronje, M., Raymond, C., Reichstein, M., Rieseberg, L. H., Rodriguez-Llanes, J. M., Rosenstock, T., Rowhani, P., Sarhadi, A., Seppelt, R., Sidhu, B. S., Snapp, S., Soma, T., Sparks, A. H., Teh, L., Tigchelaar, M., Vogel, M. M., West, P. C., Wittman, H., and You, L.: Research priorities for global food security under extreme events, *One Earth*, 5, 756–766, <https://doi.org/10.1016/j.oneear.2022.06.008>, 2022.
- Meroni, M., Fasbender, D., Rembold, F., Atzberger, C., and Klisch, A.: Near real-time vegetation anomaly detection with MODIS NDVI: Timeliness vs. accuracy and effect of anomaly computation options, *Remote sensing of environment*, 221, 508–521, <https://doi.org/10.1016/j.rse.2018.11.041>, 2019.
- Meroni, M., Waldner, F., Seguini, L., Kerdiles, H., and Rembold, F.: Yield forecasting with machine learning and small data: What gains for grains?, *Agricultural and Forest Meteorology*, 308, 108555, <https://doi.org/10.1016/j.agrformet.2021.108555>, 2021.
- Mexico: INEGI. Agricultural Census and Survey Data, <https://www.inegi.org.mx>, accessed: 2024-04-10, 2019.

- Meyer, H. and Pebesma, E.: Machine learning-based global maps of ecological variables and the challenge of assessing them, *Nature Communications*, 13, 2208, <https://doi.org/10.1038/s41467-022-29838-9>, 2022.
- 640 Minoli, S., Egli, D. B., Rolinski, S., and Müller, C.: Modelling cropping periods of grain crops at the global scale, *Global and Planetary Change*, 174, 35–46, <https://doi.org/https://doi.org/10.1016/j.gloplacha.2018.12.013>, 2019.
- Miralles, D. G., Bonte, O., Koppa, A., Villanueva, O. B., Tronquo, E., Zhong, F., Beck, H., Hulsman, P., Dorigo, W., Verhoest, N. E., et al.: GLEAM4: global land evaporation dataset at 0.1 resolution from 1980 to near present, Under review, 2024.
- 645 Mitrovic, J., McWilliams, B., Walker, J., Buesing, L., and Blundell, C.: Representation learning via invariant causal mechanisms, arXiv preprint arXiv:2010.07922, 2020.
- Nakalembe, C. and Kerner, H.: Considerations for AI-EO for agriculture in Sub-Saharan Africa, *Environmental Research Letters*, 18, 041 002, <https://doi.org/10.1088/1748-9326/acc476>, 2023.
- Neophytides, S. P., Tsoumas, I., Tsalakou, A., Christoforou, M., Mavrovouniotis, M., Eliades, M., Papoutsas, C., Kontoes, C., and Hadjimitsis, D. G.: Invariant learning as a pathway to robust potato yield prediction, in: *Remote Sensing for Agriculture, Ecosystems, and Hydrology XXVI*, vol. 13191, pp. 33–42, SPIE, 2024.
- 650 Nowakowski, A., Mrziglod, J., Spiller, D., Bonifacio, R., Ferrari, I., Mathieu, P. P., Garcia-Herranz, M., and Kim, D.-H.: Crop type mapping by using transfer learning, *International Journal of Applied Earth Observation and Geoinformation*, 98, 102 313, 2021.
- Oikonomidis, A., Catal, C., and Kassahun, A.: Deep learning for crop yield prediction: a systematic literature review, *New Zealand Journal of Crop and Horticultural Science*, pp. 1–26, <https://doi.org/10.1080/01140671.2022.2032213>, 2022.
- 655 Paudel, D., Boogaard, H., de Wit, A., Janssen, S., Osinga, S., Pylianidis, C., and Athanasiadis, I. N.: Machine learning for large-scale crop yield forecasting, *Agricultural Systems*, 187, 103 016, <https://doi.org/10.1016/j.agsy.2020.103016>, 2021.
- Paudel, D., Boogaard, H., de Wit, A., van der Velde, M., Claverie, M., Nisini, L., Janssen, S., Osinga, S., and Athanasiadis, I. N.: Machine learning for regional crop yield forecasting in Europe, *Field Crops Research*, 276, 108 377, <https://doi.org/10.1016/j.fcr.2021.108377>, 2022.
- 660 Paudel, D., Marcos, D., de Wit, A., Boogaard, H., and Athanasiadis, I. N.: A weakly supervised framework for high-resolution crop yield forecasts, *Environmental Research Letters*, 18, 094 062, 2023.
- Paudel, D., Kallenberg, M., Ofori-Ampofo, S., Baja, H., van Bree, R., Potze, A., Poudel, P., Saleh, A., Anderson, W., von Bloh, M., Castellano, A., Ennaji, O., Hamed, R., Laudien, R., Lee, D., Luna, I., Masiliūnas, D., Meroni, M., Mutuku, J. M., Mkuhlani, S., Richetti, J., Ruane, A. C., Sahajpal, R., Shuai, G., Sitokonstantinou, V., de Souza Noia Junior, R., Srivastava, A. K., Strong, R., Sweet, L.-b., Vojnović, P.,
- 665 de Wit, A., Zachow, M., and Athanasiadis, I. N.: GitHub repository CY-Bench, <https://github.com/WUR-AI/AgML-CY-Bench>, accessed: 2024-06-05, 2025b.
- Paudel, D., Kallenberg, M., Ofori-Ampofo, S., Baja, H., van Bree, R., Potze, A., Poudel, P., Saleh, A., Anderson, W., von Bloh, M., Castellano, A., Ennaji, O., Hamed, R., Laudien, R., Lee, D., Luna, I., Masiliūnas, D., Meroni, M., Mutuku, J. M., Mkuhlani, S., Richetti, J., Ruane, A. C., Sahajpal, R., Shuai, G., Sitokonstantinou, V., de Souza Noia Junior, R., Srivastava, A. K., Strong, R., Sweet, L.-b., Vojnović, P.,
- 670 de Wit, A., Zachow, M., and Athanasiadis, I. N.: CY-Bench: A comprehensive benchmark dataset for subnational crop yield forecasting, <https://doi.org/10.5281/zenodo.11502142>, 2025a.
- Paudel, D. R., Marcos, D., de Wit, A., Boogaard, H., and Athanasiadis, I. N.: A weakly supervised framework for high resolution crop yield forecasts, *Environmental Research Letters*, 18, 094 062, <https://doi.org/10.1088/1748-9326/acf50e>, 2023b.
- Pelletier, C., Webb, G. I., and Petitjean, F.: Temporal Convolutional Neural Network for the Classification of Satellite Image Time Series, *Remote Sensing*, 11, <https://doi.org/10.3390/rs11050523>, 2019.
- 675

- Poggio, L., De Sousa, L. M., Batjes, N. H., Heuvelink, G. B., Kempen, B., Ribeiro, E., and Rossiter, D.: SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty, *Soil*, 7, 217–240, <https://doi.org/10.5194/soil-7-217-2021>, 2021.
- Portmann, F. T., Siebert, S., and Döll, P.: MIRCA2000—Global monthly irrigated and rainfed crop areas around the year 2000: A new high-resolution data set for agricultural and hydrological modeling, *Global biogeochemical cycles*, 24, 2010.
- 680 Potter, N. A.: rnasqs: 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*, 2019.
- Priyatikanto, R., Lu, Y., Dash, J., and Sheffield, J.: Improving generalisability and transferability of machine-learning-based maize yield prediction model through domain adaptation, *Agricultural and Forest Meteorology*, 341, 109652, <https://doi.org/10.1016/j.agrformet.2023.109652>, 2023.
- 685 Pushkarna, M., Zaldivar, A., and Kjartansson, O.: Data Cards: Purposeful and transparent dataset documentation for responsible AI, in: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1776–1826, <https://doi.org/10.1145/3531146.3533231>, 2022.
- Richards, M., Kirichenko, P., Bouchacourt, D., and Ibrahim, M.: Does Progress On Object Recognition Benchmarks Improve Real-World Generalization?, <https://doi.org/10.48550/arXiv.2307.13136>, 2023.
- 690 Richetti, J., Diakogianis, F. I., Bender, A., Colaço, A. F., and Lawes, R. A.: 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, <https://doi.org/10.1016/j.compag.2023.107642>, 2023.
- Rodell, M., Houser, P., Jambor, U., Gottschalck, J., Mitchell, K., Meng, C.-J., Arsenault, K., Cosgrove, B., Radakovich, J., Bosilovich, M., et al.: The global land data assimilation system, *Bulletin of the American Meteorological society*, 85, 381–394, <https://doi.org/10.1175/BAMS-85-3-381>, 2004.
- 695 Rolnick, D., Aspuru-Guzik, A., Beery, S., Dilkina, B., Donti, P. L., Ghassemi, M., Kerner, H., Monteleoni, C., Rolf, E., Tambe, M., and White, A.: Application-Driven Innovation in Machine Learning, <https://doi.org/10.48550/arXiv.2403.17381>, 2024.
- 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.
- 700 Rußwurm, M., Lefèvre, S., and Körner, M.: BreizhCrops: A Satellite Time Series Dataset for Crop Type Identification, *ArXiv*, <https://doi.org/10.48550/arXiv.1905.11893>, 2019.
- Rußwurm, M. and Körner, M.: Multi-Temporal Land Cover Classification with Sequential Recurrent Encoders, *ISPRS International Journal of Geo-Information*, 7, <https://doi.org/10.3390/ijgi7040129>, 2018.
- 705 Sacks, W. J., Deryng, D., Foley, J. A., and Ramankutty, N.: Crop planting dates: an analysis of global patterns, *Global ecology and biogeography*, 19, 607–620, 2010.
- Sainte Fare Garnot, V., Landrieu, L., Giordano, S., and Chehata, N.: Satellite Image Time Series Classification with Pixel-Set Encoders and Temporal Self-Attention, *CVPR*, 2020.
- Schauberger, B., Jägermeyr, J., and Gornott, C.: A systematic review of local to regional yield forecasting approaches and frequently used data resources, *European Journal of Agronomy*, 120, 126153, <https://doi.org/10.1016/j.eja.2020.126153>, 2020.
- 710 Schlenker, W. and Roberts, M. J.: Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change, *Proceedings of the National Academy of Sciences*, 106, 15594–15598, <https://doi.org/10.1073/pnas.0906865106>, 2009.

- Schneider, K. R., Fanzo, J., Haddad, L., Herrero, M., Moncayo, J. R., Herforth, A., Remans, R., Guarin, A., Resnick, D., Covic, N., Béné, C., Cattaneo, A., Aburto, N., Ambikapathi, R., Aytakin, D., Barquera, S., Battersby, J., Beal, T., Molina, P. B., Cafiero, C., Campeau, C., Caron, P., Conforti, P., Damerau, K., Di Girolamo, M., DeClerck, F., Dewi, D., Elouafi, I., Fabi, C., Foley, P., Frazier, T. J., Gephart, J., Golden, C., Fischer, C. G., Hendriks, S., Honorati, M., Huang, J., Kennedy, G., Laar, A., Lal, R., Lidder, P., Loken, B., Marshall, Q., Masuda, Y. J., McLaren, R., Miachon, L., Muñoz, H., Nordhagen, S., Qayyum, N., Saisana, M., Suhardiman, D., Sumaila, U. R., Cullen, M. T., Tubiello, F. N., Vivero-Pol, J.-L., Webb, P., and Wiebe, K.: The state of food systems worldwide in the countdown to 2030, *Nature Food*, 4, 1090–1110, <https://doi.org/10.1038/s43016-023-00885-9>, 2023a.
- 715
- 720 Schneider, M., Schelte, T., Schmitz, F., et al.: EuroCrops: The Largest Harmonized Open Crop Dataset Across the European Union, *Scientific Data*, 10, 612, <https://doi.org/10.1038/s41597-023-02517-0>, 2023b.
- Seguini, L., Klish, A., Meroni, M., 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/](https://agricultural-production-hotspots.ec.europa.eu/data/indicators_fpar/), in preparation, In preparation.
- 725 Seidel, S., Palosuo, T., Thorburn, P., and Wallach, D.: Towards improved calibration of crop models – Where are we now and where should we go?, *European Journal of Agronomy*, 94, 25–35, <https://doi.org/https://doi.org/10.1016/j.eja.2018.01.006>, 2018.
- Singh, K., Singh, D., Goroshi, S., Yashu, S., Kumar, Y., Solanki, S., and Sagar, S.: Crop Yield forecasting under FASAL (Forecasting Agricultural output using Space Agrometeorology and Land based observations), Tech. rep., <https://doi.org/10.13140/RG.2.2.11593.42083>, 2022.
- 730 Stuart, L., Hobbins, M., Niebuhr, E., Ruane, A. C., Pulwarty, R., Hoell, A., Thiaw, W., Rosenzweig, C., Muñoz-Arriola, F., Jahn, M., et al.: Enhancing Global Food Security: Opportunities for the American Meteorological Society, *Bulletin of the American Meteorological Society*, 104, E760—E777, <https://doi.org/10.1175/BAMS-D-22-0106.1>, 2024.
- Sweet, L.-b., Müller, C., Anand, M., and Zscheischler, J.: Cross-Validation Strategy Impacts the Performance and Interpretation of Machine Learning Models, *Artificial Intelligence for the Earth Systems*, 2, <https://doi.org/10.1175/AIES-D-23-0026.1>, 2023.
- 735 Sweet, L. B., Athanasiadis, I. N., van Bree, R., Castellano, A., Martre, P., Paudel, D., Ruane, A. C., and Zscheischler, J.: Transdisciplinary coordination is essential for advancing agricultural modeling with machine learning, *One Earth*, <https://doi.org/https://doi.org/10.1016/j.oneear.2025.101233>, 2025.
- Sweet, L.-b., Athanasiadis, I. N., van Bree, R., Castellano, A., Martre, P., Paudel, D., Ruane, A. C., and Zscheischler, J.: Transdisciplinary coordination is essential for advancing agricultural modeling with machine learning, *One Earth*, Under review.
- 740 Tadesse, T., Senay, G. B., Berhan, G., Regassa, T., and Beyene, S.: Evaluating a satellite-based seasonal evapotranspiration product and identifying its relationship with other satellite-derived products and crop yield: A case study for Ethiopia, *International Journal of Applied Earth Observation and Geoinformation*, 40, 39–54, 2015.
- Tanaka, T., Sun, L., Becker-Reshef, I., Song, X.-P., and Puricelli, E.: Satellite forecasting of crop harvest can trigger a cross-hemispheric production response and improve global food security, *Communications Earth & Environment*, 4, 1–9, <https://doi.org/10.1038/s43247-023-00992-2>, 2023.
- 745 Thomas, R. L. and Uminsky, D.: Reliance on metrics is a fundamental challenge for AI, *Patterns*, 3, 100476, <https://doi.org/10.1016/j.patter.2022.100476>, 2022.
- Tsaftaris, S. A. and Scharr, H.: Sharing the Right Data Right: A Symbiosis with Machine Learning, *Trends in Plant Science*, 24, 99–102, <https://doi.org/10.1016/j.tplants.2018.10.016>, 2019.

- 750 Tseng, G., Kerner, H., Nakalembe, C., and Becker-Reshef, I.: 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), pp. 1111–1120, <https://doi.org/10.1109/CVPRW53098.2021.00122>, 2021a.
- Tseng, G., Zvonkov, I., Nakalembe, C., and Kerner, H.: CropHarvest: A global dataset for crop-type classification, in: Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, edited by Vanschoren, J. and Yeung, S., vol. 1, Curran, [https://datasets-benchmarks-proceedings.neurips.cc/paper\\_files/paper/2021/file/54229abfcfa5649e7003b83dd4755294-Paper-round2.pdf](https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/54229abfcfa5649e7003b83dd4755294-Paper-round2.pdf), 2021b.
- 755 Tseng, G., Kerner, H., and Rolnick, D.: TIML: Task-Informed Meta-Learning for Agriculture, <https://doi.org/10.48550/arXiv.2202.02124>, 2022.
- Tzachor, A., Devare, M., King, B., Avin, S., and Ó hÉigeartaigh, S.: Responsible artificial intelligence in agriculture requires systemic understanding of risks and externalities, *Nature Machine Intelligence*, 4, 104–109, <https://doi.org/10.1038/s42256-022-00440-4>, 2022.
- 760 van der Velde, M. and Nisini, L.: 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, <https://doi.org/10.1016/j.agsy.2018.06.009>, 2019.
- Van Klompenburg, T., Kassahun, A., and Catal, C.: Crop yield prediction using machine learning: A systematic literature review, *Computers and Electronics in Agriculture*, 177, 105 709, <https://doi.org/10.1016/j.compag.2020.105709>, 2020.
- 765 Van Tricht, K., Degerickx, J., Gilliams, S., Zanaga, D., Battude, M., Grosu, A., Brombacher, J., Lesiv, M., Bayas, J. C. L., Karanam, S., et al.: WorldCereal: a dynamic open-source system for global-scale, seasonal, and reproducible crop and irrigation mapping, *Earth System Science Data*, 15, 5491–5515, <https://doi.org/10.5194/essd-15-5491-2023>, 2023.
- Vermote, E.: MOD09CMG MODIS/Terra Surface Reflectance Daily L3 Global 0.05Deg CMG V006., <https://doi.org/10.5067/MODIS/MOD09CMG.006>, 2015.
- 770 Vijverberg, S., Hamed, R., and Coumou, D.: Skillful U.S. Soy Yield Forecasts at Presowing Lead Times, *Artificial Intelligence for the Earth Systems*, 2, <https://doi.org/10.1175/AIES-D-21-0009.1>, 2023.
- Waha, K., Van Bussel, L., Müller, C., and Bondeau, A.: Climate-driven simulation of global crop sowing dates, *Global Ecology and Biogeography*, 21, 247–259, <https://doi.org/https://doi.org/10.1111/j.1466-8238.2011.00678.x>, 2012.
- Wallach, D., Palosuo, T., Thorburn, P., Hochman, Z., Gourdain, E., Andrianasolo, F., Asseng, S., Basso, B., Buis, S., Crout, N., Dibari, C.,  
775 Dumont, B., Ferrise, R., Gaiser, T., Garcia, C., Gayler, S., Ghahramani, A., Hiremath, S., Hoek, S., Horan, H., Hoogenboom, G., Huang, M., Jabloun, M., Jansson, P.-E., Jing, Q., Justes, E., Kersebaum, K. C., Klosterhalfen, A., Launay, M., Lewan, E., Luo, Q., Maestrini, B., Mielenz, H., Moriondo, M., Nariman Zadeh, H., Padovan, G., Olesen, J. E., Poyda, A., Priesack, E., Pullens, J. W. M., Qian, B., Schütze, N., Shelia, V., Souissi, A., Specka, X., Srivastava, A. K., Stella, T., Streck, T., Trombi, G., Wallor, E., Wang, J., Weber, T. K., Weihermüller, L., de Wit, A., Wöhling, T., Xiao, L., Zhao, C., Zhu, Y., and Seidel, S. J.: The chaos in calibrating crop models: Lessons learned from a multi-model  
780 calibration exercise, *Environmental Modelling & Software*, 145, 105 206, <https://doi.org/https://doi.org/10.1016/j.envsoft.2021.105206>, 2021.
- Wang, Y., Albrecht, C. M., Braham, N. A. A., Mou, L., and Zhu, X. X.: Self-supervised learning in remote sensing: A review, *IEEE Geoscience and Remote Sensing Magazine*, 10, 213–247, 2022.
- Watson, P. A. G.: Machine learning applications for weather and climate need greater focus on extremes, *Environmental Research Letters*, 17,  
785 111 004, <https://doi.org/10.1088/1748-9326/ac9d4e>, 2022.
- Xu, Y., Ma, Y., and Zhang, Z.: Self-supervised pre-training for large-scale crop mapping using Sentinel-2 time series, *ISPRS Journal of Photogrammetry and Remote Sensing*, 207, 312–325, 2024.

- 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.
- 790 You, J., Li, X., Low, M., Lobell, D., and Ermon, S.: Deep Gaussian Process for Crop Yield Prediction Based on Remote Sensing Data, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31, <https://doi.org/10.1609/aaai.v31i1.11172>, 2017.
- Zelingher, R. and Makowski, D.: Investigating and forecasting the impact of crop production shocks on global commodity prices, *Environmental Research Letters*, 19, 014 026, <https://doi.org/10.1088/1748-9326/ad0dda>, 2023.
- 795 Zhang, T., van der Wiel, K., Wei, T., Screen, J., Yue, X., Zheng, B., Selten, F., Bintanja, R., Anderson, W., Blackport, R., Glomsrød, S., Liu, Y., Cui, X., and Yang, X.: Increased wheat price spikes and larger economic inequality with 2°C global warming, *One Earth*, 5, 907–916, <https://doi.org/10.1016/j.oneear.2022.07.004>, 2022.