YIPEEO: Yield Prediction and Estimation using Earth Observation

[Experimental Dataset Description v2.0]

[ID EDD]

Version 2.0

[08/05/2024]

Submitted by:

TU Wien – Department of Geodesy and Geoinformation

in cooperation with:

CzechGlobe and EODC

This document was compiled in response to the ESA/AO/1-11144/22/I-EF: ESA Express Procurement Plus – EXPRO+ Theme 2 - Yield Estimation & Forecasting.

This document provides the revised Experimental Dataset Description of the project YIPEEO (EDD).

Number of pages: 27

Authors:	Charis Chatzikyriakou (CC), Emanuel Büechi (EB), Lucie Homolová (LH), Christoph Reimer (CR), Felix Reuß (FR)
Circulation (internal):	Project consortium
External:	ESA

Issue	Date	Details	Editor
0.1	22.2.2024	Template created	CC
0.2	01.03.2024	Added results methods	EB, LH, FR, CR, CC
0.3	04.04.2024	Updated with new results	EB, LH
1.0	05.04.2024	Finalisation and submission of the document	CC
1.1	08.05.2024	Addressed comments received from Ewelina Dobrowolska (ESA)	EB, LH, FR, CR, CC
2.0	08.05.2024	Finalisation and re-submission of the document	CC, EB

For any clarifications please contact Emanuel Büechi (Emanuel.bueechi@geo.tuwien.ac.at).

Table of Contents

List of figures

List of tables

Acronyms

CR	Cross-ratio
DEM	Digital Elevation Model
EO	Earth Observation
EVI	Enhanced Vegetation Index
GRD	Ground Range Detected
LAI	Leaf Area Index
LSTM	Long-short-term memory
LT	Lead-time
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
(ST)MTL	(Spatial-Temporal) Multi-Task Learning
(G)NDVI	(Green) Normalized Difference Vegetation Index
NDWI	Normalized Difference Water Index
P	Precipitation
R^2	coefficient of determination
Rad	Radiation
RB	Requirement Baseline Document
RF	Random Forest
RMAE	Root Mean Absolute Error
(r/ub)RMSE	(Relative/Unbiased) Root Mean Squared Error
S-1 (2)	Sentinel-1 / 2
SIF	Solar Induced Fluorescence
SVR	Support Vector Regression
T	Temperature
VH	Vertical-horizontal polarized
VI	Vegetation Indices
VV	Vertical-vertical polarized
XGB	Extreme Gradient Boosting

Executive Summary

This document constitutes the Experimental Dataset Description (EDD) as a result of Task 4, as described in the proposal of the ESA YIPEEO project. In Section 1, it describes the input data sets, the methodology and the model used for the generation of the Agriculture Science Precursors Experimental Dataset (ASP ED) and it gives a detailed overview of the products and their validation. In Section 2, the crop classification and the crop yield forecasts are presents and discussed and finally, in Section 3, some conclusions and outlook for future work is provided.

1 Agriculture Science Precursors Experimental Dataset (ASP ED)

1.1 Input data

The input data sets that are used for the generation of the ASP ED have been described in the submitted *D2.2 Database Description v1.0*, in Section 2 (in-situ crop yield data), Section 3 (EO data) and Section 4 (Meteorological data). Therefore, please refer to that document for the detailed description. In short, these data sets are:

Earth observation data:

- Sentinel-1

Product	Spatial scale	Spatial coverage	Temporal scale	Temporal coverage
ARD	20 m	Selected countries (AT, CZ, potentially NL)	Several days	2016 - 2023

- Sentinel-2

Product	Spatial scale	Spatial coverage	Temporal scale	Temporal coverage
L2 reflectance	20 m	Selected countries (AT, CZ, potentially NL)	Several days	2016 - 2023

- Copernicus Global Land Service (CGLS) vegetation products

Product	Spatial scale	Spatial coverage	Temporal scale	Temporal coverage
CGLS selected vegetation products	0.05°	Selected countries (AT, CZ, potentially NL)	Daily	2016 - 2023

Meteorological data:

- ERA5-Land

Product	Spatial scale	Spatial coverage	Temporal scale	Temporal coverage
ERA5-Land selected meteorological parameters	0.1°	Selected countries (AT, CZ, potentially NL)	Daily	2016 - 2023

Crop yield data

For the ED we used crop yield data on field scale and on regional scale. For field scale, we used data from the Rostenice farm that was used in the ATBD. On a regional scale, we used the data from Austria and Czechia on NUTS4 level and NUTS2 level data from the Netherlands.

Crop classification

Crop type maps from Austria and Czechia are used for the three crops winter wheat, spring barley and maize for the years 2016-2022.

1.2 Methodology

The project foresees the implementation of two approaches for upscaling. The first approach focuses on upscaling the models to run with low resolution input datasets such as ERA5- Land, CGLS data products or "upscaled" Sentinel products. On the other hand, the second approach is based on higher resolution input data with ML model inference to happen at field scale and finally aggregated to larger spatial units such as defined by the Nomenclature of Territorial Units for Statistics (NUTS4, NUTS3, NUTS2) and per country. Hereafter, we will concentrate more closely on the overall methodology followed in respect to the model inference at field scale level.

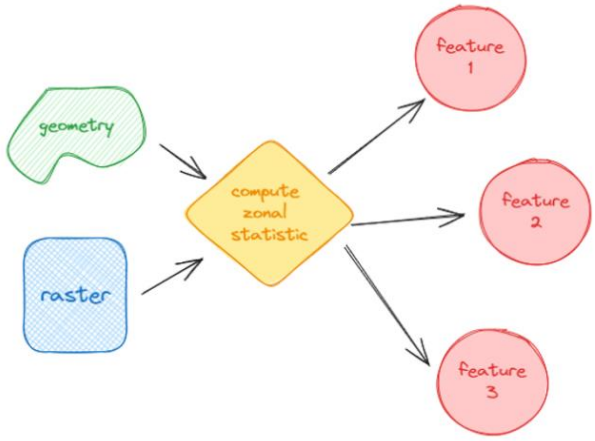

Figure 1: Overall concept of feature extraction.

ML models require certain features to be trained on to finally make predictions or draw conclusions from new, unseen data provided to it. This process is also known as feature extraction transforming raw data such as raster data, as provided by Earth Observation (EO) satellites, into a format suitable for model training. Features are the individual measurable properties or characteristics of the data that are relevant to the problem being addressed. In the EO domain these are typically deduced via

zonal statistics computing a summary of statistics or metrics for a specific geographic zone or region, defined by a geometry object, within satellite imagery or raster datasets. Nowadays, Earth Observation data is represented as raster datacubes, which are conceptually multi-dimensional arrays with additional information about their dimensions. Applying zonal statistics, commonly also referred to as spatial aggregation, to such raster datacubes will result in a new datacube holding features with certain dimensions. This newly created datacube is referred to as vector datacube.

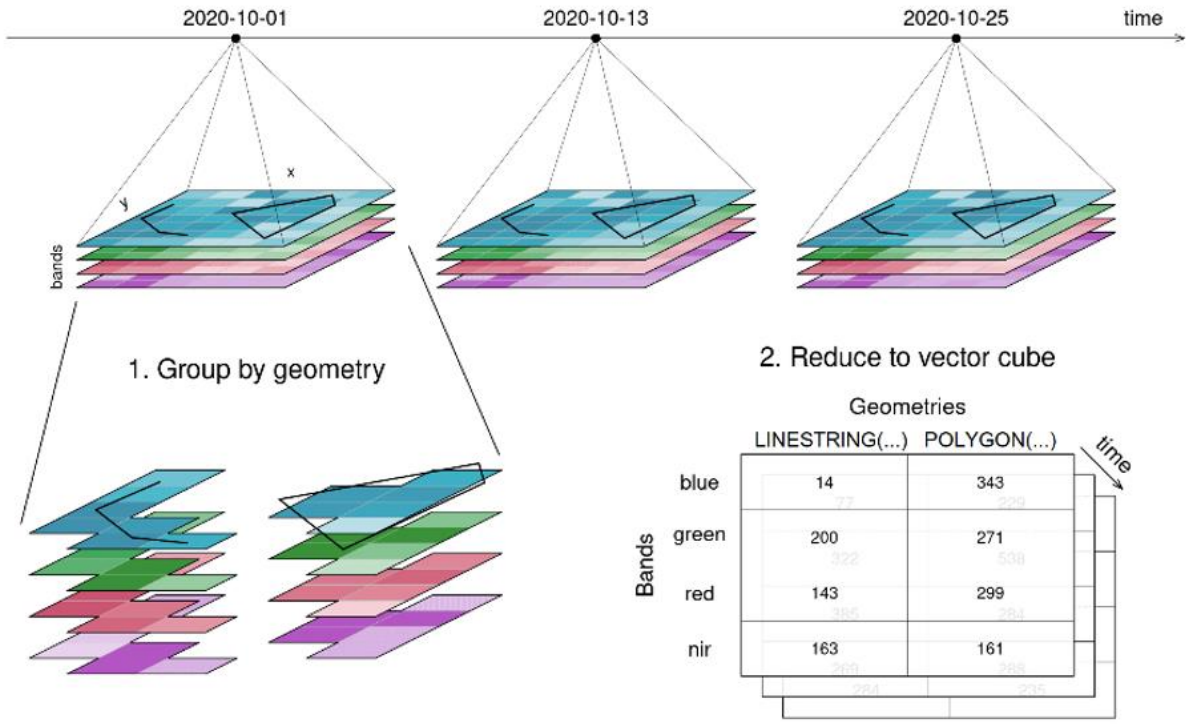

Figure 2: Vector datacube representation [https://openeo.org/documentation/1.0/datacubes.html#what-are-datacubes]

With reference to this project, field boundaries are acting as geometry objects in the vector datacube. Accordingly, features extracted from the satellite imagery (Sentinel-1, Sentinel-2, etc.) can be merged with crop yield information of the individual fields to be used for model training. In case of supervised models, such as used in this project, the following workflow is utilised to finally predict crop yield on a field scale level. For model training, agricultural field data is collected through a network of collaborative farms and entities providing information about crop yields at field scale level. This labelled dataset is used with a set of selected input features derived from satellite imagery to train a given ML model. Satellite imagery (raster datacubes) covering other fields and time periods of interest are used to generate a vector data cube holding the required input features. This vector datacube acts as input to the trained ML model to predict crop yields for those. The result can again be expressed as

a vector datacube holding the needed information about crop yields at field level. Upscaling of this information at NUTS3 and NUTS2 level is done again via zonal statistics (spatial aggregation) by computing the median of all fields within a certain territorial unit. This is done per crop type and year.

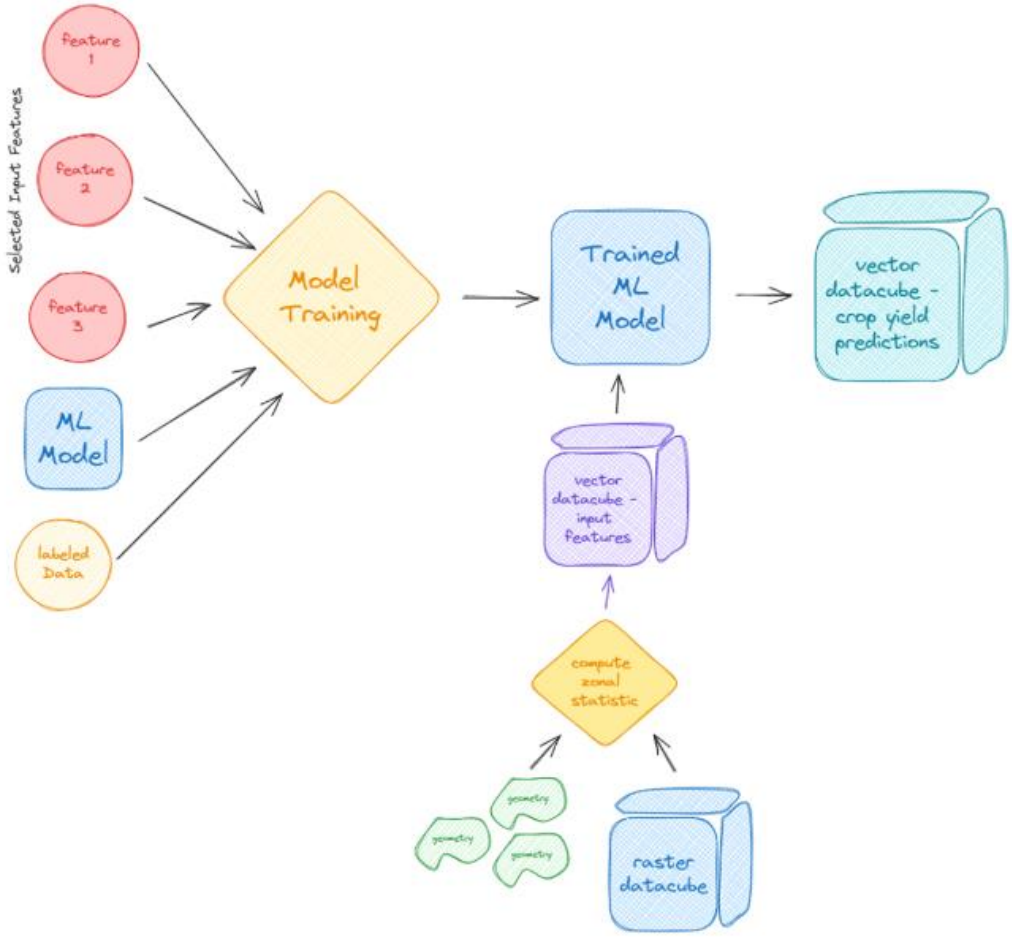

Figure 3: General ML workflow for crop yield prediction.

1.2.1 Data extraction

Data extraction, or more specifically feature extraction, is done individually per input dataset. For each input dataset a separate data extraction workflow is developed within Python. Computation of these Python functions is executed via the Python based parallel and distributed computing framework Dask. Dask allows for efficient execution of embarrassingly parallel workloads which is exactly the use case of feature extraction and model inference. Running workloads on Dask is one cloud native service offered by EODC, complementing the big data processing services such as openEO. Ultimately, EODC envisions to implement the final workflow as an openEO process graph for further reuse. Such "user defined process" (UDP) can be shared with other users following the openEO metadata description of processes. Therefore, a user can finally select to either run the code directly on Dask or execute the same code via and openEO UDP.

The collected agricultural field data was converted and ingested into an OGC API – Features compliant server to allow for easy access to the data. Exposing the data via an API standard such this allows to consume the data by various clients. OGC API – Features enable to query the given database and to extract only data which is needed for further processing as simple GeoJSON object. Hereafter, a complete list of features extracted per input dataset will be presented to be further explored for model training and inference.

1.2.1.1 Sentinel-1

Sentinel-1 data for the years 2016-2023 was pre-processed by TUW RS on the Vienna Scientific Cluster using the software SNAP8 and software packages developed by the TUW RS group. The processing workflow consists of the following steps:

1. Apply precise orbit data
2. Border-noise removal
3. Radiometric calibration
4. Radiometric terrain-flattening
5. Range-Doppler terrain correction

For steps 4. and 5. the 30 m Copernicus Digital Elevation Model (DEM) was used. To extract time series on field level from the pre-processed Sentinel-1 data, several further processing steps were performed to mitigate the impact of the viewing geometry and undesired objects within or near the fields. In a first step, an incidence angle normalization to 40° was performed. Afterwards, all pixels below a standard deviation of 5dB within one year were filtered out as they are typically stemming from radar shadow pixels or are no crop pixels. Finally, the cross-ratio was calculated by subtracting VV and VH polarized backscatter. The final time series were then stored as NetCDF files per field.

The aggregation of the time series from field level to NUTS2/NUTS3 level was performed in the linear domain to retrieve mathematically correct results.

1.2.1.2 Sentinel-2

The Sentinel-2 L2A collection is used to compute a set of features based on the provided bands as well as various vegetation indices. The Sentinel-2 L2A data cube is dynamically created by utilising the STAC API. The datacube is pre-filter with scenes of a cloud cover less than 80%. The following features are extracted per field and timestamp:

- Band Medians and Standard Deviations
    - B02, B03, B04, B05, B06, B07, B08, B8A, B11, B12

- Vegetation indices based on median bands
    - NDVI
    - EVI
    - NDWI
    - NMDI

An outlier removal was added on a field scale level utilising the SCL band and outlier removal based on 2 x inter quartile range (IQR). Finally, all the data is stored in NetCDF files for further use in the workflow. The results can also be converted to zarr files if this is required for publishing the datasets.

1.2.1.3 ERA-5 Land

ERA5-Land data was extracted directly on NUTS3 and NUTS4 level. Hence, all pixels that lay within the regions were aggregate by taking the median per timestep. The daily mean values per variable were aggregated to monthly timesteps. The used variables from ERA5-Land are:

- Sum of reference evapotranspiration
- Total precipitation
- Potential evapotranspiration
- Surface net radiation
- Mean daily temperature
- Volumetric soil water content of top layer (0-7 cm)

The extracted data can be shared via TU Wien's repository. The raw data was accessed from the climate datastore. Therefore, a further redistribution of it is not planned.

1.2.2 **Model**

The models are the same as in the ATBD, Random Forest and Extreme Gradient Boosting. Also the model setup as shown in Figure 4 is the same. The first crop yield forecasts are calculated starting 4 months before the harvest with the predictor data available by then. Every month, new data is added and a new crop yield forecast are calculated (Figure 4). We use the crop types winter wheat, spring barley, and maize.

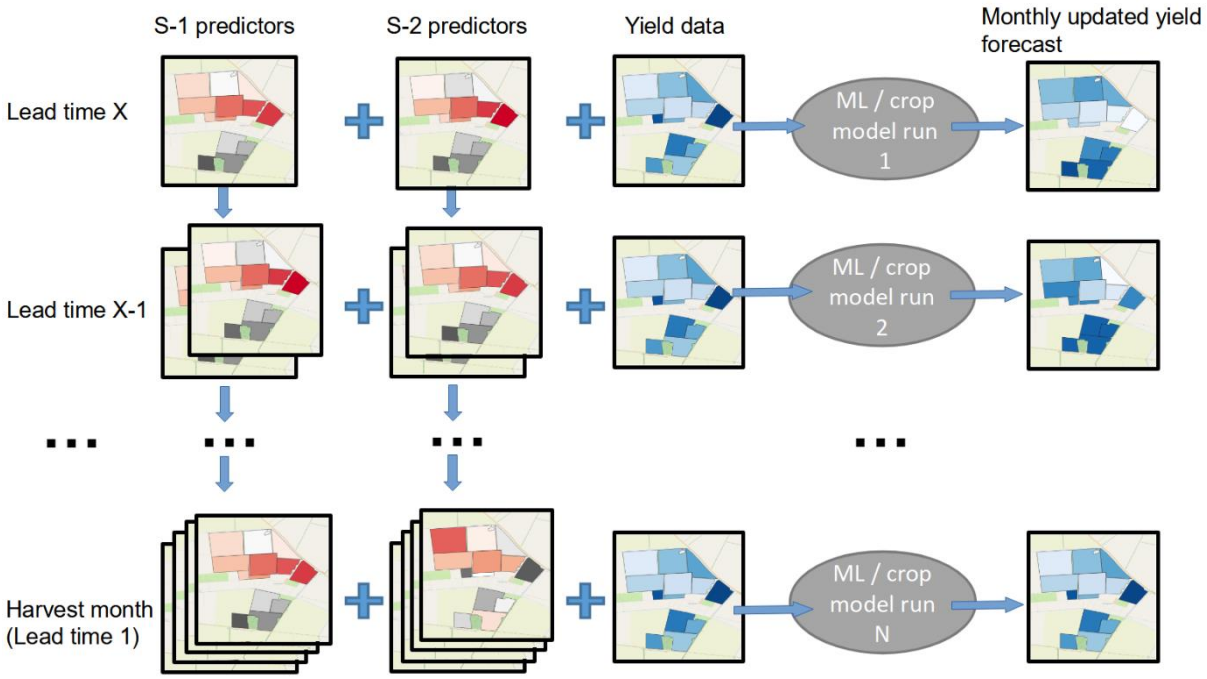

Figure 4: Setup of the crop yield forecasts on a field-scale

1.2.2.1 Crop classification

To obtain information on crop type during the season for the regions selected for upscaling, a Long Short Term Memory (LSTM) network for crop classification was developed. This approach uses Sentinel-1 and Sentinel-2 time series to predict crop type early in the season. The model was trained with time series for 6 major crops in the Czech Republic and afterwards applied in the Netherlands to test the transferability of the model. 3/5 of the available data for each crop type were used for model training, 1/5 for the model validation, and 1/5 for the testing and calculation of accuracy metrics. The achieved results are outlined in chapter 3.

1.3 Products

Resulting from the data preparation described in 2.2, we have three datasets for testing and training the crop yield forecasts per crop type. These are EO-regional (Sentinel data aggregated to NUTS4 regions), EO-field (Sentinel data for the fields of Rostenice farm), and ERA-regional (ERA5-Land data extracted per NUTS4 region). Based on these, we envisaged different ways of training and testing the crop yield forecasts (Figure 5):

1. Train and test the model on a regional scale for Austria and Czechia. Once with EO data only, once with ERA5-Land data only, and once using both.
2. Train and test the model on field-scale using EO-field from Czechia (as done in the ATBD)

3. Train the model using EO-field data from Rostenice and test it on regional scale for Austria and Czechia
4. Same as 3 but vice versa: train the model with regional data from Czechia and Austria and test it on field data in Rostenice

We selected these approaches as we want to know how well crop yields can be forecasted using Sentinel data on a field and regional level and compare it to forecasts based on ERA5-Land data. The setups 3 and 4 are tested to see if the same information can be used independently of the scale. As field-scale crop yield data is hard to get, it would help a lot if forecasts could be trained using only regional-scale crop yield data which is much easier to get.

All these approaches are tested for all three crops, winter wheat, spring barley, and maize. Most model runs on regional level are done on NUTS4 level. Only Maize in Czechia is modelled on NUTS3 level, as the crop yield data is not available on NUTS4.

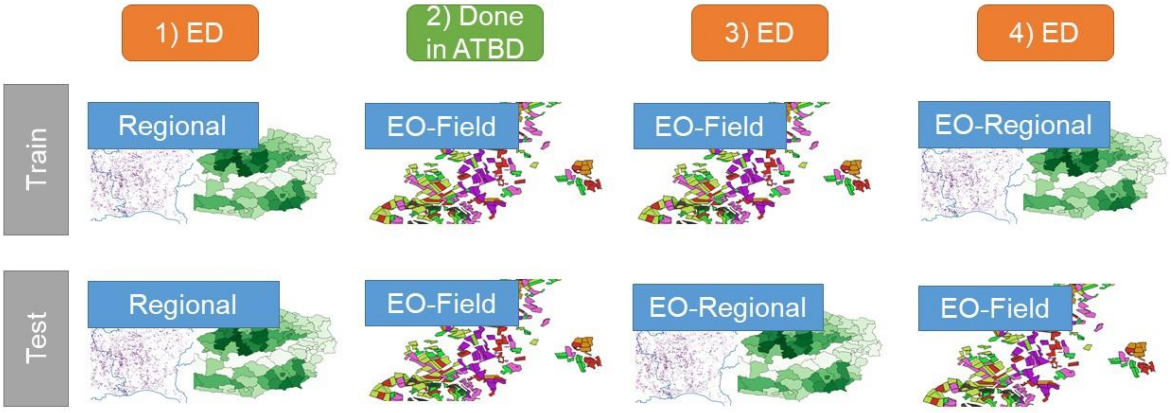

Figure 5: Overview of the different test-train splits across scales. For 2) to 4) only EO data is used (Sentinel-1 and 2), while for 1) ERA5-Land data is used as predictor too.

1.4 Product validation

The product validation uses the similar approach as described in the Product Validation Report (PVR, D3.2). Approximately 60% of the data was used for model training, 20% for validation (i.e. for the model optimization and cross-validation) and the remaining 20% for independent testing. The set-up for ED product validation is divided as follows:

1. Random training-validation split using 30-fold random cross-validation that was used during model training

2. A holdout validation was used for the final testing dataset. This was used to make it easier comparable to 3) which is based on a simple holdout testing too. For the holdout validation we randomly split the data once into a test and train set.
3. Scale split using data from field-level and apply it to regional scale and vice versa.

Following statistics, same as described in PVR (D3.2), were computed to validate the crop yield estimates:

- Pearson's coefficient of determination – R^2 (eq. 2 in D3.2),
- Bias – B (eq. 3 in D3.2),
- Root mean square error - RMSE (eq. 4 in D3.2),
- Relative root mean square error – rRMSE (eq. 5 in D3.2), and
- Unbiased root mean square error – ubRMSE (eq. 6 in D3.2)

$$explained\ variance = 1 - \frac{var(Oi-Pi)}{var(Oi)} \quad \text{eq.1}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Pi-Oi)^2}{\sum_{i=1}^{n}(Oi-\bar{O})^2} \quad \text{eq. 2}$$

$$B = \frac{1}{n}\sum_{i=1}^{n}(Pi - Oi) \quad \text{eq. 3}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Pi - Oi)^2} \quad \text{eq. 4}$$

$$rRMSE = \frac{RMSE}{\bar{O}} \quad \text{eq. 5}$$

$$ubRMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Pi - Oi)^2 - B^2} \quad \text{eq. 6}$$

were *Pi* are predicted, *Oi* observed yield values, $\bar{O}$ is mean of observed yields.

2 Results

2.1 Crop Classification

Table 1 illustrates the achieved accuracies of the model in Czech Republic for one year excluded from the training data. The columns indicate the different months at which end the crop classification was

performed. A prediction at the end of May means a time series from January until end of May was used as a model input.

Table 1: Achieved F1 scores for major crops and different month.

Crop	N samples	March	April	May
Winter wheat	182	0.62	0.66	0.72
Maize	302	0.33	0.45	0.76
Spring barley	236	0.37	0.35	0.82
Winter barley	81	0.24	0.29	0.66
Soya*	30	0.12	0.17	0.45
Winter rape	185	0.9	0.92	0.94
Overall	1006	0.55	0.59	0.78

*Lower accuracy due to low sampling size

As Table 1 indicates, the accuracy significantly declines if the prediction is done before May. In contrast, the difference between a prediction at the end of April and end of March is only minor. This can be explained that the development of heads/ears is typically happening around May and crucial to distinguish the crops in the time series. Among the crop types, winter rape achieved the highest accuracy. Soya, however, has the lowest F1 score but might be impacted by a lower number of samples. Misclassification occurs especially between the very similar crops winter wheat and winter barley.

In a next step the model was applied on the same crops for one season at the end of April in the Netherlands. As the available data for the Netherlands was very low and did not include all the included crop types, existing time series from a previous project were used for demonstration purposes. First, the model was tested without retraining it with samples from the Netherlands. As Table 2 outlines the model performed poorly in this case. With a value of 0.29, the overall accuracy is significantly lower compared to the Czech Republic. However, when the model is retrained with a small amount of data from the Netherlands from a different season, the accuracy exceeds even the one from Czech Republic. A major factor for the higher accuracy can be seen in the much higher F1 score of Soya which was not impacted by a low sampling size. Overall, this case study demonstrates the possibility to apply crop classification models in various countries by retraining them on a small sampling size. For (neighboring) countries with very similar vegetation periods, comparable weather conditions and similar sowing and

harvest dates the retraining might not even be necessary. Adding meteorological data is also expected to improve the overall accuracy and transferability of the model.

Table 2: Achieved F1 scores of the same model applied in the Netherlands with (right) and without (left) retraining at end of April.

Crop	N Samples	F1 Score (no	F1 Score (5 epochs)
Winter wheat	1000	0.36	0.47
Maize	1000	0.25	0.76
Spring barley	1000	0.21	0.63
Winter barley	1000	0.29	0.55
Soya*	1000	0.17	0.72
Winter rape	1000	0.22	0.91
Overall	6000	0.29	0.67

2.2 Crop yield forecasts

2.2.1 Preliminary results from version 0.2

The results of the crop yield forecasts for the different methods are summarized in Table 3 and Table 4. It showed that the crop yield forecasts for maize performed well starting around 3 months before the harvest. The field-level forecast using Sentinel-2 data with Random Forest (RF) achieved a Pearson´s correlation (R) of 0.76 at that point, while the regional forecasts with Sentinel-2 has an R of 0.67. The models using Sentinel-2 data also outperformed the forecast based on ERA5-Land, which achieved an R of 0.57 three months before harvest. Comparing the two models RF and Extreme Gradient Boosting (XGB) shows that RF outperformed XGB for most combinations of training data and leadtimes. The worse performance of the models for a leadtime of 4 months is consistent in all model setups and has already been concluded in the PVR. A potential way to improve this will be to include seasonal weather forecast data. Another potential improvement will be to combine Sentinel data and ERA5-Land data.

The approach to train the data on field level and apply it on regional level showed a lower performance than the models trained and tested on field or regional level. The model trained on regional level and tested on field level showed a similar performance. Only for LT3 the results were significantly lower, which could still be caused by chance, as it is not observed in the other lead-times. However, there are still some factors that seem promising for the approach of training and testing at different scales. First,

the model may still improve when Sentinel-1 data is used too. Secondly, using NUTS4 regions from Czechia could improve the performance, as the field-level data is from Czechia too. Thirdly, retraining with a small amount of field-data, like done for the crop classification, could also help. Due to the mentioned reason of limited data availability on field-level, this approach of training with regional data and testing on field level is worth being further pursued.

Table 3: Performance of the different crop yield forecasts using Random Forests. The values show the Pearson's correlation between the forecasted maize yields and the observed maize yields for the testing data. The rows that are named with EO are based on Sentinel-2 data only, while the last row is trained with ERA5-Land data. Regional shows the results of Austria NUTS4 level, field shows the results from the field level data from the Rostenice farm, while regional2field show the performance of the model trained with Austria NUTS4 level data and tested on the fields from Polkovice and vice versa is field2regional. LT1 to LT4 stand for leadtimes, where LT1 is one month before harvest, LT2 two months, and so on.

RF Maize	LT4	LT3	LT2	LT1
EO-regional	0.49	0.67	0.76	0.75
EO-field	0.4	0.76	0.78	0.79
EO-regional2 field	0.03	0.46	0.54	0.6
EO-field2 regional	0.03	0.25	0.56	0.58
ERA5-regional	0.46	0.57	0.58	0.6

Table 4: Same as Tab. 3 for using Extreme Gradient Boosting instead of Random Forest.

XGB Maize	LT4	LT3	LT2	LT1
EO-regional	0.35	0.55	0.66	0.74
EO-field	0.27	0.71	0.76	0.74
EO-regional2 field	0.09	0.43	0.57	0.5
EO-field2 regional	0.13	0.26	0.5	0.42
ERA5-regional	0.41	0.39	0.46	0.61

2.2.2 Crop yield forecasts validation

1) **Train and test the model on a regional scale for Austria and Czechia. Once with EO data only (combination of Sentinel-1 and Sentinel-2), once with ERA5-Land data only, and once using both.**

Figure 6 and Table 5 show the validation statistics at the regional scale. Regional models trained and applied at NUTS4 (NUTS3 for maize in Czechia, respectively) level show trends of improving accuracy

with lead time closer to the harvest. Winter wheat and maize show high R^2-values (>0.45) from around 3 months before harvest when only using Sentinel data. Adding ERA5-Land data decreases the performance. This is related to the high number of predictors used when combining S1, S2, and ERA5-Land (16 predictors for each timestep, i.e. 64 predictors at leadtime 1). This is rather too much considering a training dataset size of a few hundred datapoints. If the number of training datapoints are not much higher than the number of features, overfitting is very likely. Hence, feature selection will be key to reduce this. For spring barley, on the other hand, we cannot observe this. In this case, combining Sentinel data and ERA5-Land led to the best results. The overall performance is still lower than for winter wheat and maize, though. The combination of ERA5-Land data and EO data seems promising in a way that we can find the best predictors for each crop. I.e., spring barley yield forecasts perform much better when using ERA5-Land data over EO based forecasts, while the contrary is true for maize and winter wheat. For an optimal model setup joining the advantages of the three datasets, we will spend more time optimizing the number of features and test other ways of feature selection.

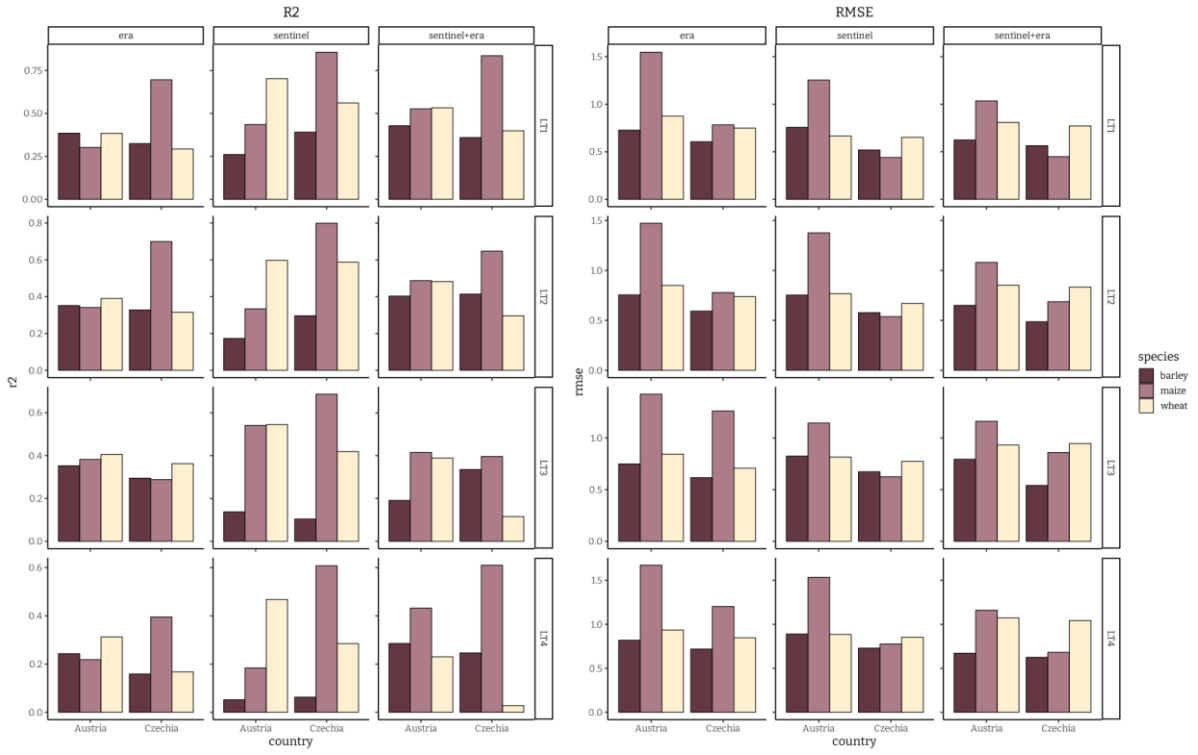

Figure 6: Statistics for XGB regional models for Austria and Czechia trained with different input data (era – using climatic variables from ERA-5 Land only, sentinel – using variables derived from Sentinel-1 and Sentinel-2 data only, all – combination of ERA-5 Land and Sentinel variables). Please, note that results for the winter wheat model are based on Sentinel-2 predictors only.

Table 5: Summary of the validation statistics for the XGB regional models for Austria and Czechia trained with different input data.

Predictors		**ERA-5 Land**					**Sentinel-1 and Sentinel-2**					**ERA and Sentinel combined**				
Crop	Lead	R^2	Bias	RMSE	RRMS	ubRM	R^2	Bias	RMSE	RRMS	ubRM	R^2	Bias	RMSE	RRMS	ubRM
Austria																
Spring	1	0.39	-0.01	0.73	0.17	0.73	0.26	-0.26	0.76	0.19	0.71	0.43	-0.09	0.62	0.15	0.62
barley	2	0.35	-0.01	0.76	0.18	0.76	0.17	-0.11	0.75	0.19	0.75	0.40	-0.13	0.65	0.16	0.64
	3	0.35	0.00	0.75	0.17	0.75	0.14	-0.23	0.82	0.21	0.79	0.19	-0.18	0.79	0.19	0.77
	4	0.24	0.05	0.82	0.19	0.82	0.05	-0.24	0.89	0.23	0.86	0.29	0.04	0.67	0.16	0.67
Grain	1	0.30	-0.02	1.55	0.15	1.55	0.44	-0.07	1.25	0.12	1.25	0.53	0.11	1.04	0.10	1.03
maize	2	0.34	-0.03	1.47	0.15	1.47	0.33	-0.14	1.38	0.13	1.37	0.49	0.01	1.08	0.10	1.08
	3	0.38	-0.13	1.42	0.14	1.42	0.54	-0.01	1.14	0.11	1.14	0.41	0.01	1.16	0.11	1.16
	4	0.22	-0.05	1.67	0.17	1.67	0.18	-0.29	1.53	0.15	1.51	0.43	0.02	1.16	0.11	1.16
Winter	1	0.38	0.02	0.87	0.14	0.87	0.70	0.07	0.67	0.11	0.66	0.53	-0.02	0.81	0.13	0.81
wheat	2	0.39	0.05	0.85	0.14	0.85	0.60	-0.02	0.77	0.13	0.77	0.48	0.05	0.85	0.14	0.85
	3	0.41	0.03	0.84	0.14	0.84	0.54	0.00	0.82	0.13	0.82	0.39	0.10	0.93	0.15	0.93
	4	0.31	-0.01	0.94	0.15	0.94	0.47	0.05	0.88	0.14	0.88	0.23	0.10	1.07	0.17	1.07
Czechia																
Spring	1	0.32	-0.03	0.61	0.12	0.61	0.39	0.08	0.52	0.10	0.51	0.36	0.08	0.56	0.11	0.56
barley	2	0.33	-0.05	0.59	0.12	0.59	0.30	0.15	0.58	0.11	0.56	0.41	0.04	0.49	0.09	0.49
	3	0.29	-0.06	0.62	0.12	0.61	0.10	0.12	0.67	0.13	0.66	0.34	0.06	0.54	0.10	0.54
	4	0.16	-0.06	0.72	0.14	0.72	0.06	0.12	0.73	0.14	0.72	0.25	0.09	0.63	0.12	0.62
Grain	1	0.70	-0.08	0.78	0.10	0.78	0.86	0.00	0.44	0.06	0.44	0.84	0.12	0.45	0.06	0.43
maize	2	0.70	0.06	0.78	0.10	0.78	0.80	0.15	0.54	0.07	0.52	0.65	0.05	0.69	0.09	0.68
	3	0.29	-0.13	1.26	0.16	1.25	0.69	0.06	0.62	0.08	0.62	0.40	-0.07	0.86	0.11	0.86
	4	0.40	-0.49	1.20	0.15	1.10	0.61	0.28	0.78	0.10	0.73	0.61	-0.10	0.68	0.09	0.67
Winte	1	0.29	0.00	0.75	0.12	0.75	0.56	-0.09	0.65	0.11	0.64	0.40	-0.08	0.77	0.13	0.77
wheat	2	0.32	0.03	0.74	0.12	0.74	0.59	-0.19	0.67	0.11	0.64	0.30	-0.17	0.83	0.14	0.82
	3	0.36	-0.01	0.71	0.12	0.71	0.42	-0.22	0.77	0.13	0.74	0.12	-0.14	0.95	0.16	0.94
	4	0.17	0.05	0.85	0.14	0.85	0.28	-0.22	0.85	0.14	0.82	0.03	-0.19	1.04	0.17	1.03

2) Train and test the model on field-scale using EO-field from Czechia (as done in the ATBD)

Figure 7 show the validation statistics at the field scale. The XGB models were trained with EO predictors at the field level (Rostenice farm, Czechia). These results are similar to the ones presented

in the PVR. The only difference is that we used a holdout validation here and a cross-validation for the PVR. Still, for a thorough comparison of the different model setups, we wanted to show the results here again. The main conclusion of the PVR remains the same: the performance of the crop yield forecasts at the field level are only satisfactory for maize. There we reach R^2-values >0.45 from 3 months before harvest. Winter wheat shows an acceptable performance for LT1 (R^2 = 0.53) but not earlier than that. Spring barley shows again the worst overall performance. Here, adding ERA5-Land data may again help, but was not tested so far due to the relatively low spatial resolution which would lead to having only one or two pixels for all fields. ERA5-Land data, however, can improve the performance of a field-scale model, if it is trained on data from different regions to consider impacts of different weather conditions and shifts in crop phenology on crop yields.

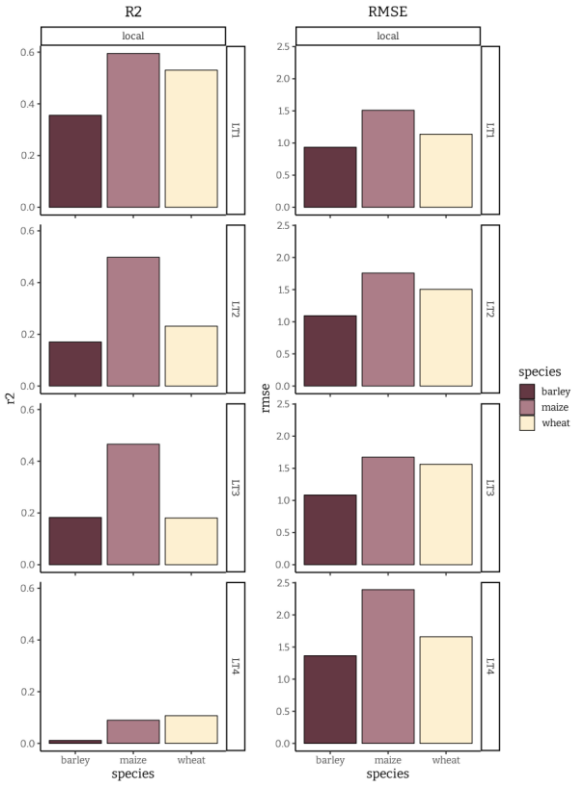

Figure 7: Statistics for XGB models trained at field-level data from Czechia (Rostenice farm) and applied at the field-level in Czechia. The same results as in described in ATBD (D3.1). Please, note that the winter wheat models are trained with Sentinel-2 predictors only. For the other two crops Sentinel-1 and Sentinel-2 data is used.

Table 6: Summary of the validation statistics for the XGB model trained at field-level data from Czechia (Rostenice farm) and applied at field-level in Czechia.

Crop type	Lead time	R^2	Bias [t/ha]	RMSE [t/ha]	rRMSE [%]	ubRMSE
Spring	1	0.36	0.00	0.93	0.16	0.93
barley	2	0.17	0.14	1.09	0.19	1.08
	3	0.18	0.19	1.08	0.19	1.07
	4	0.01	0.07	1.36	0.23	1.36
Grain	1	0.59	-0.26	1.51	0.14	1.49
maize	2	0.50	-0.39	1.76	0.16	1.71
	3	0.47	-0.11	1.67	0.15	1.67
	4	0.09	-0.54	2.39	0.22	2.33
Winter	1	0.53	-0.01	1.14	0.18	1.14
wheat	2	0.23	-0.21	1.50	0.23	1.49
	3	0.18	0.05	1.56	0.24	1.56
	4	0.11	-0.12	1.66	0.26	1.66

3) Train the model using EO-field data from Rostenice and test it on regional scale for Austria and Czechia

These two parts (3 and 4) are now focusing on the transferability of the machine learning models. I.e., we checked if the model can be trained with field data and applied to regional scale here and vice versa in the section 4. In short: the answer is no, they cannot. Figure 8 and Table 7 show that the first attempt of upscaling the model was not very successful. The highest achieved R^2-value was 0.24 for maize at LT1 for Austria and 0.47 for LT2 over Czechia. However, further ways to improve the performance can be explored. This could include a combined training (training the data with field scale and update it with a few regional observations), or again a more in-depth feature selection. Especially, for maize there seem to be some potential to achieve an acceptable performance when further improving the model.

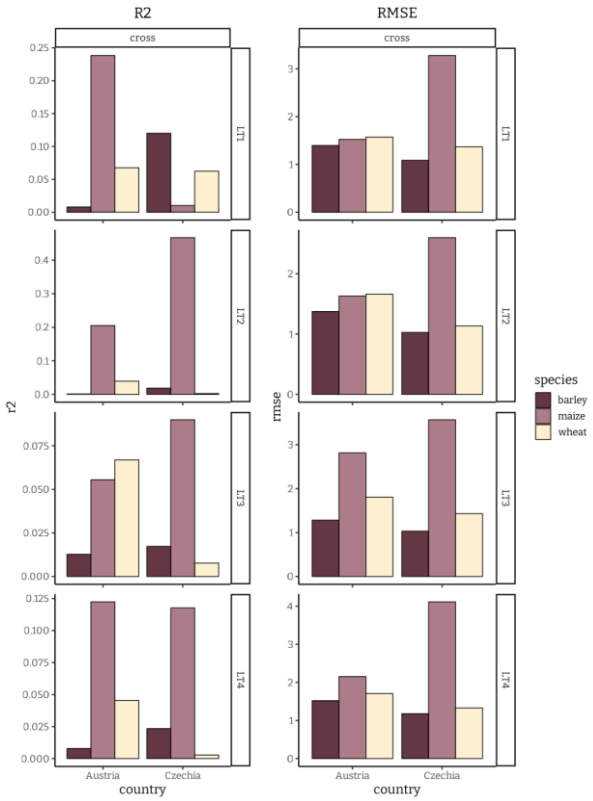

Figure 8: Statistics for XGB models trained at field-level data from Czechia (Rostenice farm) and applied at regional scale (NUTS4 data) in Austria and Czechia. Please, note that the winter wheat models are trained with Sentinel-2 predictors only. For the other two crops Sentinel-1 and Sentinel-2 data is used.

Table 7: Summary of the validation statistics for the XGB model trained at field-level data from Czechia (Rostenice farm) and applied at regional scale (NUTS4 data) in Austria and Czechia.

		Austria					Czechia				
Crop	**Lead**	**R^2**	**Bias**	**RMSE**	**rRMSE**	**ubRMS**	**R^2**	**Bias**	**RMSE**	**rRMSE**	**ubRMS**
Spring	1	0.01	0.58	1.40	0.33	1.27	0.12	0.13	1.09	0.21	1.08
barley	2	0.00	-0.63	1.37	0.33	1.22	0.02	-0.25	1.03	0.20	1.00
	3	0.01	-0.61	1.28	0.30	1.13	0.02	-0.24	1.03	0.20	1.00
	4	0.01	-0.92	1.52	0.36	1.21	0.02	-0.65	1.18	0.23	0.98
Grain	1	0.24	-0.62	1.52	0.14	1.39	0.01	-2.81	3.28	0.40	1.70
maize	2	0.21	-0.76	1.63	0.15	1.44	0.47	-2.32	2.60	0.31	1.18
	3	0.06	-1.89	2.81	0.26	2.08	0.09	-3.15	3.57	0.43	1.69
	4	0.12	-1.20	2.15	0.20	1.79	0.12	-3.65	4.11	0.50	1.90
Winter	1	0.07	0.40	1.57	0.26	1.52	0.06	-0.08	1.37	0.22	1.36
wheat	2	0.04	0.42	1.66	0.27	1.61	0.00	0.32	1.13	0.18	1.09
	3	0.07	0.04	1.81	0.30	1.81	0.01	-0.26	1.43	0.23	1.41
	4	0.05	0.11	1.71	0.28	1.70	0.00	0.33	1.33	0.22	1.29

4) Same as 3) but vice versa: train the model with regional data from Czechia and Austria and test it on field data in Rostenice

Similarly to the results in 3) the crop yield forecasts here show a much lower performance than when the model is trained and tested on solely either on the field or regional level. Again, the best performance is achieved for maize, which shows a R^2 of >0.2 for LT2 and LT1. The results of LT3 and LT4 are show bad results for all crops (R^2 around 0). This is not seen in the RMSE which remains relatively constant over the leadtimes. However, the RMSE is overall quite high (>22%) (see Tab. 8 and Fig. 9). The low overall performance of spring barley and winter wheat make further conclusions difficult. For maize, the jump in performance from LT2 to LT3 is unexpected. When training and testing on field scale this jump occurred a month earlier (between LT3 and LT4) (Fig. 7). As the overall performance is much lower too, though, further optimizations are requried. These could be achieved by either updating the trained model use feature selection or combining the Austrian and Czechian dataset to have a larger training dataset.

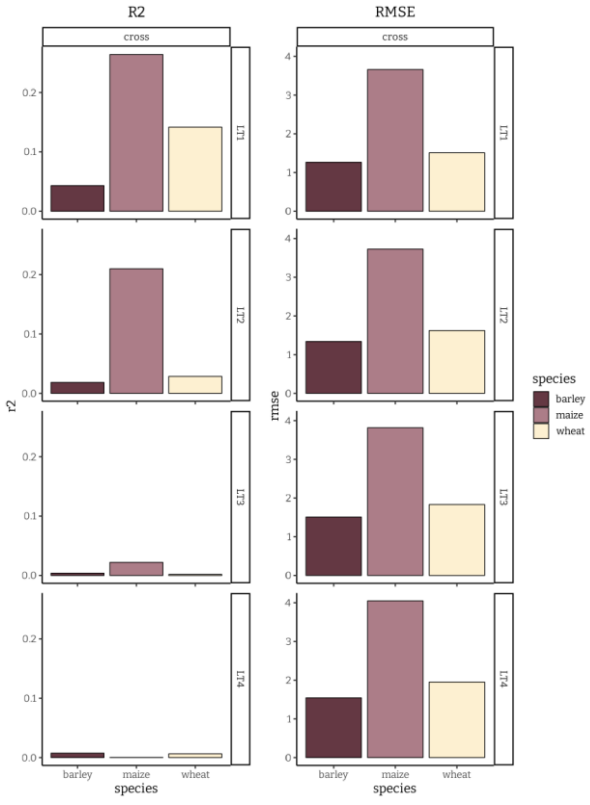

Figure 9: Statistics for XGB models trained at regional-level (NUTS4 data) from Czechia and Austria and applied to field-level (Rostenice farm). Please, note that the winter wheat models are trained with Sentinel-2 predictors only. For the other two crops Sentinel-1 and Sentinel-2 data is used.

Table 8: Summary of the validation statistics for the XGB model trained at regional-level (NUTS4 data) from Czechia and Austria and applied at field-level (Rostenice farm).

Crop type	Lead time	R^2	Bias [t/ha]	RMSE [t/ha]	rRMSE [%]	ubRMSE
Spring	1	0.04	0.24	1.27	0.22	1.24
barley	2	0.02	0.42	1.34	0.23	1.27
	3	0.00	0.72	1.51	0.26	1.33
	4	0.01	0.83	1.54	0.27	1.30
Grain	1	0.26	2.93	3.66	0.33	2.20
maize	2	0.21	2.97	3.73	0.33	2.25
	3	0.02	2.85	3.82	0.34	2.54
	4	0.00	3.02	4.05	0.36	2.70
Winter	1	0.14	0.35	1.51	0.23	1.47
wheat	2	0.03	0.34	1.62	0.25	1.59
	3	0.00	0.49	1.83	0.28	1.77
	4	0.01	0.50	1.95	0.30	1.89

3 Conclusion

For the Agriculture Science Precursors Experimental Dataset (ASP ED), we upscaled the field scale crop yield forecasts to NUTS3 and NUTS4 level for three main crops: winter wheat, spring barley, and maize. The results showed that there are large differences between the methods and the crops. Maize and winter wheat showed promising results from around 2 months before the harvest (R^2>0.5) on a regional level when the model is trained on a regional level using EO and ERA5-Land data. Spring barley yields, on the other hand, seem to be harder to forecast. Generally, a combination of ERA5 Land reanalysis and EO data helped a lot to improve the results. EO data alone provided already useful results, but with adding reanalysis data the performance often improved.

The train-test split between scales, training on field level – applying on regional level and vice versa, led to less good results. Only maize showed some potential (LT2 for Czechia R^2>0.45) that the learnings may be transferable. In a next step, we will further explore the potential of this cross-training and try to improve the performance of the models on field and regional scale. For this, we will test various methods: 1) updating the model, i.e., training the model on regional scale, updating it with some field data, and test it on the remaining field data. 2) more profound feature selection, 3) using more training data with either combining several countries or use the newly available Spain field-level dataset, 4) using different machine learning models as for example LSTM. After these potential

improvements, we will publish the experimental dataset consisting of the forecasts and the predictor dataset on TU Wien's research data repository (https://researchdata.tuwien.ac.at/).