YieldSAT: A Multimodal Benchmark Dataset for High-Resolution Crop Yield Prediction

Miro Miranda[1,2,*], Deepak Pathak[2,*], Patrick Helber[3], Benjamin Bischke[3], Hiba Najjar[1,2],
Francisco Mena[1,2], Cristhian Sanchez[2], Akshay Pai[2], Diego Arenas[2],
Matias Valdenegro-Toro[4], Marcela Charfuelan[2], Marlon Nuske[2], Andreas Dengel[1,2]
[1]RPTU Kaiserslautern-Landau [2]DFKI GmbH [3]Vision Impulse GmbH [4]University of Groningen

Abstract

Crop yield prediction requires substantial data to train scalable models. However, creating yield prediction datasets is constrained by high acquisition costs, heterogeneous data quality, and data privacy regulations. Consequently, existing datasets are scarce, low in quality, or limited to regional levels or single crop types, hindering the development of scalable data-driven solutions. In this work, we release YieldSAT, a large, high-quality, and multimodal dataset for high-resolution crop yield prediction. YieldSAT spans various climate zones across multiple countries, including Argentina, Brazil, Uruguay, and Germany, and includes major crop types, including corn, rapeseed, soybeans, and wheat, across 2,173 expert-curated fields. In total, over 12.2 million yield samples are available, each with a spatial resolution of $10\,\mathrm{m}$. *Each field is paired with multispectral satellite imagery, resulting in 113,555 labeled satellite images, complemented by auxiliary environmental data. We demonstrate the potential of large-scale and high-resolution crop yield prediction as a pixel regression task by comparing various deep learning models and data fusion architectures. Furthermore, we highlight open challenges arising from severe distribution shifts in the ground truth data under real-world conditions. To mitigate this, we explore a domain-informed Deep Ensemble approach that exhibits significant performance gains. The dataset is available at* https://yieldsat.github.io/.

1. Introduction

Digital agriculture has emerged as an essential tool for addressing current challenges in the agricultural sector, providing data-driven solutions for informed decision-making and ultimately for achieving the UN's Sustainable Development Goals (SDGs) [42], specifically SDG 2 (no hunger) and SDG 13 (climate action) [42]. A key component is crop yield prediction at large scale and high spatial resolution, supporting the management and optimization of crop productivity, the implementation of regional policies, insurance concepts, and the adaptation to changing climate conditions [29]. Yield prediction can be considered as an image regression task involving the processing of multimodal time series data. Nevertheless, yield prediction requires large amounts of high-quality data to train data-driven models [25]. In this context, the exponential growth of openly available Remote Sensing (RS) and Earth Observation (EO) data has become an essential driver of recent advances in crop yield prediction. For instance, satellite programs like the *Sentinel-2 (S2) mission* of the *Copernicus Program* continuously deliver imagery in high spatial resolution and high temporal frequency [44]. Jointly, satellite data thoroughly captures crop development from seeding to harvesting by delivering information on soil properties, vegetation, water content, nutrient supply, and plant biochemistry [3, 14, 27, 48, 56].

Although the amount of openly accessible EO data is unlimited, labeled EO datasets are highly scarce, making up only 0.1% of the total volume of unlabeled data [60], a reason why regression models remain largely underexplored [57]. In crop yield prediction, the lack of dedicated datasets is a serious concern, leading to models trained on single crop types, limited geographic locations, and individual years. Such models frequently exhibit severe performance collapse in real-world scenarios, leading to skepticism about their deployment in practice [29, 37].

To fill this gap, we created *YieldSAT*, a high-quality dataset for large-scale, high-resolution crop yield prediction spanning 4 countries, 4 crop types, and 9 years. *YieldSAT* is the first multimodal dataset for crop yield prediction at both field and subfield levels (i.e., pixel with $10\,\mathrm{m}$ resolution), designed for supervised learning using only globally and publicly available input data. We provide an in-depth comparison of benchmark results obtained from multiple Deep Learning (DL) architectures and data fusion methods. Moreover, we highlight open challenges in crop yield pre-

*corresponding authors: {miro.miranda_lorenz, deepak_kumar.pathak}@dfki.de

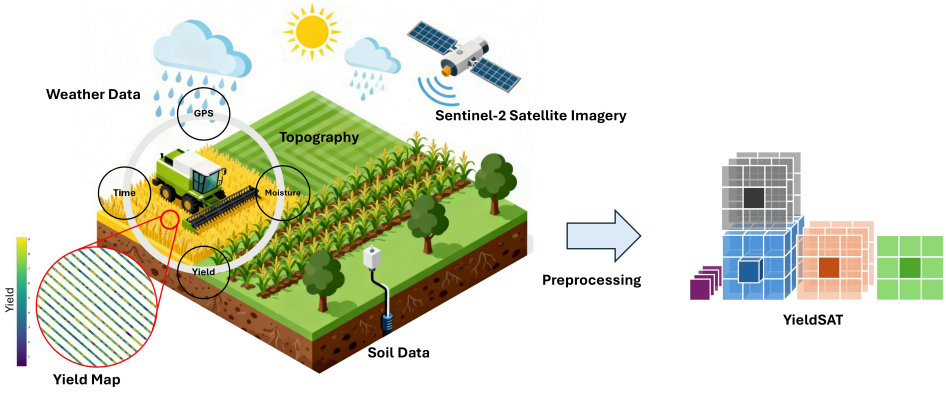


Figure 1. Schematic overview of the *YieldSAT* dataset and the data collection and preprocessing. At harvest, a combine harvester collects point data containing various information (yield, geolocation, time, and moisture), referred to as a yield map. For each yield map, Earth Observation (EO) data is collected (satellite imagery, weather, soil, and topography data). The target yield map and the input data are preprocessed into an ML-ready data format, formulating yield prediction as a pixel-wise regression task.

diction arising from severe distribution shifts in the ground-truth data, which lead to significant performance collapse in DL models. To find explanations and potential mitigation strategies, we propose a domain-informed *Deep Ensemble (DE)* approach [20] and investigate the weight space distribution of the ensemble members. Our approach demonstrates significant improvement relative to the baseline.

By releasing *YieldSAT*, we aim to accelerate digital farming and EO research. We believe this work supports the computer vision community in developing robust image regression methods that produce physically meaningful outputs under challenging, real-world conditions.

2. Related Work

Subfield (i.e., pixel) level crop yield prediction can be considered as a dense, structured image (pixel) regression task, similar to monocular depth estimation [5, 18]. Nevertheless, yield prediction requires processing multimodal data, including long time series with varying temporal and spatial resolutions, making it a multimodal fusion task [24]. Additionally, yield prediction is a physically grounded task that involves predicting biophysical quantities, connecting it to scientific methods like *Physics-Informed Neural Networks (PINNs)* [4, 28, 29, 39] and explainable AI [15, 33]. However, many EO regression applications, like high-resolution yield prediction, remain underexplored, primarily due to the lack of dedicated datasets [57]. Still, yield prediction using DL and EO has gained widespread interest [51]. For this, many methods for yield prediction use multispectral satellite imagery and advanced model architectures, like LSTM and ConvLSTM [12, 36, 46], Transformers [13, 52], and Diffusion Models [30]. Moreover, to process multimodal input data, simple and advanced data fusion methods are used to handle different temporal, spatial, and spectral resolutions [24, 25, 32]. As ground truth data, regional [49] and field-level data [2] are mainly used. Only a few studies use subfield-level yield data [37], mainly because of the high acquisition costs. Consequently, studies are limited to specific regions, crop types, and individual years [9, 35, 49, 53, 58]. Moreover, yield data is often affected by shifts in data distribution, driven by differences in management practices, environmental conditions, and climate variability [40]. Consequently, generalization to unknown years and regions often causes a severe performance reduction [25, 29, 37] and, consequently, skepticism of deploying models into practice. Only a few publicly available yield datasets exist, ready for training DL models and to study yield prediction at large scale. A comparison of available datasets for yield prediction is provided in Tab. 1. For instance, the *SwissYield* [37] dataset contains only 73 fields from a single crop type and country (Switzerland). In addition, the *CropNet* [22] dataset provides only regional-level yield data [11, 22]) for the US, with low temporal and spatial resolution. The data is coupled only with single bands and derivatives from S2.

YieldSAT is, to our knowledge, the first public dataset that combines pixel-level yield maps with multimodal, globally-available EO inputs across multiple crops and multiple countries ready for training DL models. Moreover, only a few studies use DEs [20] for yield prediction [45], primarily for uncertainty estimation. Exploring DEs to explore the impact of distribution shifts, such as evidenced in [17, 54], is an open challenge.

3. Dataset Overview

The *YieldSAT* dataset is a multimodal dataset comprising high-resolution ($10\,\mathrm{m} \times 10\,\mathrm{m}$) yield data, coupled with mul-

Table 1. Comparison of the YieldSAT dataset with other crop yield prediction datasets.

Dataset	Countries	Crops	Years	Fields	Pixel-Level	Resolution (Optical)		Features	Curated
						Spatial	Temporal		
SwissYield [37]	1	2	2017–2021	73	✓	10 m	∼ **5 days**	14	✗
CropNet [22]	1	**4**	2017-2022	0	✗	9 km	∼ 14 days	13	✗
YieldSAT (ours)	**4**	**4**	**2016-2024**	**2,173**	✓	10 m	∼ **5 days**	**72**	✓

Table 2. Overview of the available ground truth yield data. In total, 2,173 labeled fields (yield images) are available.

YieldSAT	Fields					Pixels	Area (ha)		Years	S2 Images
	Corn	Rapeseed	Soybean	Wheat	Total		Total	Average		
Argentina	185	✗	440	126	751	∼5.3 M	∼55,898	74.3	2017-2024	42,931
Brazil	118	✗	293	140	551	∼4.2 M	∼43,125	78.2	2017–2024	19,308
Uruguay	✗	✗	572	✗	572	∼2.1 M	∼32,804	57.3	2018–2022	24,318
Germany	✗	111	✗	188	299	∼0.6 M	∼6,460	21.6	2016–2022	26,998
Summary	**303**	**111**	**1,305**	**454**	**2,173**	**∼12.2 M**	**∼138,288**	**57.8**	**2016-2024**	**113,555**

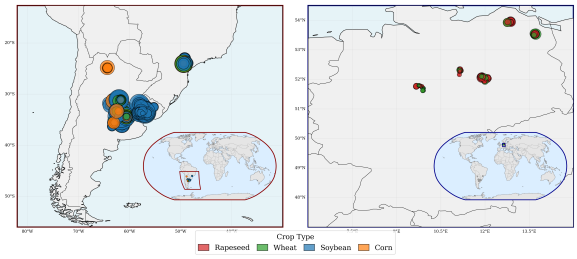


Figure 2. Spatial distribution of the collected yield data for each region and crop type. The yield data is colored by crop type. Left: South America, right: Europe. The marker size indicates the field size. (Map source: [26])

timodal EO data sources. The dataset spans a large area across Argentina, Brazil, Uruguay, and Germany. The available countries are major contributors to the global food production of the selected crop types [41]. In total, the data covers a labeled area of approximately $138\,288\,\mathrm{ha}$ ($\sim 1384\,\mathrm{km}^2$) with crop types of soybean (*Glycine max L.*), corn (*Zea mays*), rapeseed (*Brassica napus subsp. napus*), and wheat (*Triticum aestivum*). In Fig. 2, the geographic locations of the available fields are displayed, colored by crop type. Note that the availability of crop types varies by country. For example, Germany includes rapeseed and wheat, whereas Argentina provides data for corn, soybeans, and wheat. The dataset spans nine years, from 2016 to 2024, thereby covering large climate and yield variability. In total, 2,173 labeled fields are available. Soybean is the most represented crop, with 1,305 fields, while rapeseed is the least represented, with 111 fields. However, notable differences in the field size exist between countries. For example, Brazil exhibits the largest average field size, with $78.8\,\mathrm{ha}$. In contrast, Germany has smaller field sizes, with an average of $26.6\,\mathrm{ha}$. The average field size is $57.8\,\mathrm{ha}$ across the entire dataset. Altogether, the dataset includes approximately 12.2 million yield samples (labeled pixels) with $10\,\mathrm{m}$ resolution each. A detailed description of the available ground truth data is depicted in Tab. 2. Note that pixel counts are nominal and yield measurements are inherently spatial autocorrelated due to environmental factors. Each field is coupled with multimodal data sources, including multispectral satellite imagery from S2, weather data, soil data, and topography information. In total, 72 features are available for each sample.

4. Data Collection

The data collection consists of (1) the ground truth yield data collection, (3) preprocessing, and (2) EO data collection.

4.1. Yield Data Collection

YieldSAT contains subfield-level yield data from combine harvesters, collected at high spatial resolution. During harvest, a combine harvester equipped with yield monitors drives through the field and collects georeferenced data points as point vector data at a consistent frequency. Each data point contains various information, such as the geographic coordinate (latitude and longitude) of each measurement, the amount of wet yield, and the moisture content. Together, all data points form a yield map, a collection of data points in vector format for a single field. A schematic overview of the yield collection and a yield map is given in Fig. 1. Yield maps provide valuable information about the spatial variability and productivity of a single field and enables farmers to identify productivity zones, estimate yield quality, and quantify damages and losses, and serves as a foundation for future activities.

4.2. Yield Data Preprocessing

Raw yield data is commonly collected as georeferenced point vector data in the *shapefile* format. Still, many other data formats exist that can store geospatial yield data. Therefore, if the data is not provided in the shapefile format, a format conversion is performed as a first step. Moreover, yield maps are manually inspected and curated. The curation and other metadata are stored for each yield map, providing information on data quality, location, and other insights relevant to training yield prediction models. Only yield data was considered for further processing that appeared realistic to agricultural experts. Nevertheless, combine harvester yield data is heavily inhomogeneous across farmers, regions, and countries due to the use of different machines, languages, units, and management practices and therefore requires careful data preprocessing [21, 43]. To harmonize the data, a standardized preprocessing pipeline is used. This includes automatic and semi-automatic translation of feature naming over various languages and conversion to the metric system. Additionally, automatic transformation from the Geographic WGS84 to a projected UTM coordinate reference system is performed.

Combine harvester data is often mis-calibrated and associated with sensor errors, positioning inconsistencies, missing information, delays between the grain collection and measurement event, and focuses during turns [50]. Consequently, removing erroneous values related to position, timestamp, yield, moisture, and inactive harvesters is essential to improve data quality and prevent misleading outcomes [21]. For this, based on expert rules, zero yield points and biologically infeasible points are removed. This includes crop-specific maximum yield values. In addition, data points are filtered by three standard deviations ($\pm 3\sigma$), following [43]. Finally, the scaled yield (i.e., dry yield) is calculated based on the provided wet yield, adjusted to a fixed standard moisture, as $y_s = y_w * (1 - m_m/1 - m_s)$, where y_s is the scaled yield (dry yield), y_w is the wet yield, m_m is the measured moisture, and m_s is the standard moisture. The Appendix 6 gives maximum yield values and the standard moisture. The scaled yield is calculated because it is less affected by measurement noise (weather and time of harvest). Additionally, the scaled (dry) yield is the true indicator of the grain output used to estimate revenue potential for farmers, traders, and crop insurances.

4.3. Earth Observation Data Collection

We acquired EO data for every yield map based on a stringent selection criteria: (1) demonstrated or theoretically established influence on crop development and yield, (2) open and freely accessible, (3) global coverage, and (4) high spatial resolution if possible.

4.3.1. Optical & Multispectral Satellite Imagery

For each yield map, S2 Level-2A multispectral satellite imagery, including all 13 spectral bands, was acquired for the period between seeding and harvesting. The yield map boundary serves as a geo-reference for the image acquisition process, resulting in a time series with a temporal resolution of approximately one image every five days. Low-resolution bands were nearest-neighbor upsampled to 10 m to ensure a uniform spatial resolution across all bands. The S2 bands are preserved in the original form, and no specific indices are calculated (e.g., NDVI, NDWI). This increases the flexibility of the dataset for downstream tasks. Additionally, the *Scene Classification Layer (SCL)* was acquired for each time step, providing per-pixel class information for the S2 product at 20 m resolution. In total, the SCL provides 12 class labels, including "vegetated," "non-vegetated," and "clouded".

4.3.2. Additional Data Modalities

We further provide Additional Data Modalities (ADMs) (weather, soil, and topography) to complement and compensate for inconsistencies and shortcomings of the S2 data. For example, S2 often suffers from missing time steps (e.g., due to cloud occlusion), which introduce uncertainty into the model [31]. Weather data for each field were derived from the ECMWF Reanalysis (ERA5) program [14] between seeding and harvesting, at a daily resolution. Soil data was acquired from the *SoilGrids* archive in 250 m resolution [38]. Topography data was acquired from the SRTM mission [7] in 30 m resolution. For soil and topography data, raster images are generated for each feature and upsampled to 10 m resolution using cubic spline interpolation to match the S2 image resolution. For soil, all soil properties are sampled at depths of 0-5, 5-15, 15-30, 30-60, 60-100, and 100-200 cm with the uncertainty provided for every layer. For the topography data, the *RichDEM* [1] library was used for feature engineering to derive additional features, including aspect, curvature, slope, and the *Topographic Wetness Index (TWI)*. The TWI is derived following [19]. A detailed overview of the available data modalities is given in Tab. 3.

Table 3. Overview of all available data modalities in the *YieldSat* dataset and their characteristics.

Modality	Source	Product	Resolution	
			Spatial	Temporal
Multispectral	Sentinel-2 L2A	B01 - Coastal Aerosol	60 m	~ 5 days
		B02 - Blue	10 m	
		B03 - Green	10 m	
		B04 - Red	10 m	
		B05 - Red Edge 1	20 m	
		B06 - Red Edge 2	20 m	
		B07 - Red Edge 3	20 m	
		B08 - NIR	10 m	
		B8A - Narrow NIR	20 m	
		B09 - Water vapour	60 m	
		B11 - SWIR 1	20 m	
		B12 - SWIR 2	20 m	
		Scene Classification Layer	20 m	
Weather	Era5 [14]	Max Temperature	30 km	Daily
		Mean Temperature		
		Min Temperature		
		Total Precipitation		
Soil	SoilGrids [38]	Soil Organic Carbon	250 m	Static
		Nitrogen		
		Cation Exchange Capacity		
		Clay		
		Silt		
		Sand		
		pH		
		Coarse fragments		
Topography	SRTM [7]	Topography (DEM)	30 m	
		Slope		
		Curvature		
		TWI		
		Aspect		

4.4. Rasterization & Data Quality

To align the S2 imagery and the ADMs with the yield data, yield maps are rasterized so that each S2 product pixel aligns spatially with the corresponding yield data pixel across the entire time series. For this, a rasterization grid derived from S2 imagery at 10 m resolution is overlaid on the yield map. All yield points within a given pixel are averaged, yielding a rasterized yield image with the same spatial resolution as the S2 data. Consequently, each pixel in the S2 product is spatially aligned with a pixel in the target yield image. An example time series of S2 imagery with the final

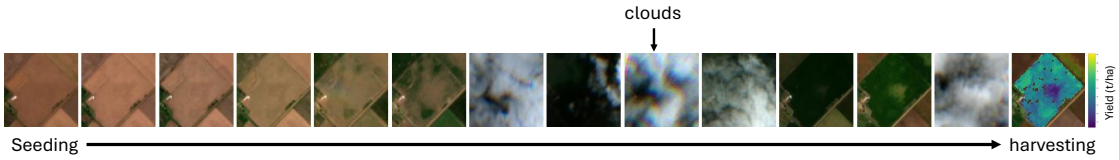


Figure 3. Example satellite time series of a soybean field from seeding (left) to harvesting (right). The last image shows the collected yield at harvest.

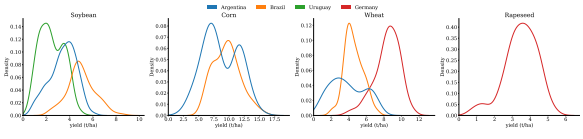


Figure 4. Yield data distribution plots for each crop type and country, averaged per field.

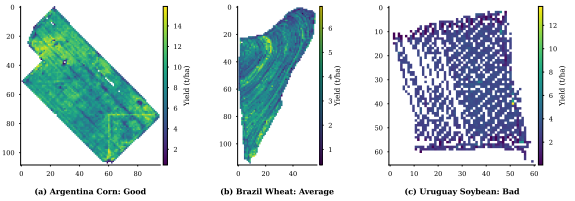


Figure 5. Examples of rasterized yield maps with the curation for each quality level (*good*, *average*, and *bad*).

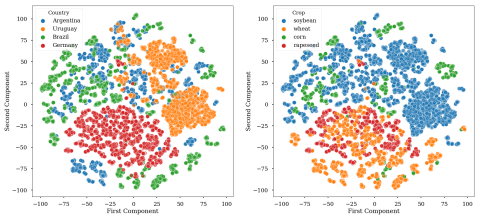


Figure 6. t-SNE plot of the S2 surface reflectance colored by countries (left) and crops (right).

rasterized yield image is depicted in Fig. 3. Averaging per pixel follows state-of-the-art processing for harvester data and preserves both field-level means and subfield patterns. Pixels without vector points are masked and not used for training. Nevertheless, rasterization can introduce variable support sizes for each pixel due to harvester path density, swath width, speed, logging frequency, and positional delay, which are often correlated in space. This can impact yield modeling and introduce spatially correlated uncertainty. To account for variable support sizes and spatially correlated uncertainty in future research, we additionally provide label information for each image: (i) the number of yield points and (ii) the standard deviation, both per $10\,\mathrm{m}$ pixel.

Every yield map contains a quality label from manual expert labeling. Fig. 5 presents randomly selected rasterized yield maps for each quality level (*good*, *average*, and *bad*), highlighting the varying quality levels within the dataset. Low-quality yield maps are often characterized by sparse data distribution, spatial misalignment, or erroneous measurements, such as unrealistically high yield values. Additionally, artifacts may be present, including patterns caused by harvester turns or delays in the measurement process. A more in-depth analysis of the rasterization, data quality, and support sizes is provided in the Appendix (see A.2.1).

In summary, the *YieldSAT* dataset is highly diverse, with distinct yield distributions for each crop type and country. For instance, corn and wheat tend to exhibit the highest yields, with a broad spread, while soybeans and rapeseed generally show lower yields. This is depicted in Fig. 4. More importantly, the yield data distributions are significantly different between countries and crops and between years and regions for a single crop type and country (p-values < 0.0001, see Appendix A.2.2). Additionally, we observe high diversity in the patterns of the data modalities (e.g., surface reflectance for the S2 time series) between countries and crop types. In Fig. 6 a t-Distributed Stochastic Neighbor Embedding (t-SNE) [23] of the surface reflectance of the S2 time series is shown, colored by countries (left) and crops (right). The plot shows that the surface reflectance differs significantly between countries and crops, increasing the difficulty of generalizing between different environments.

4.5. Data Availability

The final dataset is available in two formats: (1) a preprocessed version using an input fusion strategy, described in [12, 36]. Here, the input modalities are aligned via concatenation and temporal and spatial repetition, resulting in a unified time series of 24 time steps, encompassing all available data modalities. The dataset is stored in the *xarray* data format [16] and ready to train DL models. (2) A version where each field is stored jointly with the described input modalities and further meta information. This version enables the development of advanced DL models by offering high flexibility. For both data formats, the preprocessing and clean-

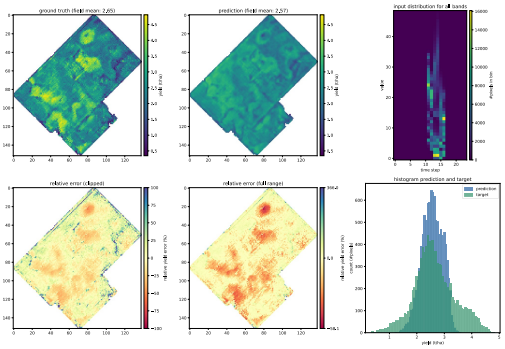


Figure 7. Qualitative results for a single soybean field from Argentina, generated with the 3D-LSTM model. Top left to bottom right: ground truth yield map, predicted yield map, prediction over target plot, input distribution over time, clipped relative pixel-wise error (100%), relative pixel-wise error (full range), histogram of predicted (blue) and target (green) values.

ing were done as described. Further information about the data, limitations, legal and ethical considerations, and data distribution is given in the data sheet in the Appendix (see B).

5. Experiments

In this section, we demonstrate the potential of large-scale, high-resolution crop yield prediction with different DL models using previously published methods. All models are trained at the pixel level, as in the state-of-the-art approaches, treating each pixel as an independent sample. Since pixels are highly autocorrelated due to e.g., shared soil properties, topography, management practices, or microclimate, some methods include spatial neighborhood information by processing a pixel with its surroundings using a 3D-CNN block or a ConvLSTM [47] approach. Such approaches account for spatial dynamics and spatial correlations within the field (image) [32]. To prevent overfitting and information leakage, models were trained using a stratified, grouped 10-fold cross-validation (cv), where pixels are grouped by field and stratified by region. This is done to ensure that pixels from the same field are either in the training or testing set. The metrics are presented as the average across the folds and are used for relative model ranking. All models are evaluated based on their ability to predict yield at both the subfield (i.e., pixel) and field levels. For field-level performance, all pixels in the same field are averaged and compared with the field's averaged ground truth.

Results are reported on subsets per country and crop due to the heterogeneity of the dataset, which is how the data will be used in practice.

We compare models trained only on S2 data and models that further incorporate ADMs using a simple input fusion method [36] and advanced fusion techniques [24].

5.1. Benchmark Results

An overview of benchmark results for different model architectures and selected subsets is provided in Tab. 4 for the R^2-score and for the RMSE in tons/hectare (t/ha). The presented subsets contain key characteristics (crops/countries/quality) while providing key insights and maintaining readability. A full replication of all available subsets is provided in the Appendix (see A.3).

Notably, the results depend strongly on the country and crop type, likely due to differences in the quality of the ground truth data. For instance, soybean in Argentina (ARG-S) exhibits high performance across all models, with a maximum R^2 score of 0.84 and an RMSE of 0.49 t/ha. Additionally, there are significant differences between subfield- and field-level performance, with the field-level exhibiting consistently higher scores, especially for advanced fusion methods. In general, modeling spatial correlation using 3D-CNN blocks (3D-LSTM, 3D-ConvLSTM, AFF) significantly improves the performance compared to modeling each pixel independently. Such approaches even perform considerably well when using only S2 data as input. However, we emphasize that integrating ADMs commonly improves performance compared to using S2 data alone. Nevertheless, the benefit of ADMs depends on the model architecture and the fusion strategy. Related studies demonstrate that optical data, e.g., S2 is crucial for pixel-wise yield prediction [25, 32, 33]. For instance, coupling spatial information (3D-LSTM, 3D-ConvLSTM) with an input fusion strategy even results in a performance reduction, which is related to the difference in temporal and spatial resolution of the input data [32]. Consequently, coupling multimodal data with a more complex architecture requires advanced fusion methods to harmonize different spatial, temporal, and spectral resolutions. A qualitative example of a predicted field is shown in Fig. 7, which depicts high subfield variability and a good match between the input and target distributions, while also highlighting areas of high pixel-wise error.

5.2. Distribution Shift & Deep Ensembles

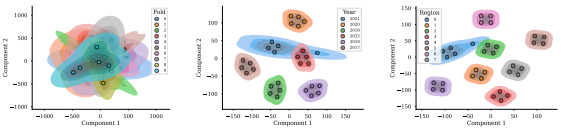
DL models are susceptible to distribution shifts and exhibit overconfidence and degraded performance when exposed to such data [6, 59]. We already reported that the *YieldSAT* dataset is affected by distribution shifts, even between years and regions for a single country (see Appendix A.2.3). In this section, we explore the impact of distribution shift in DL models using a real-world scenario, namely a *Leave-One-Year-Out (LOYO)* and *Leave-One-Region-Out (LORO)* CV experiments. A region is defined by a set of fields belonging to a single farmer or to a local data provider. Subsequently, a model is evaluated on a held-out year or region to

Table 4. Results for the RMSE (t/ha) (↓) and the R^2-score (↑) for different models and datasets. The best model is highlighted in bold, and the best overall score is underlined. ARG = Argentina, BRA = Brazil, GER = Germany, URG = Uruguay. C = corn, R = rapeseed, S = soybean, W = wheat.

Evaluation			Field-Level										Subfield (Pixel)-Level									
			ARG-S		BRA-C		GER-R		GER-W		URG-S		ARG-S		BRA-C		BRA-S		GER-R		GER-W		URG-S	
Modalities	Fusion Method	Model	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
S2	✗	3D-ConvLSTM [13]	0.79	0.55	0.82	0.74	0.81	0.58	0.65	1.12	0.77	0.51	0.65	0.90	0.45	2.13	0.39	0.94	0.49	1.20	0.34	2.40	0.41	1.22
	✗	3D-LSTM [32]	0.77	0.58	0.82	0.74	0.82	0.57	0.54	1.28	0.73	0.56	0.65	0.90	0.46	2.13	0.39	0.93	0.48	1.20	0.30	2.46	0.39	1.23
	✗	LSTM [36]	0.72	0.64	0.75	0.88	0.62	0.83	0.55	2.60	0.66	0.62	0.60	0.96	0.42	2.20	0.34	0.98	0.36	1.33	0.32	2.47	0.37	1.26
	✗	Transformer [12]	0.73	0.63	0.79	0.82	0.75	0.67	0.56	1.26	0.72	0.56	0.62	0.94	0.44	2.16	0.38	0.94	0.44	1.25	0.32	2.43	0.38	1.24
S2 + ADM	Input Fusion	3D-ConvLSTM [13]	0.82	0.52	0.83	0.72	0.78	0.63	0.70	1.03	0.78	0.50	0.68	0.87	0.46	2.12	0.38	0.94	0.42	1.27	0.39	2.30	0.41	1.22
		3D-LSTM [32]	0.76	0.59	0.84	0.71	0.81	0.59	0.62	1.17	0.76	0.54	0.64	0.91	0.46	2.12	0.41	0.93	0.49	1.20	0.37	2.34	0.40	1.22
		LSTM [36]	0.72	0.64	0.81	0.78	0.81	0.58	0.63	1.16	0.72	0.56	0.59	0.98	0.43	2.18	0.33	0.99	0.47	1.21	0.35	2.38	0.39	1.23
		Transformer [12]	0.72	0.64	0.79	0.81	0.76	0.65	0.61	1.19	0.73	0.55	0.58	0.98	0.44	2.17	0.37	0.96	0.45	1.24	0.38	2.32	0.39	1.23
	Feature Fusion	AFF [32]	**0.84**	0.49	**0.84**	0.70	0.80	0.60	0.74	0.96	0.81	**0.46**	**0.73**	**0.79**	0.46	2.12	0.44	0.90	0.49	1.20	0.44	2.20	0.43	1.19
		MMGF [25]	0.82	0.51	0.76	0.86	0.75	0.68	0.77	0.90	0.75	0.53	0.70	0.84	0.42	2.19	0.42	0.92	0.44	1.26	0.44	2.21	0.40	1.22

Table 5. Overview for crop yield prediction at the field level using temporal (Leave-One-Year-Out) and spatial splitting (Leave-One-Region-Out). All models, except the baseline, are defined as a *Deep Ensemble* [20] with 5 ensemble members. All models were trained on S2 data only on subsets from Argentina. DE = Deep Ensemble. The best score is highlighted in bold.

Evaluation		Leave-One-Year-Out		Leave-One-Region-Out	
Dataset	Model	RMSE (↓) t/ha	R^2 (↑) -	RMSE (↓) t/ha	R^2 (↑) -
Soybean	DE-LSTM	0.70 ± 0.25	0.55 ± 0.64	0.65 ± 0.20	0.65 ± 0.87
	DE-3D-LSTM	**0.23** ± 0.11	**0.63** ± 0.40	**0.19** ± 0.14	**0.73** ± 0.53
	Baseline LSTM	0.85 ± 0.23	0.50 ± 0.50	0.72 ± 0.15	0.64 ± 0.15
Corn	DE-LSTM	2.06 ± 0.82	0.46 ± 0.52	1.81 ± 0.86	0.59 ± 5.28
	DE-3D-LSTM	**1.70** ± 0.61	**0.63** ± 0.36	**1.36** ± 0.83	**0.76** ± 2.31
	Baseline LSTM	2.06 ± 1.07	0.46 ± 0.52	2.05 ± 0.75	0.47 ± 1.36
Wheat	DE-LSTM	0.98 ± 0.45	0.79 ± 0.95	1.11 ± 0.47	0.73 ± 1.69
	DE-3D-LSTM	**0.99** ± 0.35	**0.79** ± 1.24	**0.98** ± 0.36	**0.79** ± 1.24
	Baseline LSTM	1.02 ± 0.37	0.77 ± 1.49	1.18 ± 0.26	0.70 ± 0.38



(a) 10-Fold CV (b) LOYO (c) LORO

Figure 8. t-SNE of the network parameters for the Deep Ensemble model for different CV scenarios. Left: standard 10-CV, center: LOYO, right: LORO. The weights are colored by fold.

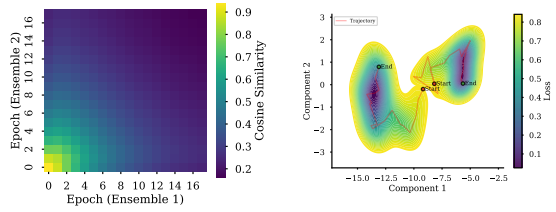
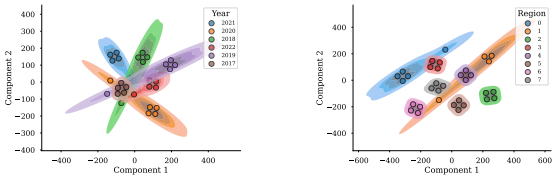


Figure 9. Visualization of the weight space diversity during model training. Left: Cosine similarity between two ensemble members during training. Right: PCA plot of the weight space during training, together with the loss. The trajectory in the weight space is highlighted in red from start to end.



(a) LOYO (b) LOFO

Figure 10. t-SNE of the network parameters for the Deep Ensemble model trained with prior knowledge (3D-LSTM) under distribution shift. Left: LOYO, right: LORO.

assess its generalization to unknown distributions. We focus on LOYO and LORO experiments, as this reflects how the dataset will be used in practice. Nevertheless, cross-crop and cross-country evaluations should be investigated in future work. In preliminary experiments, however, these settings led to model collapse without domain adaptation due to severe data heterogeneity.

To explore the impact of distribution shifts, we employ a DE approach [20]. Specifically, we evaluate an LSTM model in an ensemble setting and the 3D-LSTM [32] that incorporates spatial correlations. This architecture was chosen because of its good performance in the earlier experiments and for its lightweight design compared to the advanced fusion methods, since training DEs is computationally expensive. Moreover, the model is trained solely on S2 data based on insights from our earlier findings. The results are shown on crops from Argentina only as the largest and best-curated subset (see Appendix A.2.1) with strong year/region shifts, which provides statistical power for computationally expensive experiments and weight space analysis under real-world conditions. The results are presented in Tab. 5 and confirm the hypothesis that model performance significantly decreases under distributional shift. For instance, for the LOYO experiment, an overall reduction in the R^2-score of 22 p.p. is reported compared to the standard CV experiments for soybean (see results for LSTM in Tab. 4). Likewise, for the LORO experiment, a severe reduction of 8 p.p. is shown. In the Appendix we show that the im-

pact of distribution shifts is consistent over all model architectures and datasets (see Appendix A.3). In contrast, the DE demonstrates improved performance compared to the baseline model in both settings, as shown in Tab. 5. In the LOYO setting, the DE improves by 5 p.p. R^2 over the baseline model for soybean. Likewise, the DE model exhibits a 12 p.p. increase over the baseline in the LORO experiment for corn. This underscores the superiority of ensemble methods, which are consistently more robust to distribution shifts. Interestingly, including additional inductive bias in the form of spatial locality (3D-LSTM) further increases the gap between the baseline and the ensemble approach. In the LORO scenario, an improvement of 29 p.p. is achieved for corn, resulting in an overall R^2 of 0.76. This score is almost equal to the R^2 score of the baseline model in the standard CV scenario. Similarly, in the LOYO setting, an improvement of 17 p.p. in the R^2-score is observed compared to the baseline model. This improvement is proportionally higher than the standard CV experiment (see Tab. 4).

5.3. Distribution shift & Weight Space Diversity

To explore the degrading performance and the difference between probabilistic and deterministic models, we analyze the model weights in Fig. 8. The plot illustrates a low-dimensional embedding of the trained model weights of the DE model using a t-SNE. The weights are displayed for the standard 10-fold CV (left), LOYO (center), and LORO (right) scenarios. Additionally, the weights are colored by folds. For example, in the LOYO scenario, a model is colored by the year in the validation set. The plot reveals interesting insights. While the weight distributions of the ensemble members overlap entirely in the standard 10-fold CV, a clear separation between model parameters is observed under distribution shift (LOYO and LORO). Each fold forms a distinct cluster in weight space with no overlapping. We conclude two main things from this. First, we argue that this may explain the poor performance under data shift, as the model is less capable of generalizing to unknown data distributions due to the separation in weight space. Secondly, DEs explore multiple modes in weight space that may explain the better performance compared to the deterministic baseline, which only explores single modes [8]. This is underlined by Fig. 9. The plot illustrates the trajectory in weight space during training. The left row shows the cosine similarity between two ensemble members over the training epochs. Each comparison shows that the similarity in weight space is high at the start of the training and decreases throughout the training. At the end of the training, the weight space is clearly separated between the two ensemble members. The right plot shows the trajectories in weight space for two ensemble members, along with the validation loss. The plot underlines that the ensemble members are initialized randomly in nearby regions. Additionally, the plot shows that each ensemble member explores distinct modes in weight space that are characterized by lower loss values. This indicates that each ensemble member explores several optimal solutions, explaining the higher robustness under distribution shifts. This is underlined by [8]. More examples are given in the Appendix (see A.3). Although DEs provide a notable advantage over deterministic models by capturing multiple modes in weight space, their performance still degrades under distribution shift. To address this limitation, we explore the incorporation of additional functional forms into the model. In particular, we employ a 3D-LSTM architecture that captures spatial correlation within the image, a crucial aspect to capture in-field dynamics. The resulting weight space is illustrated in Fig. 10. As shown in the figure, including more expressive functional forms of the model anchors the solution within closer regions of the weight space. Consequently, generalization under distribution shift improves, as evidenced in Tab. 5, where the distributions exhibit greater overlap, similar to Fig. 8 (10-fold cv). Nonetheless, elongated clusters indicate residual uncertainty.

6. Conclusion & Open Challenges

This work introduced *YieldSAT*, a multimodal dataset for crop yield prediction at both the field and subfield levels. The dataset covers multiple countries, crop types, and years. We provided benchmark results across several model architectures and highlighted open challenges, including distribution shifts in the ground truth data. Although the dataset enables scalable yield prediction, it does not provide global coverage. For this, more data is required.

Several promising research directions remain open. First, most existing approaches operate at the pixel level due to significant variations in field size and the limited data. Modeling entire fields independent of their spatial extent remains an open challenge. Second, most models are designed for specific regions or crop types. Recently, Foundation Models (FMs) have emerged as a promising direction, offering the potential to handle multiple EO tasks and diverse data sources within a single unified model [10, 55]. Despite their promise, FMs are still in their early stages, especially for regression tasks [57]. Moreover, integrating uncertainty quantification, physical consistency, and explainability is often overlooked in the current literature but is highly required [33, 34, 60]. The current analysis is mainly performance-centric, leaving more detailed physical analyses to future work. Finally, although this study addressed the impact of distribution shifts arising in real-world settings, more attention must be paid to this topic. Addressing these challenges is fundamental to improving explainability, fostering trust, and ultimately enabling broader adoption of DL to support and advance digital farming.

Acknowledgments

This work was partly funded through the ESA InCubed Programme (https://incubed.esa.int/) as part of the project AI4EO Solution Factory (https://www.ai4eo-solution-factory.de/). M.M., H.N., and F.M. acknowledge support through a scholarship from the RPTU University of Kaiserslautern-Landau.

References

[1] Richard Barnes. *RichDEM: Terrain Analysis Software*, 2016. 4

[2] Juan Cao, Zhao Zhang, Yuchuan Luo, Liangliang Zhang, Jing Zhang, Ziyue Li, and Fulu Tao. Wheat yield predictions at a county and field scale with deep learning, machine learning, and google earth engine. *European Journal of Agronomy*, 123:126204, 2021. 2

[3] Christopher J Crawford, David P Roy, Saeed Arab, Christopher Barnes, Eric Vermote, Glynn Hulley, Aaron Gerace, Mike Choate, Christopher Engebretson, Esad Micijevic, et al. The 50-year landsat collection 2 archive. *Science of Remote Sensing*, 8:100103, 2023. 1

[4] Salvatore Cuomo, Vincenzo Schiano Di Cola, Fabio Giampaolo, Gianluigi Rozza, Maziar Raissi, and Francesco Piccialli. Scientific machine learning through physics–informed neural networks: Where we are and what's next. *Journal of Scientific Computing*, 92(3):88, 2022. 2

[5] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014. 2

[6] Burak Ekim, Girmaw Abebe Tadesse, Caleb Robinson, Gilles Hacheme, Michael Schmitt, Rahul Dodhia, and Juan M. Lavista Ferres. Distribution shifts at scale: Out-of-distribution detection in earth observation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2265–2274, June 2025. 6

[7] Tom G Farr and Mike Kobrick. Shuttle Radar Topography Mission produces a wealth of data. *Eos, Transactions American Geophysical Union*, 81(48):583–585, 2000. 4

[8] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. arxiv 2019. *arXiv preprint arXiv:1912.02757*, 2019. 8

[9] Keyhan Gavahi, Peyman Abbaszadeh, and Hamid Moradkhani. Deepyield: A combined convolutional neural network with long short-term memory for crop yield forecasting. *Expert Systems with Applications*, 184:115511, 2021. 2

[10] Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, et al. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27672–27683, 2024. 8

[11] Erhu He, Yiqun Xie, Licheng Liu, Weiye Chen, Zhenong Jin, and Xiaowei Jia. Physics guided neural networks for time-aware fairness: an application in crop yield prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14223–14231, 2023. 2

[12] Patrick Helber, Benjamin Bischke, Peter Habelitz, Cristhian Sanchez, Deepak Pathak, Miro Miranda, Hiba Najjar, Francisco Mena, Jayanth Siddamsetty, Diego Arenas, et al. Crop yield prediction: An operational approach to crop yield modeling on field and subfield level with machine learning models. In *IGARSS - IEEE International Geoscience and Remote Sensing Symposium*, pages 2763–2766. IEEE, 2023. 2, 5, 7

[13] Patrick Helber, Benjamin Bischke, Carolin Packbier, Peter Habelitz, and Florian Seefeldt. An operational approach to large-scale crop yield prediction with spatio-temporal machine learning models. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 4299–4302. IEEE, 2024. 2, 7

[14] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. 1, 4

[15] Adrian Höhl, Ivica Obadic, Miguel-Angel Fernandez-Torres, Hiba Najjar, Dario Augusto Borges Oliveira, Zeynep Akata, Andreas Dengel, and Xiao Xiang Zhu. Opening the black box: A systematic review on explainable artificial intelligence in remote sensing. *IEEE Geoscience and Remote Sensing Magazine*, 2024. 2

[16] S. Hoyer and J. Hamman. xarray: N-D labeled arrays and datasets in Python. *In revision, J. Open Res. Software*, 2017. 5

[17] Pavel Izmailov, Sharad Vikram, Matthew D Hoffman, and Andrew Gordon Gordon Wilson. What are bayesian neural network posteriors really like? In *International conference on machine learning*, pages 4629–4640. PMLR, 2021. 2

[18] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9492–9502, 2024. 2

[19] Martin Kopeckỳ, Martin Macek, and Jan Wild. Topographic wetness index calculation guidelines based on measured soil moisture and plant species composition. *Science of the Total Environment*, 757:143785, 2021. 4

[20] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017. 2, 7

[21] Corentin Leroux, Hazaël Jones, Anthony Clenet, Benoit Dreux, Maxime Becu, and Bruno Tisseyre. A general method to filter out defective spatial observations from yield mapping datasets. *Precision Agriculture*, 19:789–808, 2018. 3, 4

[22] Fudong Lin, Kaleb Guillot, Summer Crawford, Yihe Zhang, Xu Yuan, and Nian-Feng Tzeng. An open and large-scale dataset for multi-modal climate change-aware crop yield predictions. In *Proceedings of the 30th ACM SIGKDD Con-*

ference on Knowledge Discovery and Data Mining (KDD), pages 5375–5386, 2024. 2, 3

[23] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 5

[24] Francisco Mena, Diego Arenas, Marlon Nuske, and Andreas Dengel. Common practices and taxonomy in deep multiview fusion for remote sensing applications. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:4797–4818, 2024. 2, 6

[25] Francisco Mena, Deepak Pathak, Hiba Najjar, Cristhian Sanchez, Patrick Helber, Benjamin Bischke, Peter Habelitz, Miro Miranda, Jayanth Siddamsetty, Marlon Nuske, Marcela Charfuelan, Diego Arenas, Michaela Vollmer, and Andreas Dengel. Adaptive fusion of multi-modal remote sensing data for optimal sub-field crop yield prediction. *Remote Sensing of Environment*, 318:114547, 2025. 1, 2, 6, 7

[26] Met Office. *Cartopy: a cartographic python library with a matplotlib interface*. Exeter, Devon, 2010 - 2015. 3

[27] Nando Metzger, Mehmet Ozgur Turkoglu, Stefano D'Aronco, Jan Dirk Wegner, and Konrad Schindler. Crop classification under varying cloud cover with neural ordinary differential equations. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–12, 2021. 1

[28] Miro Miranda, Marcela Charfuelan, and Andreas Dengel. Exploring physics-informed neural networks for crop yield loss forecasting. *arXiv preprint arXiv:2501.00502*, 2024. 2

[29] Miro Miranda, Marcela Charfuelan, Matias Valdenegro-Toro, and Andreas Dengel. Informed learning for estimating drought stress at fine-scale resolution enables accurate yield prediction. In *ECAI 2025 – 27th European Conference on Artificial Intelligence*, pages 5384–5391. IOS Press, 2025. 1, 2

[30] Miro Miranda, Akshay Dinesh, David N. Lesmes-Leon, Fernando Mena, Mauricio Charfuelan, and Andreas Dengel. regdiff: Regression diffusion for earth observation. In *Proceedings of the IGARSS 2025 – IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2025. Accepted for publication. 2

[31] Miro Miranda, Francisco Mena, and Andreas Dengel. An analysis of temporal dropout in earth observation time series for regression tasks. In *Advances in Intelligent Data Analysis XXIII: 23rd International Symposium on Intelligent Data Analysis, IDA 2025, Konstanz, Germany, May 7–9, 2025, Proceedings*, volume 15669, page 389. Springer Nature, 2025. 4

[32] Miro Miranda, Deepak Pathak, Marlon Nuske, and Andreas Dengel. Multi-modal fusion methods with local neighborhood information for crop yield prediction at field and subfield levels. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 4307–4311. IEEE, 2024. 2, 6, 7

[33] Hiba Najjar, Miro Miranda, Marlon Nuske, Ribana Roscher, and Andreas Dengel. Explainability of sub-field level crop yield prediction using remote sensing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025. 2, 6, 8

[34] Hiba Najjar, Deepak Pathak, Marlon Nuske, and Andreas Dengel. Intrinsic explainability of multimodal learning for crop yield simulation. *Computers and Electronics in Agriculture*, 239:111003, 2025. 8

[35] Xanthoula Eirini Pantazi, Dimitrios Moshou, Thomas Alexandridis, Rebecca L Whetton, and Abdul Mounem Mouazen. Wheat yield prediction using machine learning and advanced sensing techniques. *Computers and electronics in agriculture*, 121:57–65, 2016. 2

[36] Deepak Pathak, Miro Miranda, Francisco Mena, Cristhian Sanchez, Patrick Helber, Benjamin Bischke, Peter Habelitz, Hiba Najjar, Jayanth Siddamsetty, Diego Arenas, Michaela Vollmer, Marcela Charfuelan, Marlon Nuske, and Andreas Dengel. Predicting Crop Yield with Machine Learning: An Extensive Analysis of Input Modalities and Models on a Field and Sub-Field Level. In *IGARSS- IEEE International Geoscience and Remote Sensing Symposium*, pages 2767–2770, 2023. 2, 5, 6, 7, 16

[37] Gregor Perich, Mehmet Ozgur Turkoglu, Lukas Valentin Graf, Jan Dirk Wegner, Helge Aasen, Achim Walter, and Frank Liebisch. Pixel-based yield mapping and prediction from Sentinel-2 using spectral indices and neural networks. *Field Crops Research*, 292:108824, 2023. 1, 2, 3

[38] Laura Poggio, Luis M De Sousa, Niels H Batjes, Gerard Heuvelink, Bas Kempen, Eloi Ribeiro, and David Rossiter. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. *Soil*, 7(1):217–240, 2021. 4

[39] Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics informed deep learning (part i): Data-driven solutions of nonlinear partial differential equations. *arXiv preprint arXiv:1711.10561*, 2017. 2

[40] Deepak K Ray, James S Gerber, Graham K MacDonald, and Paul C West. Climate variation explains a third of global crop yield variability. *Nature communications*, 6(1):5989, 2015. 2

[41] Hannah Ritchie, Pablo Rosado, and Max Roser. Agricultural production. *Our World in Data*, 2023. https://ourworldindata.org/agricultural-production. 3

[42] Ribana Roscher, Lukas Roth, Cyrill Stachniss, and Achim Walter. Data-centric digital agriculture: A perspective. *arXiv preprint arXiv:2312.03437*, 2023. 1

[43] Cristhian Sanchez, Deepak Pathak, Miro Miranda, Marcela Charfuelan, Patrick Helber, Marlon Nuske, Benjamin Bischke, Peter Habelitz, Nafisur Rahman, Francisco Mena, et al. Influence of data cleaning techniques on sub-field yield predictions. In *IGARSS- IEEE International Geoscience and Remote Sensing Symposium*, pages 4852–4855. IEEE, 2023. 3, 4

[44] Serco Gael consortium for ESA. Copernicus sentinel data access annual report, 2023. (Accessed: March 28, 2026). 1

[45] Mohsen Shahhosseini, Guiping Hu, Saeed Khaki, and Sotirios V Archontoulis. Corn yield prediction with ensemble cnn-dnn. *Frontiers in plant science*, 12:709008, 2021. 2

[46] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai kin Wong, and Wang chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting, 2015. 2

[47] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015. 6

[48] Pierre Soille, Armin Burger, D De Marchi, Pieter Kempeneers, D Rodriguez, Vassilis Syrris, and Veselin Vasilev. A versatile data-intensive computing platform for information retrieval from big geospatial data. *Future Generation Computer Systems*, 81:30–40, 2018. 1

[49] Amit Kumar Srivastava, Nima Safaei, Saeed Khaki, Gina Lopez, Wenzhi Zeng, Frank Ewert, Thomas Gaiser, and Jaber Rahimi. Winter wheat yield prediction using convolutional neural networks from environmental and phenological data. *Scientific reports*, 12(1):3215, 2022. 2

[50] Tanha Talaviya, Dhara Shah, Nivedita Patel, Hiteshri Yagnik, and Manan Shah. Implementation of artificial intelligence in agriculture for optimisation of irrigation and application of pesticides and herbicides. *Artificial intelligence in agriculture*, 4:58–73, 2020. 4

[51] Thomas Van Klompenburg, Ayalew Kassahun, and Cagatay Catal. Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, 177:105709, 2020. 2

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[53] Xinlei Wang, Jianxi Huang, Quanlong Feng, and Dongqin Yin. Winter wheat yield prediction at county level and uncertainty analysis in main wheat-producing regions of China with deep learning approaches. *Remote Sensing*, 12(11):1744, 2020. 2

[54] Andrew G Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *Advances in neural information processing systems*, 33:4697–4708, 2020. 2

[55] Aoran Xiao, Weihao Xuan, Junjue Wang, Jiaxing Huang, Dacheng Tao, Shijian Lu, and Naoto Yokoya. Foundation models for remote sensing and earth observation: A survey. *IEEE Geoscience and Remote Sensing Magazine*, 2025. 8

[56] Qiaoyun Xie, Jadu Dash, Alfredo Huete, Aihui Jiang, Gaofei Yin, Yanling Ding, Dailiang Peng, Christopher C Hall, Luke Brown, Yue Shi, et al. Retrieval of crop biophysical parameters from sentinel-2 remote sensing imagery. *International Journal of Applied Earth Observation and Geoinformation*, 80:187–195, 2019. 1

[57] Xizhe Xue and Xiao Xiang Zhu. Regression in earth observation: Are vlms up to the challenge? *Geoscience and Remote Sensing Magazine*, 2025. 1, 2, 8

[58] Jiaxuan You, Xiaocheng Li, Melvin Low, David Lobell, and Stefano Ermon. Deep gaussian process for crop yield prediction based on remote sensing data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017. 2

[59] Aurick Zhou and Sergey Levine. Bayesian adaptation for covariate shift. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 914–927. Curran Associates, Inc., 2021. 6

[60] Xiao Xiang Zhu, Zhitong Xiong, Yi Wang, Adam J Stewart, Konrad Heidler, Yuanyuan Wang, Zhenghang Yuan, Thomas Dujardin, Qingsong Xu, and Yilei Shi. On the foundations of earth and climate foundation models. *arXiv preprint arXiv:2405.04285*, 2024. 1, 8