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

A.1. Crop Values for Data Preprocessing

The maximum accepted yield values for each crop type are given in Tab. 6. Yield values above this threshold were removed. Additionally, the standard moisture is given in Tab. 6 for each crop type that is used to calculate the scaled yield as shown in the main paper.

A.2. Data Analysis

A.2.1. Data Quality and Support Size

The rasterization workflow for a single field is shown in Fig. 11. Combine harvester in vector format must be processed to match high-resolution EO inputs (e.g., S2). For this, every yield measurement that falls into a grid cell of $10\,\mathrm{m} \times 10\,\mathrm{m}$ are averaged. Due to harvester path density, swath width, speed, logging frequency, and positional delay, different numbers of yield measurements can fall into a single grid cell, resulting in variable support sizes for every pixel. Those support sizes are often correlated in space. An example of different support sizes for every yield map quality level, shown in Fig. 5, is illustrated in Fig. 12. The figure shows the number of points that fall into a grid cell and the corresponding standard deviation per pixel. It is evident that the yield maps have different support sizes and variability per pixel. In Fig. 13, we show that these characteristics depend on the crop type and dataset. For instance, Brazil has consistently more yield points per pixel, while Uruguay has fewer samples. This is also reflected in a lower standard deviation per pixel. For each yield map, we also provide the raster grid for the variable support size.

Each field was manually inspected and curated. For this, each field is assigned a quality score of either "good", "average", or "bad", following a stringent guideline:

1. Good yield maps are characterized by high data quality and consistency. They do not exhibit strong visible striping artifacts (small can be present). The data coverage is dense, with no significant gaps or sparsity. Additionally, good yield maps show meaningful variability across the field, reflecting realistic yield differences, while maintaining smooth spatial transitions without abrupt or isolated changes.

2. Bad yield maps display clear quality issues. These may include visible striping patterns, often caused by sensor or machinery inconsistencies, as well as sparse or incomplete data coverage. Such maps typically lack meaningful vari-

Table 6. Maximum accepted yield values in t/ha for every crop type and the standard moisture.

Standard Values	Wheat	Rapeseed	Soybean	Corn
Max. Yield (t/ha)	20	10	15	45
Standard Moisture (%)	15	9	15	16

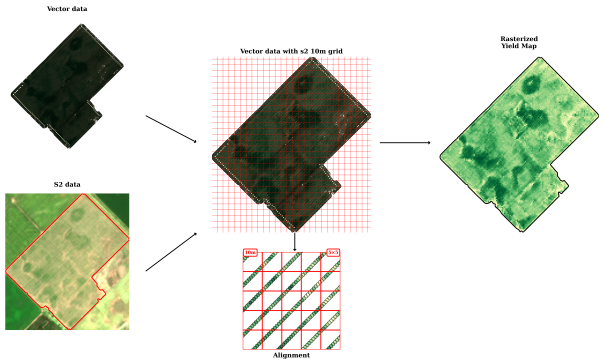


Figure 11. Schematic overview of data collection and preprocessing: A combine harvester collects point vector data of the yield, which is preprocessed and rasterized to align with S2 $10\,\mathrm{m}$ grid for pixel-wise regression.

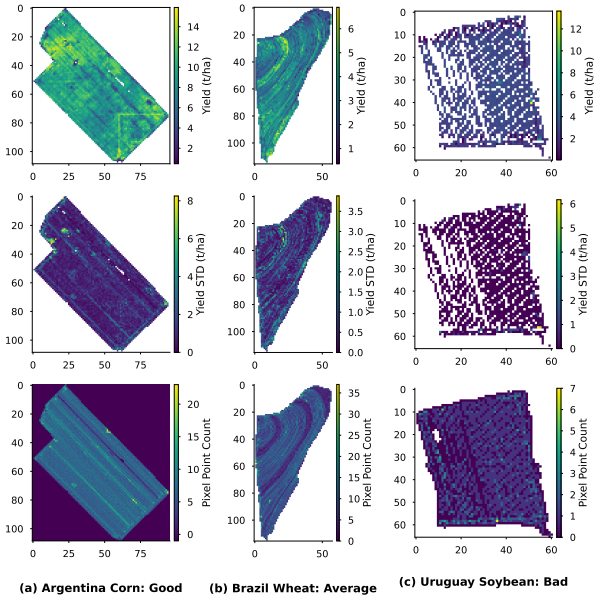


Figure 12. Label support size and standard deviation per $10\,\mathrm{m}$ pixel caused by the rasterization. Top row: sample count, bottom row: standard deviation.

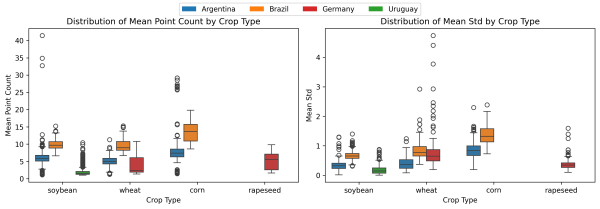


Figure 13. Boxplots for pixel point count (left) and mean standard deviation (right) per crop and country.

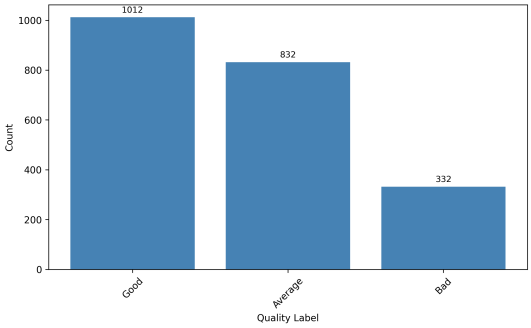


Figure 14. Overall distribution of yield map quality over the entire dataset.

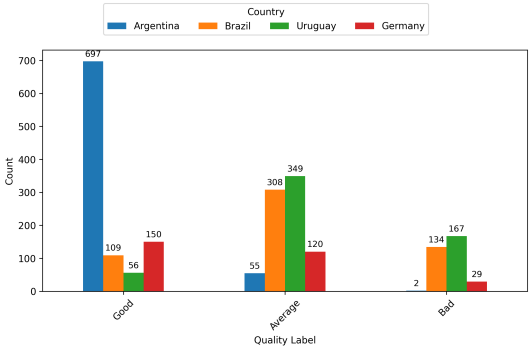


Figure 15. Overall distribution of yield map quality over the entire dataset grouped by country.

ability, with large areas showing nearly constant values. They may also contain isolated pixels or small regions with sudden, unrealistic changes in yield values that are inconsistent with the surrounding data.

3. Average yield maps fall between these two extremes. They may exhibit minor striping, moderate sparsity, or limited variability, but not to the extent observed in bad maps. While generally usable, they do not meet the quality standards of good yield maps and may require additional processing or careful interpretation.

An example of each class is shown in the main paper in Fig. 5. The distribution of yield map quality across the entire dataset is shown in Fig. 14. The plot indicates that the majority of yield maps are of good quality, followed by average and bad categories, based on a total of 332 fields. Fig. 15 presents the distribution of yield map quality grouped by country. Argentina exhibits the highest number of good-quality yield maps, whereas Uruguay has the largest number of low-quality (bad) yield maps. Finally, Fig. 16 shows the distribution of yield map quality by crop type. Soybean accounts for the largest number of high-quality yield maps, followed by wheat and corn.

The quality labels are stored in the metadata of each field.

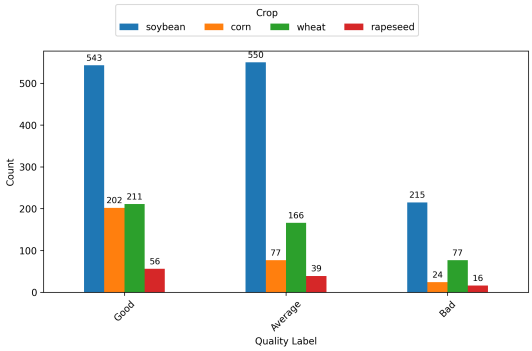


Figure 16. Overall distribution of yield map quality over the entire dataset grouped by country.

A.2.2. Distribution Shifts: Countries and Crop Types

To test whether the distributions of crop types and countries are significantly different, significance testing is performed to compare the distributions between groups. Both tests demonstrate significantly different distributions between countries and crop types (p-values < 0.0001) as shown in Tab. 7. Following, we perform a post-hoc analysis that compares pairs within each group using Dunn's significance test using a Holm-Bonferroni correction. The results for the pairwise comparison between each country are shown in Tab. 8. Notably, each group has a distinct yield distribution, with most p-values < 0.0001. Except for between Germany and Brazil, no significance is present. Additionally, the Dunn's test for the pairwise comparison between crop types is shown in Tab. 9. Similarly, most distributions differ significantly between crop types, with all p-values < 0.0001, except for rapeseed and soybean.

A.2.3. Distribution Shifts: Regions and Years

Importantly, we also find significant differences in data distribution and surface reflectance across regions and years within a single country and crop type. In Fig. 17 t-SNE plot of the surface reflectance for only soybean in Argentina is shown. The data are colored by year (left) and region (right). We highlight that, in both cases, a separation between individual years and regions is evident, as shown by clusters within each group. Below is the kernel density estimation plot of the target yield distribution. The data are grouped by year (left) and region (right). Note that each group's distribution is unique. Individual distributions do not follow a normal distribution. Moreover, the plot shows that the yield distribution differs between years and regions. The results indicate unique patterns within each year-region combination across countries and crop types.

To test the hypothesis that all groups have the same yield distribution, a Kruskal-Wallis test is performed to compare the distributions across groups. The results are depicted in Tab. 10. For both tests, a significantly high

Table 7. Kruskal–Wallis H-test between yield distributions grouped by countries and by crop type. $**** = p < 0.0001$.

Evaluation	Country	Crop
p-value	****	****

Table 8. Pairwise post-hoc comparisons of the yield distributions between individual **countries**. Each cell displays the statistical significance level of the difference between two countries based on the Dunn's test using the Holm–Bonferroni correction. ns = no significance ($p \geq 0.05$), $* = p < 0.05$, $** = p < 0.01$, $*** = p < 0.001$, $**** = p < 0.0001$.

Comparision	Argentina	Brazil	Germany	Uruguay
Argentina	-	****	****	****
Brazil	****	-	ns	****
Germany	****	ns	-	****
Uruguay	****	****	****	-

Table 9. Pairwise post-hoc comparisons of the yield distributions between individual **crops**. Each cell displays the statistical significance level of the difference between two crops based on the Dunn's test using the Holm–Bonferroni correction. ns = no significance ($p \geq 0.05$), $* = p < 0.05$, $** = p < 0.01$, $*** = p < 0.001$, $**** = p < 0.0001$.

Comparision	Corn	Rapeseed	Soybean	Wheat
Corn	-	****	****	****
Rapeseed	****	-	ns	****
Soybean	****	ns	-	****
Wheat	****	****	****	-

Table 10. Kruskal–Wallis H-test between yield distributions for soybean in Argentina, grouped by years and by regions. $**** = p < 0.0001$.

Evaluation	Year	Region
p-value	****	****

Table 11. Pairwise post-hoc comparisons of the yield distributions for soybean in Argentina between individual **years**. Each cell displays the statistical significance level of the difference between two years based on Dunn's test using the Holm–Bonferroni correction. ns = no significance ($p \geq 0.05$), $* = p < 0.05$, $** = p < 0.01$, $*** = p < 0.001$, $**** = p < 0.0001$.

Year	2017	2018	2019	2020	2021	2022
2017	-	****	***	****	****	****
2018	****	-	ns	***	****	****
2019	***	ns	-	**	****	****
2020	****	***	**	-	****	****
2021	****	****	****	****	-	****
2022	****	****	****	****	****	-

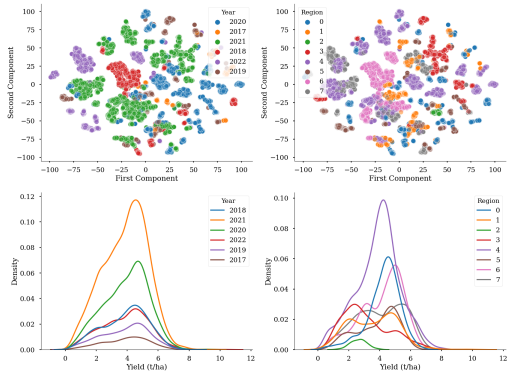


Figure 17. Visualization of the surface reflectance and yield data distribution only for soybean in Argentina, grouped by years and regions. Top left: t-SNE plot of the S2 surface reflectance data grouped by **years**. Top right: t-SNE plot of the S2 surface reflectance data grouped by **region**. Bottom left: Kernel density estimation plot for the target yield data grouped by **years**. Bottom right: Kernel density estimation plot for the target yield data grouped by **regions**.

Table 12. Pairwise post-hoc comparisons of the yield distributions for soybean in Argentina between individual **regions**. Each cell displays the statistical significance level of the difference between two regions based on Dunn's test using the Holm–Bonferroni correction. ns = no significance ($p \geq 0.05$), $* = p < 0.05$, $** = p < 0.01$, $*** = p < 0.001$, $**** = p < 0.0001$.

Region	0	1	2	3	4	5	6	7
0	-	****	****	****	****	****	**	****
1	****	-	****	****	****	****	****	****
2	****	****	-	***	****	****	****	****
3	****	****	***	-	****	****	****	****
4	****	****	****	****	-	ns	****	ns
5	****	****	****	****	ns	-	ns	ns
6	**	****	****	****	****	ns	-	***
7	****	****	****	****	ns	ns	***	-

p-value is reported ($p < 0.0001$). This means that at least one year and one region have a significantly different yield distribution compared to the rest of the groups. To compare each pair of yield distributions, a post hoc test is performed using Dunn's test. The pairwise comparison between the years is depicted in Tab. 11. We highlight that most pairwise comparisons indicate significantly different distributions. Only between 2018 and 2019 is there no significance. The pairwise comparison between the regions is depicted in Tab. 12. As in previous years, the data distribution across regions is mostly significantly different. However, individual regions do not show a significantly different distribution. For instance, region 5 shows no significant difference between regions 4, 6, and 7.

In conclusion, the results undermine the assumption that the distribution across years and regions is mostly significantly different, which increases the difficulty of generalization.

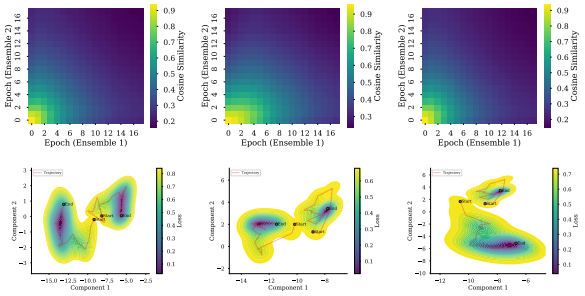


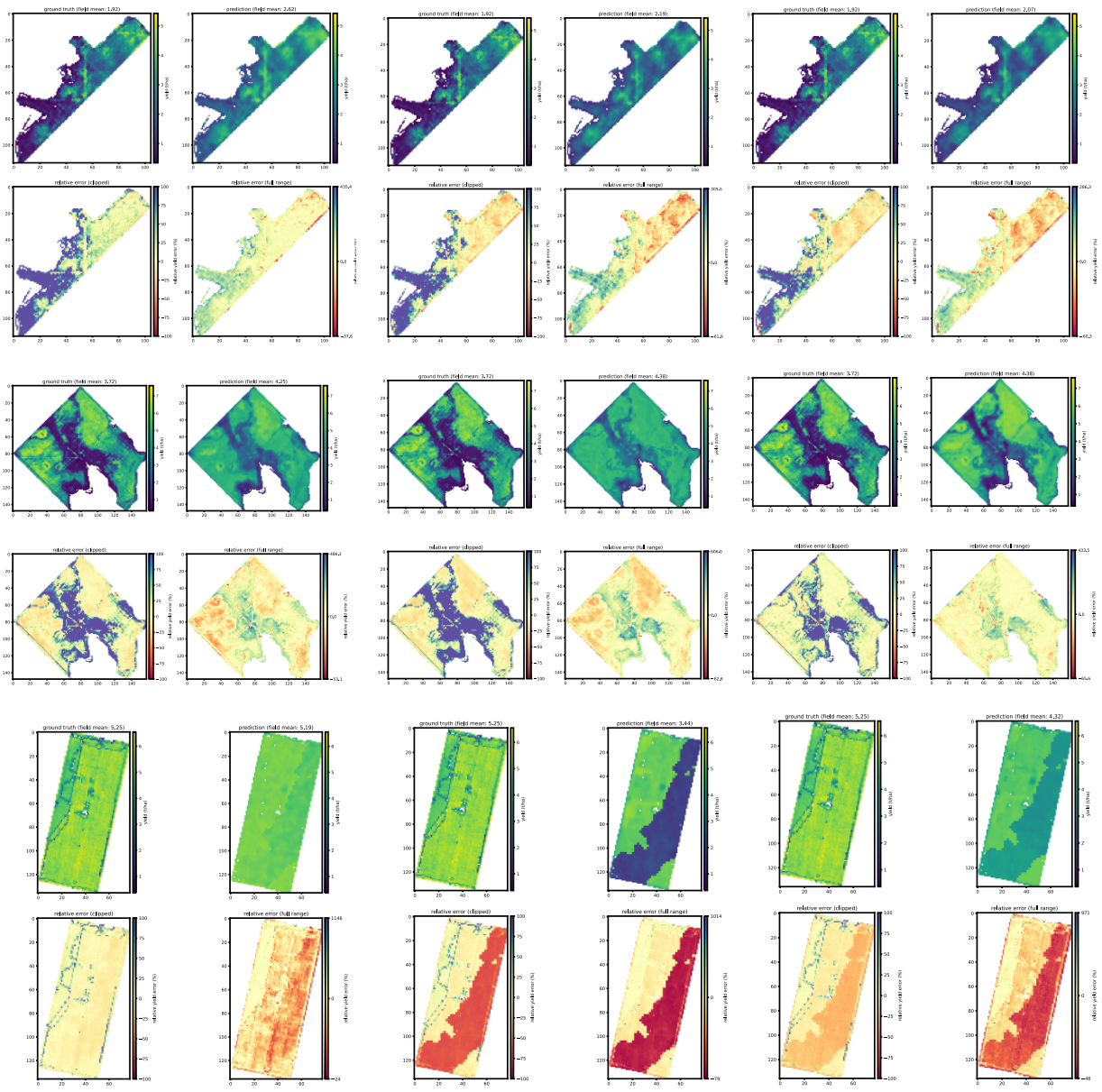
Figure 18. Visualization of the weight space diversity during model training. Top: Cosine similarity between pairs of ensemble members during training. Bottom: PCA plot of the weight space during training, together with the loss. The trajectory in the weight space is highlighted in red from start to end.

A.3. Further Experiments

Fig. 19 shows additional example predictions for single fields (image), illustrating the ground-truth yield map, the predicted yield map, and the pixel-wise error (clipped and full range). We show the results for the standard 10-fold cv experiment and the temporal (LOYO) and spatial (LORO) experiments. Note that for some fields, the LOYO and LORO experiments exhibit severe performance collapse, as evidenced by high pixel-wise errors and areas of mode collapse (only single-scalar predictions with no spatial variability).

In Tab. 13 we provide a full replication for the 10-fold cv baseline experiment of all benchmark models for each dataset at the field level and at the subfield (pixel)-level in Tab. 13. We further provide a full replication of all benchmark models on the OOD experiments, namely the temporal (LOYO) and spatial (LORO) splitting. For the temporal splitting, the field level results are given in Tab. 15 and for the subfield (pixel)-level in Tab. 16. For the spatial splitting (LORO), the field level results are given in Tab. 17 and for the subfield (pixel)-level in Tab. 18.

Fig. 18 provides more examples of the cosine similarity in weight space over the training epochs, together with the trajectory in weight space.



(a) Standard 10-Fold CV (b) LOYO CV (c) LOFO CV

Figure 19. Example ground truth and predicted yield maps together with the pixel-wise error for the standard CV, the LORO, and LOYO CV under distribution shift. The results are generated with the LSTM model [36], trained on S2 data.

Table 13. Results for the RMSE (t/ha) (↓) and the R^2-score (↑) for different models and datasets for the standard 10-fold cross-validation at the field level. ARG = Argentina, BRA = Brazil, GER = Germany, URG = Uruguay. C = corn, R = rapeseed, S = soybean, W = wheat.

Evaluation				Field-Level																	
				ARG-C		ARG-S		ARG-W		BRA-C		BRA-S		BRA-W		GER-R		GER-W		URG-S	
Splitting	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
CV10	Sentinel-2	X	3D-ConvLSTM	0.84 ± 0.07	1.13 ± 0.22	0.79 ± 0.12	0.55 ± 0.16	0.92 ± 0.08	0.62 ± 0.24	0.82 ± 0.09	0.74 ± 0.13	0.76 ± 0.09	0.34 ± 0.05	0.73 ± 0.09	0.42 ± 0.16	0.81 ± 0.10	0.58 ± 0.19	0.65 ± 0.16	1.12 ± 0.34	0.77 ± 0.06	0.51 ± 0.10
		X	3D-LSTM	0.74 ± 0.15	1.44 ± 0.39	0.77 ± 0.11	0.58 ± 0.15	0.89 ± 0.10	0.70 ± 0.25	0.82 ± 0.11	0.74 ± 0.11	0.76 ± 0.08	0.34 ± 0.06	0.63 ± 0.16	0.48 ± 0.21	0.82 ± 0.09	0.57 ± 0.14	0.54 ± 0.26	1.28 ± 0.40	0.73 ± 0.09	0.56 ± 0.12
		X	LSTM	0.79 ± 0.11	1.29 ± 0.24	0.72 ± 0.12	0.64 ± 0.14	0.85 ± 0.10	0.84 ± 0.23	0.75 ± 0.2	0.88 ± 0.22	0.68 ± 0.12	0.39 ± 0.05	0.62 ± 0.16	0.49 ± 0.18	0.62 ± 0.25	0.83 ± 0.26	0.55 ± 0.41	2.60 ± 1.28	0.66 ± 0.11	0.62 ± 0.12
		X	Transformer	0.77 ± 0.07	1.35 ± 0.21	0.73 ± 0.11	0.63 ± 0.16	0.87 ± 0.11	0.78 ± 0.23	0.79 ± 0.19	0.82 ± 0.19	0.75 ± 0.12	0.35 ± 0.08	0.62 ± 0.12	0.49 ± 0.19	0.75 ± 0.14	0.67 ± 0.21	0.56 ± 0.33	1.26 ± 0.24	0.72 ± 0.10	0.56 ± 0.10
	Sentinel-2 + ADM	Feature Fusion	AFF	0.84 ± 0.05	1.12 ± 0.18	0.84 ± 0.09	0.49 ± 0.14	0.92 ± 0.08	0.6 ± 0.22	0.84 ± 0.04	0.70 ± 0.15	0.8 ± 0.08	0.31 ± 0.07	0.71 ± 0.12	0.43 ± 0.14	0.80 ± 0.11	0.60 ± 0.13	0.74 ± 0.17	0.96 ± 0.37	0.81 ± 0.09	0.46 ± 0.11
			MMGF	0.79 ± 0.13	1.30 ± 0.34	0.82 ± 0.09	0.51 ± 0.14	0.89 ± 0.09	0.69 ± 0.25	0.76 ± 0.15	0.86 ± 0.21	0.80 ± 0.06	0.31 ± 0.05	0.59 ± 0.15	0.51 ± 0.17	0.75 ± 0.14	0.68 ± 0.26	0.77 ± 0.16	0.90 ± 0.21	0.75 ± 0.08	0.53 ± 0.09
		Input Fusion	3D-ConvLSTM	0.12 ± 0.26	2.63 ± 0.44	0.82 ± 0.09	0.52 ± 0.14	0.89 ± 0.1	0.72 ± 0.25	0.83 ± 0.13	0.72 ± 0.17	0.76 ± 0.1	0.34 ± 0.06	0.72 ± 0.10	0.42 ± 0.17	0.78 ± 0.09	0.63 ± 0.18	0.70 ± 0.20	1.03 ± 0.31	0.78 ± 0.10	0.50 ± 0.10
			3D-LSTM	0.40 ± 0.50	2.17 ± 0.79	0.76 ± 0.10	0.59 ± 0.14	0.87 ± 0.11	0.77 ± 0.27	0.84 ± 0.1	0.71 ± 0.13	0.75 ± 0.1	0.35 ± 0.06	0.72 ± 0.09	0.42 ± 0.14	0.81 ± 0.07	0.59 ± 0.12	0.62 ± 0.31	1.17 ± 0.27	0.77 ± 0.07	0.52 ± 0.09
			LSTM	0.76 ± 0.12	1.39 ± 0.29	0.72 ± 0.11	0.64 ± 0.13	0.85 ± 0.10	0.83 ± 0.24	0.81 ± 0.20	0.78 ± 0.2	0.64 ± 0.15	0.42 ± 0.07	0.58 ± 0.21	0.52 ± 0.22	0.81 ± 0.09	0.58 ± 0.15	0.63 ± 0.32	1.16 ± 0.35	0.72 ± 0.06	0.56 ± 0.07
			Transformer	0.69 ± 0.18	1.55 ± 0.34	0.72 ± 0.12	0.64 ± 0.15	0.83 ± 0.11	0.87 ± 0.22	0.79 ± 0.16	0.81 ± 0.18	0.71 ± 0.09	0.37 ± 0.06	0.61 ± 0.21	0.49 ± 0.19	0.76 ± 0.11	0.65 ± 0.15	0.61 ± 0.24	1.19 ± 0.22	0.73 ± 0.07	0.55 ± 0.10

Table 14. Results for the RMSE (t/ha) (↓) and the R^2-score (↑) for different models and datasets for the standard 10-fold cross-validation at the subfield (pixel)-level. ARG = Argentina, URG = Uruguay, GER = Germany, BRA = Brazil. S = soybean, R = rapeseed, W = wheat, C = Corn.

Evaluation				Subfield (Pixel)-Level																	
				ARG-C		ARG-S		ARG-W		BRA-C		BRA-S		BRA-W		GER-R		GER-W		URG-S	
Splitting	Modalities	Fusion Method	Model	R2	RMSE	R2	RMSE	R2	RMSE	R2	RMSE	R2	RMSE	R2	RMSE	R2	RMSE	R2	RMSE	R2	RMSE
CV10	Sentinel-2	X	3D-ConvLSTM	0.65 ±0.11	2.2 ±0.27	0.65 ±0.06	0.9 ±0.11	0.79 ±0.1	1.1 ±0.18	0.45 ±0.12	2.13 ±0.31	0.39 ±0.07	0.94 ±0.06	0.23 ±0.05	1.38 ±0.18	0.49 ±0.1	1.2 ±0.22	0.34 ±0.09	2.4 ±0.34	0.41 ±0.06	1.22 ±0.12
		X	3D-LSTM	0.59 ±0.13	2.37 ±0.38	0.65 ±0.07	0.9 ±0.11	0.79 ±0.11	1.08 ±0.2	0.46 ±0.14	2.13 ±0.37	0.39 ±0.07	0.93 ±0.06	0.24 ±0.05	1.37 ±0.19	0.48 ±0.12	1.2 ±0.23	0.3 ±0.2	2.46 ±0.43	0.39 ±0.07	1.23 ±0.12
		X	LSTM	0.59 ±0.07	2.4 ±0.26	0.6 ±0.08	0.96 ±0.1	0.74 ±0.12	1.21 ±0.19	0.42 ±0.13	2.2 ±0.35	0.34 ±0.09	0.98 ±0.06	0.22 ±0.06	1.39 ±0.17	0.36 ±0.14	1.33 ±0.23	-0.32 ±0.64	3.38 ±1.04	0.37 ±0.07	1.26 ±0.12
		X	Transformer	0.61 ±0.05	2.33 ±0.23	0.62 ±0.08	0.94 ±0.1	0.76 ±0.12	1.17 ±0.22	0.44 ±0.13	2.16 ±0.31	0.38 ±0.08	0.94 ±0.07	0.22 ±0.06	1.39 ±0.17	0.44 ±0.1	1.25 ±0.24	0.32 ±0.34	2.43 ±0.27	0.38 ±0.08	1.24 ±0.13
	Sentinel-2 + ADM	Feature Fusion	AFF	0.7 ±0.06	2.03 ±0.27	0.73 ±0.06	0.79 ±0.1	0.84 ±0.09	0.96 ±0.17	0.46 ±0.11	2.12 ±0.33	0.44 ±0.08	0.9 ±0.07	0.24 ±0.06	1.37 ±0.18	0.49 ±0.09	1.20 ±0.22	0.44 ±0.16	2.2 ±0.44	0.43 ±0.07	1.19 ±0.10
			MMGF	0.65 ±0.08	2.2 ±0.33	0.7 ±0.06	0.84 ±0.1	0.8 ±0.09	1.07 ±0.19	0.42 ±0.14	2.19 ±0.42	0.42 ±0.07	0.92 ±0.06	0.2 ±0.08	1.4 ±0.19	0.44 ±0.16	1.26 ±0.23	0.44 ±0.1	2.21 ±0.36	0.4 ±0.07	1.22 ±0.12
			3D-ConvLSTM	0.08 ±0.24	3.58 ±0.64	0.68 ±0.06	0.87 ±0.1	0.78 ±0.1	1.11 ±0.18	0.46 ±0.12	2.12 ±0.31	0.38 ±0.07	0.94 ±0.06	0.23 ±0.06	1.38 ±0.18	0.42 ±0.14	1.27 ±0.24	0.39 ±0.16	2.3 ±0.4	0.41 ±0.06	1.22 ±0.12
		Input Fusion	3D-LSTM	0.36 ±0.29	2.97 ±0.63	0.64 ±0.07	0.91 ±0.11	0.78 ±0.1	1.13 ±0.18	0.46 ±0.12	2.12 ±0.34	0.41 ±0.06	0.93 ±0.06	0.24 ±0.06	1.37 ±0.17	0.49 ±0.1	1.2 ±0.2	0.37 ±0.13	2.34 ±0.35	0.4 ±0.07	1.22 ±0.12
			LSTM	0.58 ±0.09	2.41 ±0.23	0.59 ±0.08	0.98 ±0.1	0.75 ±0.11	1.2 ±0.2	0.43 ±0.14	2.18 ±0.37	0.33 ±0.09	0.99 ±0.07	0.21 ±0.06	1.4 ±0.18	0.47 ±0.1	1.21 ±0.2	0.35 ±0.19	2.38 ±0.4	0.39 ±0.07	1.23 ±0.12
			Transformer	0.52 ±0.13	2.57 ±0.2	0.58 ±0.07	0.98 ±0.1	0.7 ±0.15	1.3 ±0.24	0.44 ±0.14	2.17 ±0.33	0.37 ±0.09	0.96 ±0.07	0.22 ±0.06	1.39 ±0.17	0.45 ±0.1	1.24 ±0.2	0.38 ±0.13	2.32 ±0.34	0.39 ±0.07	1.23 ±0.13

Table 15. Results for the RMSE (t/ha) (↓) and the R^2-score (↑) for different models and datasets for the Leave-One-Year-Out cross-validation at the field level. ARG = Argentina, BRA = Brazil, GER = Germany, URG = Uruguay. C = corn, R = rapeseed, S = soybean, W = wheat.

Evaluation				Field-Level																	
				ARG-C		ARG-S		ARG-W		BRA-C		BRA-S		BRA-W		GER-R		GER-W		URG-S	
Splitting	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
LOYO	Sentinel-2	X	3D-ConvLSTM	0.61 ±0.55	1.75 ±0.51	0.65 ±0.11	0.71 ±0.18	0.78 ±0.88	1.00 ±0.60	0.20 ±0.12	1.58 ±0.75	0.36 ±0.42	0.55 ±0.15	-0.06 ±1.00	0.82 ±0.42	0.43 ±0.24	1.02 ±0.23	0.23 ±0.60	1.67 ±0.60	0.63 ±0.25	0.65 ±0.15
		X	3D-LSTM	0.53 ±0.49	1.93 ±0.71	0.66 ±0.24	0.70 ±0.21	0.77 ±1.94	1.02 ±0.77	0.36 ±0.66	1.42 ±0.42	0.12 ±0.86	0.65 ±0.22	-0.33 ±1.02	0.92 ±0.56	0.41 ±0.23	1.03 ±0.24	0.00 ±0.74	1.90 ±0.72	0.62 ±0.30	0.66 ±0.17
		X	LSTM	0.46 ±0.59	2.06 ±0.54	0.50 ±0.22	0.85 ±0.25	0.77 ±1.70	1.02 ±0.30	0.19 ±1.50	1.59 ±0.48	0.06 ±1.66	0.67 ±0.21	-0.05 ±1.61	0.82 ±0.40	-0.07 ±0.77	1.39 ±0.86	-1.74 ±2.65	3.14 ±1.32	0.61 ±0.23	0.67 ±0.12
		X	Transformer	0.57 ±0.64	1.83 ±0.33	0.52 ±0.21	0.84 ±0.20	0.78 ±0.94	0.99 ±0.51	0.19 ±0.61	1.59 ±0.59	0.28 ±0.64	0.59 ±0.16	-0.09 ±1.04	0.83 ±0.42	0.33 ±0.44	1.10 ±0.30	0.14 ±0.91	1.76 ±0.50	0.60 ±0.33	0.68 ±0.16
	Sentinel-2 + ADM	Feature Fusion	AFF	0.33 ±10.65	2.30 ±1.47	0.77 ±0.64	0.59 ±0.23	0.81 ±0.42	0.93 ±0.25	0.42 ±0.19	1.35 ±0.59	0.38 ±0.70	0.54 ±0.16	0.14 ±0.51	0.74 ±0.43	0.44 ±0.34	1.00 ±0.28	0.19 ±0.49	1.71 ±0.64	0.53 ±0.47	0.73 ±0.22
			MMGF	0.31 ±4.38	2.33 ±1.24	0.66 ±0.77	0.70 ±0.67	0.36 ±0.54	1.71 ±0.85	0.39 ±1.04	1.38 ±0.44	0.38 ±0.40	0.55 ±0.15	-0.08 ±1.06	0.83 ±0.38	0.15 ±0.47	1.24 ±0.43	-0.03 ±0.75	1.92 ±0.87	0.48 ±0.96	0.77 ±0.12
			3D-ConvLSTM	-0.01 ±1.09	2.82 ±1.01	0.67 ±0.21	0.69 ±0.16	0.77 ±1.48	1.03 ±0.42	0.45 ±0.63	1.32 ±0.55	0.34 ±0.71	0.56 ±0.12	-0.15 ±0.78	0.86 ±0.53	0.49 ±0.28	0.96 ±0.27	0.46 ±0.29	1.39 ±0.35	0.56 ±0.28	0.71 ±0.15
		Input Fusion	3D-LSTM	0.18 ±0.63	2.54 ±0.81	0.62 ±0.24	0.74 ±0.14	0.78 ±0.64	0.99 ±0.35	0.28 ±0.36	1.50 ±0.62	0.45 ±0.47	0.52 ±0.11	-0.07 ±0.67	0.82 ±0.49	0.26 ±0.48	1.15 ±0.39	0.15 ±0.68	1.75 ±0.63	0.52 ±0.33	0.74 ±0.19
			LSTM	0.42 ±0.60	2.13 ±0.60	0.62 ±0.31	0.74 ±0.16	0.74 ±1.67	1.08 ±0.14	0.29 ±1.99	1.49 ±0.41	0.26 ±1.10	0.59 ±0.13	0.12 ±1.21	0.75 ±0.32	0.58 ±0.15	0.87 ±0.27	0.33 ±0.55	1.56 ±0.38	0.54 ±0.35	0.73 ±0.12
			Transformer	0.33 ±0.88	2.30 ±0.98	0.65 ±0.21	0.72 ±0.15	0.75 ±0.85	1.08 ±0.3	0.25 ±1.07	1.54 ±0.43	-0.09 ±1.31	0.72 ±0.30	0.11 ±1.47	0.75 ±0.33	0.47 ±0.11	0.98 ±0.21	0.02 ±1.48	1.88 ±0.47	0.56 ±0.33	0.71 ±0.16

Table 16. Results for the RMSE (t/ha) (↓) and the R^2-score (↑) for different models and datasets for the Leave-One-Region-Out (LORO) cross-validation at the field level.ARG = Argentina, BRA = Brazil, GER = Germany, URG = Uruguay. C = corn, R = rapeseed, S = soybean, W = wheat.

Evaluation				Subfield (Pixel)-Level																	
				ARG-C		ARG-S		ARG-W		BRA-C		BRA-S		BRA-W		GER-R		GER-W		URG-S	
Splitting	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
LOYO	Sentinel-2	X	3D-ConvLSTM	0.48 ±0.29	2.68 ±0.13	0.54 ±0.14	1.04 ±0.18	0.67 ±0.18	1.38 ±0.32	0.15 ±0.11	2.67 ±0.69	0.19 ±0.11	1.08 ±0.17	0.06 ±0.19	1.52 ±0.45	0.21 ±0.15	1.49 ±0.33	0.15 ±0.27	2.71 ±0.79	0.34 ±0.09	1.28 ±0.13
		X	3D-LSTM	0.47 ±0.25	2.70 ±0.39	0.55 ±0.14	1.03 ±0.18	0.68 ±0.5	1.35 ±0.32	0.21 ±0.19	2.56 ±0.73	0.13 ±0.21	1.12 ±0.18	0.03 ±0.24	1.54 ±0.53	0.18 ±0.26	1.52 ±0.40	0.01 ±0.3	2.93 ±0.91	0.35 ±0.11	1.27 ±0.14
		X	LSTM	0.39 ±0.21	2.90 ±0.27	0.46 ±0.16	1.12 ±0.17	0.61 ±0.73	1.49 ±0.46	0.15 ±0.25	2.66 ±0.66	0.04 ±0.44	1.18 ±0.22	0.08 ±0.17	1.50 ±0.46	-0.05 ±0.35	1.71 ±0.28	-0.68 ±0.77	3.82 ±1.21	0.33 ±0.08	1.29 ±0.1
		X	Transformer	0.42 ±0.35	2.83 ±0.29	0.48 ±0.14	1.10 ±0.16	0.67 ±0.62	1.38 ±0.41	0.19 ±0.16	2.60 ±0.8	0.16 ±0.20	1.10 ±0.14	0.05 ±0.19	1.53 ±0.45	0.14 ±0.19	1.55 ±0.32	0.13 ±0.30	2.74 ±0.70	0.33 ±0.11	1.29 ±0.12
	Sentinel-2 + ADM	Feature Fusion	AFF	0.35 ±1.83	2.99 ±1.31	0.66 ±0.36	0.89 ±0.19	0.72 ±0.31	1.27 ±0.24	0.29 ±0.89	2.43 ±0.81	0.29 ±0.18	1.02 ±0.17	0.11 ±0.15	1.48 ±0.48	0.25 ±0.14	1.45 ±0.36	0.16 ±0.22	2.70 ±0.82	0.32 ±0.16	1.30 ±0.16
			MMGF	0.30 ±0.97	3.10 ±1.13	0.57 ±1.32	1.00 ±0.48	0.28 ±2.97	2.03 ±0.76	0.23 ±0.25	2.53 ±0.67	0.24 ±0.16	1.05 ±0.15	0.06 ±0.16	1.52 ±0.41	0.05 ±0.28	1.63 ±0.43	0.06 ±0.30	2.86 ±0.97	0.28 ±0.12	1.34 ±0.13
			3D-ConvLSTM	-0.02 ±0.47	3.75 ±0.79	0.56 ±0.16	1.01 ±0.13	0.64 ±0.75	1.43 ±0.36	0.27 ±0.15	2.47 ±0.81	0.19 ±0.19	1.08 ±0.16	0.04 ±0.20	1.54 ±0.52	0.22 ±0.17	1.47 ±0.36	0.22 ±0.17	2.60 ±0.63	0.31 ±0.11	1.31 ±0.16
		Input Fusion	3D-LSTM	0.25 ±0.40	3.23 ±0.68	0.53 ±0.14	1.05 ±0.14	0.68 ±0.47	1.34 ±0.34	0.22 ±0.11	2.55 ±0.82	0.22 ±0.17	1.06 ±0.19	0.06 ±0.21	1.52 ±0.51	0.13 ±0.40	1.56 ±0.51	0.15 ±0.28	2.72 ±0.79	0.30 ±0.13	1.32 ±0.17
			LSTM	0.44 ±0.26	2.78 ±0.24	0.49 ±0.19	1.09 ±0.15	0.61 ±0.67	1.49 ±0.37	0.19 ±0.33	2.60 ±0.67	0.17 ±0.28	1.09 ±0.17	0.09 ±0.17	1.50 ±0.43	0.31 ±0.10	1.38 ±0.27	0.18 ±0.38	2.67 ±0.65	0.30 ±0.10	1.32 ±0.13
			Transformer	0.24 ±0.44	3.24 ±0.71	0.49 ±0.15	1.09 ±0.13	0.60 ±0.58	1.51 ±0.29	0.22 ±0.21	2.54 ±0.80	0.08 ±0.29	1.15 ±0.17	0.11 ±0.18	1.48 ±0.41	0.25 ±0.19	1.45 ±0.29	0.06 ±0.48	2.86 ±0.53	0.30 ±0.12	1.32 ±0.14

Table 17. Results for the RMSE (t/ha) (↓) and the R^2-score (↑) for different models and datasets for the Leave-One-Region-Out (LORO) cross-validation at the field level. ARG = Argentina, BRA = Brazil, GER = Germany, URG = Uruguay. C = corn, R = rapeseed, S = soybean, W = wheat.

Evaluation				Field-Level																	
				ARG-C		ARG-S		ARG-W		BRA-C		BRA-S		BRA-W		GER-R		GER-W		URG-S	
Splitting	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
LORO	Sentinel-2	X	3D-ConvLSTM	0.68 ±1.89	1.59 ±0.48	0.69 ±0.12	0.67 ±0.22	0.80 ±0.21	0.95 ±0.24	0.55 ±0.78	1.19 ±0.36	0.37 ±1.63	0.55 ±0.18	0.47 ±0.30	0.58 ±0.24	0.25 ±0.51	1.16 ±0.31	0.14 ±1.05	1.76 ±0.75	0.68 ±0.52	0.61 ±0.30
		X	3D-LSTM	0.49 ±2.60	2.00 ±0.73	0.70 ±0.13	0.67 ±0.18	0.81 ±0.21	0.92 ±0.16	0.59 ±1.35	1.13 ±0.49	0.41 ±1.89	0.53 ±0.15	0.48 ±0.27	0.57 ±0.23	0.26 ±0.30	1.16 ±0.36	-0.08 ±1.35	1.98 ±1.09	0.65 ±0.90	0.63 ±0.26
		X	LSTM	0.47 ±1.36	2.05 ±0.75	0.64 ±0.15	0.72 ±0.15	0.7 ±0.38	1.18 ±0.26	0.39 ±1.80	1.39 ±0.37	0.28 ±2.07	0.59 ±0.21	0.27 ±0.19	0.68 ±0.23	-0.16 ±0.65	1.45 ±0.49	-2.71 ±3.03	3.66 ±1.26	0.63 ±0.43	0.65 ±0.31
		X	Transformer	0.55 ±2.11	1.88 ±0.46	0.59 ±0.17	0.77 ±0.16	0.71 ±0.46	1.16 ±0.33	0.35 ±0.29	1.43 ±0.59	0.33 ±1.85	0.57 ±0.17	0.37 ±0.28	0.63 ±0.25	0.17 ±0.84	1.22 ±0.39	0.01 ±1.89	1.89 ±0.69	0.63 ±1.11	0.65 ±0.21
	Sentinel-2 + ADM	Feature Fusion	AFF	0.53 ±0.52	1.92 ±0.74	0.78 ±0.07	0.57 ±0.10	0.87 ±0.08	0.78 ±0.13	0.55 ±0.48	1.19 ±0.52	0.63 ±1.71	0.42 ±0.18	0.52 ±0.15	0.55 ±0.20	0.07 ±0.61	1.29 ±0.45	-1.85 ±1.79	3.21 ±1.51	0.60 ±1.44	0.68 ±0.28
			MMGF	0.59 ±0.66	1.80 ±0.54	0.65 ±0.17	0.71 ±0.15	0.78 ±0.12	1.00 ±0.18	0.10 ±0.51	1.68 ±0.66	0.3 ±1.82	0.58 ±0.15	0.23 ±0.82	0.70 ±0.23	-0.56 ±0.90	1.68 ±0.69	-0.95 ±1.02	2.65 ±1.2	0.54 ±0.59	0.73 ±0.30
			3D-ConvLSTM	0.68 ±1.89	1.59 ±0.48	0.69 ±0.12	0.67 ±0.22	0.80 ±0.21	0.95 ±0.24	0.55 ±0.74	1.19 ±0.36	0.37 ±1.63	0.55 ±0.18	0.47 ±0.30	0.58 ±0.24	0.25 ±0.51	1.16 ±0.31	0.14 ±1.05	1.76 ±0.75	0.68 ±0.52	0.61 ±0.3
		Input Fusion	3D-LSTM	0.23 ±2.02	2.46 ±0.99	0.66 ±0.11	0.70 ±0.05	0.74 ±0.14	1.09 ±0.19	0.65 ±0.41	1.05 ±0.37	0.47 ±0.72	0.5 ±0.16	0.48 ±0.32	0.58 ±0.21	0.03 ±0.31	1.32 ±0.44	-0.16 ±0.81	2.05 ±0.83	0.62 ±1.91	0.66 ±0.22
			LSTM	0.43 ±1.68	2.12 ±0.84	0.66 ±0.09	0.71 ±0.10	0.61 ±0.50	1.34 ±0.11	0.38 ±0.59	1.39 ±0.56	0.29 ±1.75	0.58 ±0.12	0.26 ±0.39	0.69 ±0.28	0.20 ±0.91	1.2 ±0.50	-1.39 ±1.02	2.94 ±1.19	0.56 ±0.52	0.71 ±0.3
			Transformer	0.44 ±1.60	2.10 ±0.49	0.67 ±0.09	0.70 ±0.05	0.63 ±0.32	1.30 ±0.31	0.52 ±0.78	1.22 ±0.38	0.20 ±1.47	0.62 ±0.16	0.36 ±0.29	0.64 ±0.27	0.18 ±0.49	1.22 ±0.43	-0.34 ±0.76	2.20 ±0.69	0.59 ±0.29	0.68 ±0.28

Table 18. Results for the RMSE (t/ha) (↓) and the R^2-score (↑) for different models and datasets for the Leave-One-Region-Out (LORO) cross-validation at the subfield (pixel)-level. ARG = Argentina, BRA = Brazil, GER = Germany, URG = Uruguay. C = corn, R = rapeseed, S = soybean, W = wheat.

Evaluation				Subfield (Pixel)-Level																	
				ARG-C		ARG-S		ARG-W		BRA-C		BRA-S		BRA-W		GER-R		GER-W		URG-S	
Splitting	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
LORO	Sentinel-2	X	3D-ConvLSTM	0.54 ±0.36	2.52 ±0.33	0.57 ±0.11	1.01 ±0.10	0.69 ±0.24	1.33 ±0.23	0.33 ±0.14	2.35 ±0.37	0.19 ±0.22	1.08 ±0.13	0.16 ±0.07	1.44 ±0.31	0.12 ±0.29	1.57 ±0.30	0.07 ±0.43	2.84 ±0.69	0.36 ±0.19	1.26 ±0.29
		X	3D-LSTM	0.39 ±0.30	2.91 ±0.56	0.59 ±0.10	0.98 ±0.10	0.71 ±0.24	1.30 ±0.18	0.34 ±0.26	2.33 ±0.53	0.22 ±0.16	1.06 ±0.34	0.20 ±0.07	1.41 ±0.30	0.17 ±0.20	1.52 ±0.45	0.02 ±0.43	2.92 ±0.94	0.36 ±0.31	1.26 ±0.28
		X	LSTM	0.44 ±0.28	2.79 ±0.35	0.55 ±0.10	1.03 ±0.09	0.61 ±0.32	1.48 ±0.24	0.27 ±0.06	2.46 ±0.44	0.16 ±0.21	1.10 ±0.35	0.11 ±0.09	1.48 ±0.34	-0.05 ±0.33	1.71 ±0.54	-0.78 ±1.18	3.93 ±1.16	0.34 ±0.17	1.28 ±0.30
		X	Transformer	0.43 ±0.40	2.80 ±0.40	0.52 ±0.10	1.05 ±0.09	0.61 ±0.37	1.49 ±0.20	0.27 ±0.12	2.47 ±0.50	0.21 ±0.18	1.07 ±0.32	0.11 ±0.10	1.48 ±0.32	0.09 ±0.49	1.59 ±0.31	0.10 ±0.68	2.79 ±1.61	0.31 ±0.27	1.31 ±0.26
	Sentinel-2 + ADM	Feature Fusion	AFF	0.49 ±0.23	2.64 ±0.57	0.65 ±0.07	0.90 ±0.11	0.78 ±0.12	1.12 ±0.17	0.34 ±0.15	2.35 ±0.53	0.35 ±0.15	0.97 ±0.34	0.18 ±0.05	1.42 ±0.30	0.15 ±0.28	1.55 ±0.51	-0.58 ±0.68	3.70 ±1.25	0.34 ±0.23	1.28 ±0.28
			MMGF	0.50 ±0.22	2.63 ±0.46	0.54 ±0.15	1.04 ±0.19	0.68 ±0.14	1.35 ±0.21	0.13 ±0.20	2.69 ±0.68	0.20 ±0.17	1.07 ±0.34	0.07 ±0.15	1.51 ±0.33	-0.13 ±0.51	1.77 ±0.60	-0.34 ±0.46	3.41 ±1.13	0.31 ±0.44	1.31 ±0.29
			3D-ConvLSTM	0.54 ±0.38	2.52 ±0.33	0.57 ±0.11	1.01 ±0.1	0.69 ±0.24	1.33 ±0.23	0.33 ±0.14	2.35 ±0.37	0.19 ±0.22	1.08 ±0.33	0.16 ±0.07	1.44 ±0.31	0.12 ±0.29	1.57 ±0.30	0.07 ±0.43	2.84 ±0.69	0.36 ±0.19	1.26 ±0.29
		Input Fusion	3D-LSTM	0.39 ±0.45	2.90 ±0.53	0.57 ±0.11	1 ±0.08	0.64 ±0.24	1.44 ±0.16	0.37 ±0.12	2.29 ±0.44	0.26 ±0.16	1.04 ±0.32	0.17 ±0.07	1.43 ±0.3	0.03 ±0.32	1.64 ±0.49	-0.02 ±0.20	2.97 ±0.83	0.34 ±0.33	1.28 ±0.26
			LSTM	0.48 ±0.35	2.69 ±0.35	0.54 ±0.09	1.03 ±0.09	0.53 ±0.40	1.64 ±0.31	0.28 ±0.11	2.45 ±0.52	0.12 ±0.24	1.13 ±0.35	0.11 ±0.07	1.48 ±0.32	0.05 ±0.47	1.63 ±0.49	-0.36 ±0.72	3.44 ±0.95	0.32 ±0.39	1.30 ±0.31
			Transformer	0.38 ±0.41	2.93 ±0.42	0.53 ±0.08	1.04 ±0.09	0.53 ±0.31	1.63 ±0.26	0.31 ±0.17	2.39 ±0.49	0.15 ±0.17	1.11 ±0.3	0.12 ±0.1	1.47 ±0.33	0.13 ±0.33	1.56 ±0.43	-0.02 ±0.23	2.97 ±0.68	0.30 ±0.16	1.31 ±0.32

B. Dataset Datasheet

In the following, we will refer to the authors and their institutions in the paper as the "publishing institution."

B.1. Motivation

The motivation to create the dataset was to advance large-scale crop yield prediction. Data creation was funded by industry and public stakeholders.

B.2. Dataset Creation

The dataset can be used for crop yield prediction as described in the main paper. The dataset can be used for further research or other directions, such as crop classification, time series analysis, or land cover classification, provided that data privacy and the data license are complied with.

B.3. Data Collection Process

All the yield data was collected in collaboration with local farmers, data providers, and industry companies. Each data provider had to sign a data-sharing agreement with the publishing institution. Data providers were compensated for data collection and data sharing on an area basis.
The EO data was acquired by the publishing institution from only publicly available data providers under the licenses of those providers.
The dataset does not contain any other data from other publicly available data sources or libraries and is entirely self-contained. The data contains sensitive information about the geographic locations of the local data providers. Each data point is georeferenced. We provide georeferences to facilitate downstream research, including additional data acquisition activities beyond the current dataset.
The dataset may have errors and inconsistencies arising during the data collection process. This may include errors in the yield data collection and preprocessing processes. The dataset may include local biases to individual regions, crop types, and years. Moreover, the data may include data imbalances that could bias downstream models. We provide detailed metadata to help researchers account for these limitations.

B.4. Data Preprocessing

The data preprocessing was described in detail in Sec. 4.2. Statistics to calculate the target scaled yield (dry yield) are given in Tab. 6.

B.5. Data Distribution

The publishing institution owns the data in its entirety and will grant access under a custom license.

B.6. Data Maintenance

The dataset is entirely self-contained. The publishing institution will maintain the data entirely on its internal servers. The publishing institution will maintain the data access portal and provide sufficient information to contact it. The data will be continuously maintained and updated. Updates will be documented on the publishing institution's webpage. We welcome community contributions that improve the data quality.

The dataset is available at https://yieldsat.github.io/

B.7. Legal and Ethical Considerations

The data is owned entirely by the publishing institution. Access will be granted under the described license and data privacy regulation. The data does not contain any information about human subjects or any other confidential information. However, sensitive information is provided in the georeferences of the data.

This dataset is intended for academic research in a non-commercial setting. There are no expectations for this dataset to produce models that can generalize to real-world data, and the data usage is aimed for academic research, for example: to build new yield estimation models, evaluate yield estimation models, and study the yield estimation process itself.