YIPEEO: Yield Prediction and Estimation using Earth Observation

[Algorithm Theoretical Basis Document v2.0]

[ID ATBDv2]

Version 2.0

[03/03/2025]

Submitted by:

TU Wien – Department of Geodesy and Geoinformation

in cooperation with:

CzechGlobe and EODC

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

This document provides the Algorithm Theoretical Basis Document of the project YIPEEO (ATBDv1).

Number of pages: 40

Authors:	Emanuel Büechi (EB), Felix Reuß (FR), Lenka Bartošová (LB), Milan Fischer (MF), Lucie Homolová (LH), Miroslav Pikl (MP), Wouter Dorigo (WD)		
Circulation (internal):	Project consortium		
External:	ESA		
Issue	Date	Details	Editor
0.1	12.1.2024	First draft	EB
0.2	19.1.2024	Adding Chapters 1 and 3	EB
0.3	29.1.2024	Adding Chapter 2	FR
1.0	31.1.2024	Final adjustments	EB
2.0	22.4.2024	Including feedback from ESA	EB

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

Table of content

List of figures

List of tables

Acronyms

ANN	Artificial Neural Network
ATBD	Algorithm Theoretical Basis Document
CR	Cross-ratio
DEM	Digital Elevation Model
EO	Earth Observation
EVI	Enhanced Vegetation Index
HP	Hyperparameter
LST	Land Surface Temperature
LT	Lead-time
NDVI	Normalized Difference Vegetation Index
NDWI	Normalized Difference Water Index
NMDI	Normalized Multi-band Drought Index
PVR	Product Validation Report
RB	Requirement Baseline Document
RF	Random Forest
RFE	Recursive Feature Elimination
S1 (2)	Sentinel-1 / 2
SCL	Scene Classification Layer
SSM	Surface Soil Moisture
VH	Vertical-horizontal polarized

VV	Vertical-vertical polarized
XGB	Extreme Gradient Boosting

Executive summary

This document constitutes the Algorithm Theoretical Basis Document (ATBD) as a result of task 3 described in the proposal of the ESA YIPEEO project. This document contains the theoretical basis of the models used during task 3:

- 3.1 Sensitivity of remote sensing observations to yield
  - Assessment of the sensitivity to field management practices (e.g., ploughing)
  - Assessing the dependence of the sensitivity to crop yield on the crop type, weather conditions, the time-of-season, etc.
- 3.2 Crop yield forecast generation using different input data from D2.1 and different models
  - Extreme Gradient Boosting, Random Forest
  - Process-based models for Polkovice (CZR)

The ATBD document is divided into three chapters describing data preparation, tasks 3.1, and 3.2. Task 3.3 is described in D3.2 PVR.

1 Data

1.1 Crop yield data

The available crop yield data is described in D2.2: Database description. Since there were some updates since the deliverable of D2.2, we are going to briefly describe the used crop yield data here. Most crop yield data is available in the Czech Republic (Fig. 1). There are three crops with more than 250 observations (maize, spring barley, and winter wheat). Of these three crops, only winter wheat data was available for all other countries too. As the amount of training data is key for machine learning, we decided to focus on the three crops with the most data. Maize and spring barley were trained and tested in Czechia only. Winter wheat, on the other hand, trained in Czechia and tested in the all regions (more details about the testing and training in Chapter 3.2.4).

In addition, we have just received the field-scale data from the Spain fields. This contains a total number of 11587 observations, among others of winter barley (5210 observations), winter wheat (3071), and maize (2185). Unfortunately, the data was obtained too late to include it in the modelling for this deliverable but will be used for the upcoming tasks.

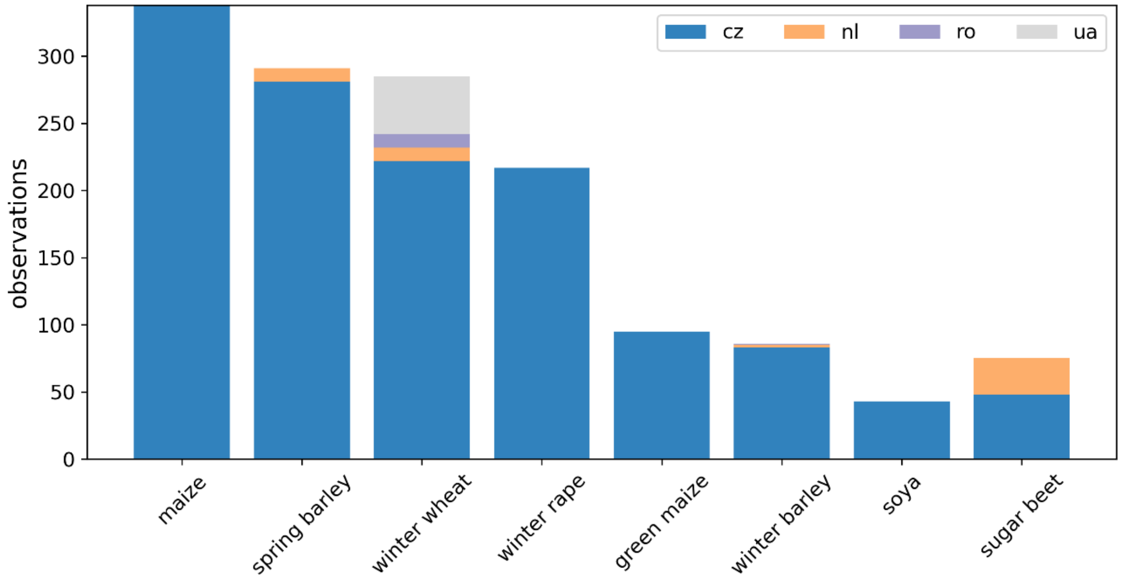

Fig. 1: Data availability per country of all crops with more than 50 observations.

1.2 Predictor datasets

The data preparation of the predictors followed a similar workflow for all datasets (Fig. 2). The datasets were accessed from a datacube (Sentinel-1 data from the EODC datacube, Sentinel-2 from Microsoft's Planetary Computer and ECOSTRESS from NASA's Earthdata). The data was extracted and resampled to the field polygons of the fields where we have crop data. Per field polygon and time step, we calculated the median and the standard deviation of all satellite observations lying within it. Hence, for each field we obtained several time-series per dataset (e.g. for Sentinel-2 L2A one per band). The considered timespan was 2016-2022. To get more reliable data, these time-series required post-processing which will be described in the following subchapters. As a last step before the modelling, the predictors were temporally resampled to monthly and biweekly observations. Starting from the harvest date the first lead time (LT1) included the last month or two-week period before the harvest date. LT2 then includes the month or two-week period before that and so on until 4 months before the harvest.

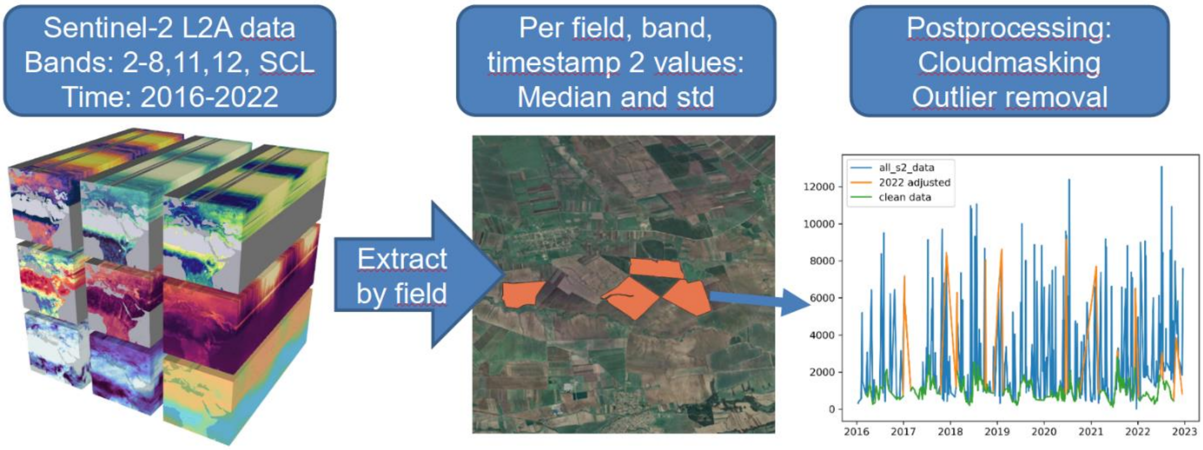

Fig. 2: Workflow of the data extraction on the example of Sentinel-2 data. First image on the left from Mahecha et al. (2020)

1.2.1 Sentinel-1

Sentinel-1 data for the years 2016-2023 was pre-processed by TUW RS 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 CopernicusDEM (Digital Elevation Model) 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, similar to (Bauer-Marschallinger et al., 2021). 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.

1.2.2 ***Sentinel-2***

Sentinel-2 L2A data was accessed from the Microsoft Planetary Computer. Data was retrieved for the bands 2-8, 11, 12, and the scene classification layer (SCL). Bands 1, 9, and 10 were not considered, as they are not required for the calculation of the used vegetation indices (see below), have a lower spatial resolution (60m compared to 10 and 20 for the remaining bands) (Son et al., 2022), and are affected by atmospheric disturbances caused by aerosols and water vapour (Perich et al., 2023). The bands were directly extracted on their native resolution. We used a total of 485 Sentinel-2 images in the Czech sites, 274 in the Netherlands, 307 in Romania, and 900 for Ukraine. The extracted data required two main steps for the post-processing: cloud masking and adjusting for the new dynamic range shifting since January 2022. The cloud masking was done using the SCL. The SCL classifies all Sentinel-2 L2A images into different classes from clouds, over cloud and topographic shadows to vegetation (see Fig. 3). Only pixels which are classified as vegetation, non-vegetated, and water were used for further processing. All other pixels were masked. Pixels classified as water were included in the analysis to not lose information on fields that are extensively irrigated (Zhen et al., 2023).

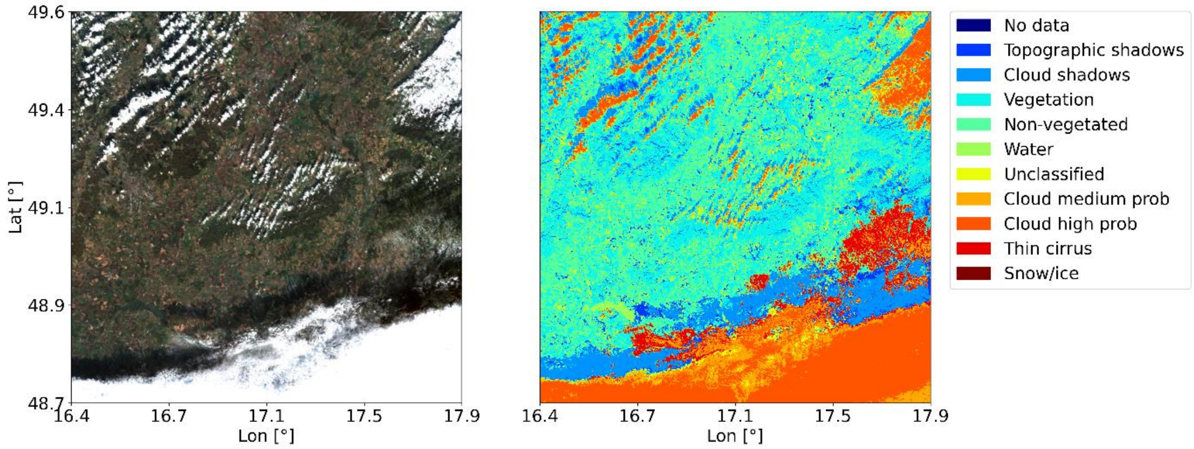

Fig. 3: Example of a Sentinel-2 L2A image (true colour composite on the left) over the Rostenice site and its scene classification on the right.

The dynamic range shifting since 25 January 2022 led to a higher offset in the data. From all observations since then an offset of 1000 was subtracted from the observations to have it in the same value range as before (Siesto 2022). As a last step, outliers were removed, as there were still some unrealistically high values observed after cloud masking (Fig. 4).

Derived from the individual bands we calculated four indices: Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI), Normalized Difference Water Index (NDWI), and the Normalized Multiband Drought Index (NMDI). The used formulas for those are:

NDVI = (B8-B4)/(B8+B4)

EVI = 2.5*(B8-B4)/((B8+6*B4-7.5*B2)+1)

NDWI = (B8-B12)/(B8+B12)

NMDI = (B8-(B11-B12))/(B8+(B11-B12))

More complex estimations like deriving biomass or leaf area index from the S2 bands were not yet included in this part of the project due to time constraints. However, for the upcoming tasks this could be considered.

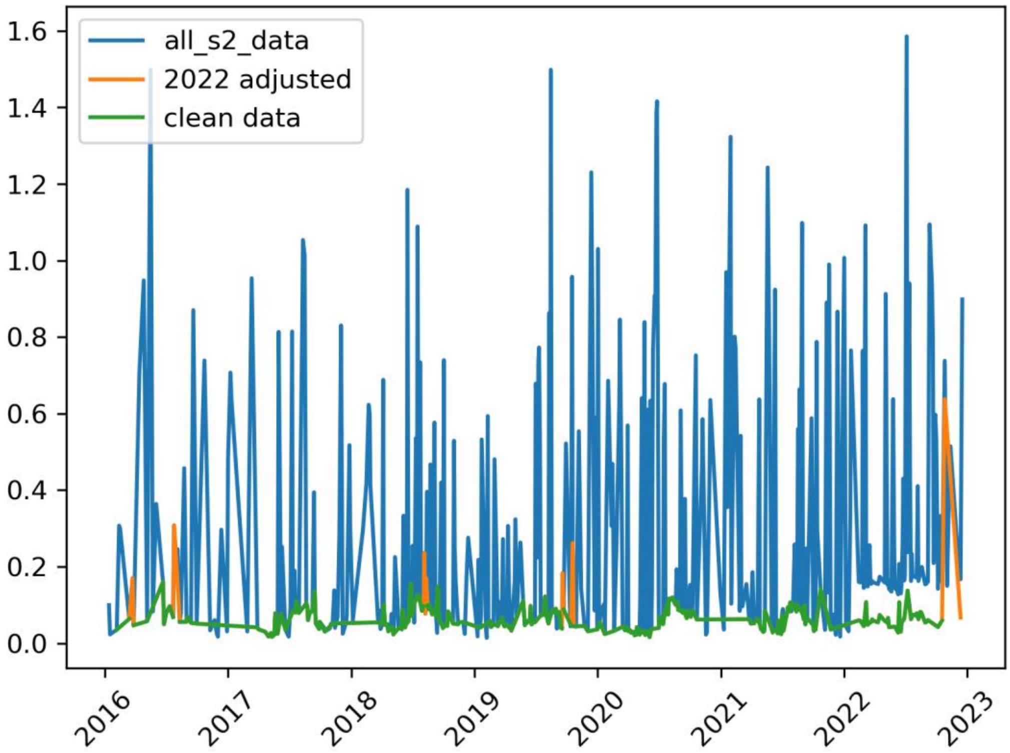

Fig. 4: Example of a Sentinel-2 L2A Band 2 time-series for a selected field of the Rostenice farm in Czech Republic. It shows the originally extracted time-series in blue, the cloud masked data in orange, and the final data in green.

1.2.3 ECOSTRESS

ECOSTRESS land surface temperature (LST) data, its uncertainties, and cloud cover layer were accessed from NASA´s Earthdata portal for the fields of Rostenice farm only. Data is available since June 2018 (Fisher et al., 2020). To start with a full growing season, data was extracted starting in January 2019. The temporal resolution of the data is varying. For many days there are several observations available, but sometimes there are only observations every few days. In total, we used 1481 observations between January 2019 and December 2022, hence, on average one observation per day. This data was still very noisy (Fig. 4) as many observations are not usable because of cloud cover or other disturbances. In addition, the observation time varies a lot. There are observations at every time of the day and the night. This makes it hard to compare several days with each other due to the diurnal temperature cycle. To obtain a less noisy dataset we decided to use various steps of post-processing: excluding unrealistic values (values below -20°C and above 40°C), checking the cloud mask of ECOSTRESS, and only use data from a similar time of the day (11 am to 5 pm). The resulting time-series is shown in Fig. 4 (right plot). The data still seemed to have some issues like abrupt changes of the temperature within some days and temperatures below 0°C in summer. This could be related to cloud cover. Hence, we used the ECOSTRESS cloud cover mask (ECO2CLD). This provides a binary flag if the observations are likely affected by clouds. However, the cloud cover mask did not help to improve this. The ECOSTRESS cloud mask is on for most of the observations, which would leave us with almost no data (Fig. 5).

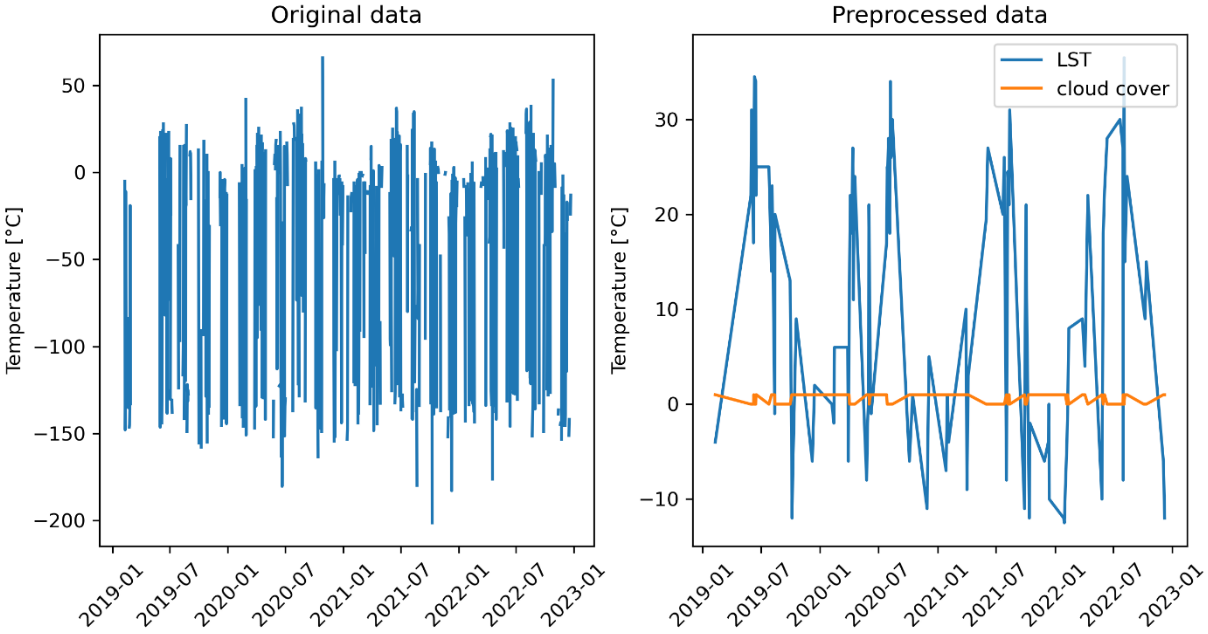

Fig. 5: ECOSTRESS land surface temperature time-series for a selected field of the Rostenice farm in Czech Republic. The left plot shows all observations while the right plot shows the cleaned time-series and the cloud flag in orange.

1.2.4 Further dataset

The last predictor we used for the field-scale crop yield forecasts is the crop type of the antecedent year. This information is included since the soil quality is impacted by the grown crop type which can affect the crop yields of the upcoming year (Gebeltová et al., 2020). For most available fields the crop type of the antecedent year was not defined in our database, though.

In the YIPEEO database are several further datasets, which have not been used for this task. The meteorological data ERA5-Land (spatial resolution of 0.1°) and C3S Seasonal forecasts (1°) were not used for this task because of the rather coarse spatial resolution. The same applies to Sentinel-3 SLSTR LST (0.01°) and Sentinel-5 TROPOMI Solar-induced chlorophyll fluorescence (~0.05°). These datasets will be used for the upscaled forecasts on a regional scale. PRISMA and EnMAP have not yet been obtained but will be updated in the modelling when we get the data.

2 Sensitivity of remote sensing observations to yield

2.1 Sensitivity of EO variables to field management practices

In a first step, the sensitivity of EO variables to field management practices was assessed. The basis of the analysis is a comprehensive data record of exact dates of field management practices including different types of ploughing (spring ploughing, vibroflexen, winter ploughing), as well as manure events. The different types of ploughing are hereby connected to different agricultural machinery and result in different soil conditions on the fields. The data set includes field management events from multiple fields in the Netherlands for the time period 2016 - 2023. In addition to the reference data, S1 and S2 time series as well as data from a weather station was used. For Sentinel-2 however, the number of observations was too low to detect short term dynamics. On average, only 12 cloud-free observations per year were available, leading to high temporal distances between the closest observation before and after a field management event. For this reason, we excluded Sentinel-2 data from our analysis.

First, the signal changes after field management events were extracted from the S1 time series by calculating the difference between the closest observation before the event and the closest observation after the event. Figure 8 shows a scatter plot of the differences in VV and VH polarized backscatter for the four different field management practices. As the scatter plot indicates, winter ploughing has the most distinct pattern in Sentinel-1 time series with an average increase of ~2.8dB in VH and 2.3 dB in VV. In contrast, no consistent pattern is observed for spring ploughing with some fields showing an increase in VV and VH after ploughing and some a decrease. Vibroflexen and manure are associated with no significant changes and show in most cases a minor increase of around 1dB in VV and VH. In almost all cases, the increase in VH polarized backscatter is more profound than for VV polarized backscatter. To better understand the differences between spring and winter ploughing, daily temperature and precipitation data from a weather station close to the fields was added to the analysis (data source: https://catalogue.ceda.ac.uk/uuid/5a94b3130e5e49b2a5a69afdc5493a4f). These data were plotted next to the backscatter time series. Figure 6 shows the backscatter time series and the daily temperature for a spring barley field in the 2022. As the plot illustrates, at the time of winter ploughing the temperatures drop below 0° causing a freezing of the soil and a subsequent decrease of backscatter. Tilling in combination with warmer temperatures results afterwards in significant increase of VV and VH backscatter. The same can be observed for the other winter ploughing events in 2022. This indicates that the soil state is a major co-founding factor for the sensitivity of backscatter to farming practices. In Figure 7, the precipitation data is plotted next to the backscatter time series. The precipitation sums indicate that the backscatter dynamics in the beginning of April are most likely related to changes in the soil moisture. In conclusion, it can be said that Vibroflexen and manure seem to have no significant impact on C-band backscatter. After winter ploughing a significant increase in backscatter can be observed, this is however to some degree related to a freezing and thawing of the soil. A more detailed analysis and additional data is required to understand the factors that cause the inconsistent behaviour of spring ploughing in the time series. Especially pictures before and after ploughing, in situ SSM, and information on the soil type (grain size of the soil particles) are examples to be named here

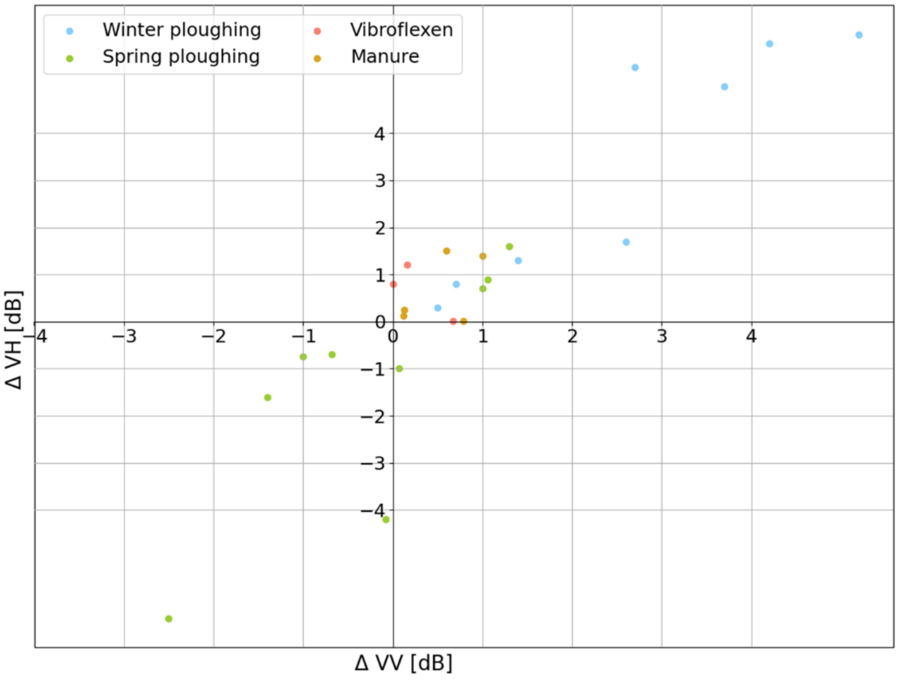

Fig. 6: Scatterplot showing differences in VV and VH polarized backscatter for various field management practices.

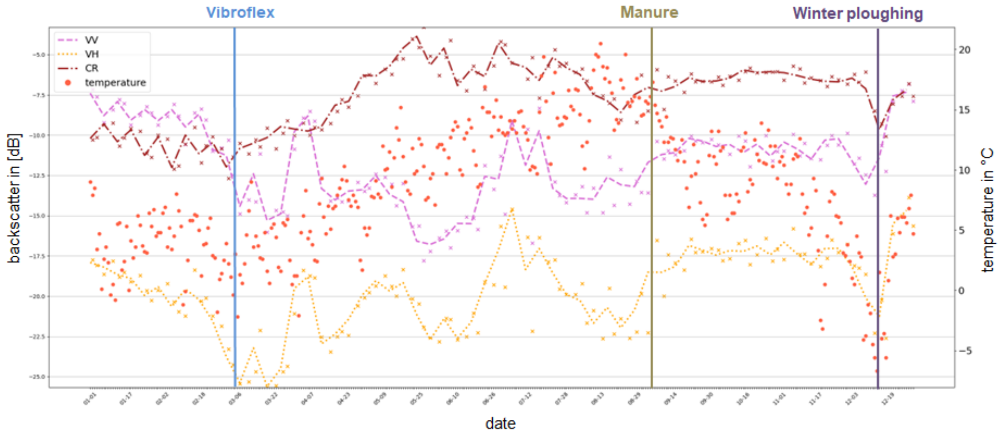

Fig. 7: Sentinel-1 time series for a spring barley field in the Netherlands in 2022. In addition to the SIG0 VV, VH, and CR time series, the field management events, and the daily temperature is illustrated.

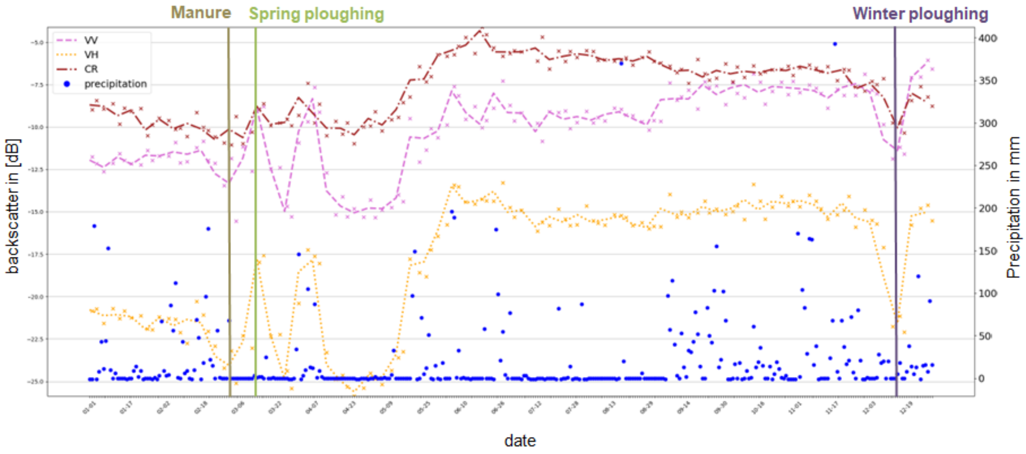

Fig. 8: Sentinel-1 time series plot for a sugar beet field in the Netherlands in 2022. In addition to the SIG0 VV, VH, and CR time series, the field management events, and daily precipitation sums are illustrated.

2.2 Sensitivity of remote sensing observations to yield

In a next step, the sensitivity of EO variables to yield was investigated. For this analysis, the variables from the Sentinel-1 and Sentinel-2 time series were resampled to monthly mean values. These were then correlated to the yield at the end of the season. This analysis was done for all major crops with at least 100 yield measurements and always for the four months before the harvest. To mitigate a potential impact of different climate conditions on the analysis, only data from the Czech Republic was used as this country has the highest number of observations.

2.2.1 Correlation per crop and lead time

Table 1 shows the Pearson R coefficient for all major crops and various EO variables. As the NMDI and NDWI results were very similar, with NDMI achieving slightly higher accuracies, the NDWI results are not included in table 1. Due to the low sensitivity of VV backscatter to vegetation and the low overall correlation, this variable is also not included in the table. As the table indicates, optical indices show overall significantly higher correlations. Depending on the crop type, either the EVI, NDVI, or NMDI shows the highest correlation. In most cases the values for the NDVI and EVI are very similar. Looking at the individual crops, soya has the lowest correlation for all EO variables. For no variable and no single season its correlation coefficient exceeds 0.16. Green Maize has on average the highest correlation for all EO variables. Except for soya the remaining crops have a very similar mean correlation for the optical variables. For the SAR variables, spring barley has the second highest value.

Tab. 1: Pearson R correlation for major crops and various EO variables for the lead times –3 to 0. The highest value per crop is highlighted in green.

	NDVI				EVI				NMDI				VH				CR			
Lead time (month)	-3	-2	-1	0	-3	-2	-1	0	-3	-2	-1	0	-3	-2	-1	0	-3	-2	-1	0
winter wheat	0.21	0.25	0.30	0.24	0.21	0.19	0.22	0.17	0.03	0.09	0.15	0.36	-0.02	-0.04	-0.11	-0.36	-0.03	0.01	0.07	-0.13
green maize	-0.05	0.49	0.45	0.37	-0.10	0.55	0.52	0.29	0.16	0.04	0.52	0.53	0.05	0.46	-0.02	-0.12	0.08	0.19	0.34	0.41
grain maize and corn-cob-mix	0.02	0.43	0.32	0.28	-0.07	0.30	0.41	0.33	0.05	-0.10	0.30	0.28	-0.11	0.23	0.31	0.18	-0.03	0.10	0.09	0.18
spring barley	0.46	0.40	-0.01	0.18	0.48	0.40	0.03	0.13	0.48	0.56	0.19	0.07	-0.22	0.02	0.18	-0.12	0.26	0.06	0.14	0.08
winter barley	0.35	0.23	0.08	0.25	0.23	0.16	0.11	0.17	0.07	-0.06	0.39	0.57	-0.20	-0.34	-0.41	-0.57	-0.16	-0.22	0.04	0.04
soya	-0.21	0.06	0.13	-0.21	-0.16	0.17	0.19	-0.17	-0.30	0.05	0.15	-0.44	-0.14	-0.04	0.07	0.01	-0.12	-0.28	-0.2	0.01

The most significant observation is, that for the majority of crops, the highest correlation is not observed at the month closest to harvest (lead time 0) but already two months before (lead time –2). This can to some degree be explained by the fact, that always all EO observations from a month were included in the analysis. This means, that for the harvest month observations after the harvest were also included. However, also the month before the harvest (lead time –1) shows in many cases lower correlations than two months before the harvest. Here, also significant differences between the optical indices are visible. Whereas the NMDI shows in most cases the highest correlation at month 0 or –1, NDVI and EVI show typically the highest correlation for the months –2 and –1. An explanation for this is that the latter two indices often start to decrease already several weeks before the harvest at the ripening stage of the plant when the crops lose their green colour. As the NDVI and EVI are chlorophyll sensitive their correlation also decreases. A similar trend can be observed for the SAR indices VH and CR when looking at single season correlation values. VH typically reaches the highest correlation during the month –2 when due to the earing and flowering the depolarization of the emitted radar wave is high and VH backscatter increases faster than VV backscatter. CR on the other hand in most cases has the highest correlation for the months –1 and 0 as it is a proxy of the plant biomass (Vreugdenhil et al., 2020)

Tab. 2: Pearson R correlation for green maize and winter barley and various EO variables for the lead times –3 to 0. The highest value per variable is highlighted in green.

Green Maize 2019				
	-3	**-2**	**-1**	**0**
NDVI	**-0.34**	**0.49**	**0.51**	**0.28**
EVI	**-0.33**	**0.73**	**0.72**	**0.49**
NMDI	**0.38**	**-0.03**	**0.41**	**0.50**
VH	**-0.30**	**0.34**	**-0.17**	**-0,52**
CR	**0.07**	**0.14**	**0.46**	**0.70**

Winter barley 2021			
-3	**-2**	**-1**	**0**
0.22	**0.19**	**0.17**	**0.30**
0.01	**0.28**	**0.16**	**0.12**
0.1	**0.14**	**0.51**	**0.74**
-0.43	**-0.66**	**-0.65**	**-0.75**
-0.49	**-0.41**	**-0.32**	**-0.31**

Table 2 shows the Pearson R coefficient for the five EO variables NDVI, EVI, NMDI, VH, and CR for green maize in the year 2022 and winter barley in 2021. Looking at the single year correlation values, the differences between the SAR variables and optical indices are getting less profound. Here, in the case of green maize, CR reaches correlation coefficients similar to the EVI. The tendency of CR and NMDI having the highest correlation at the harvest and the other variables 2-3 months before the harvest can be observed again. For winter barley 2021 the correlations are significantly lower for all variables except for NMDI. The year 2021 observed below average precipitation in the winter months which lead to the lower correlations. The next chapter investigates the impact of drought on the correlation to yield in more detail.

2.2.2 *Impact of droughts on EO variables*

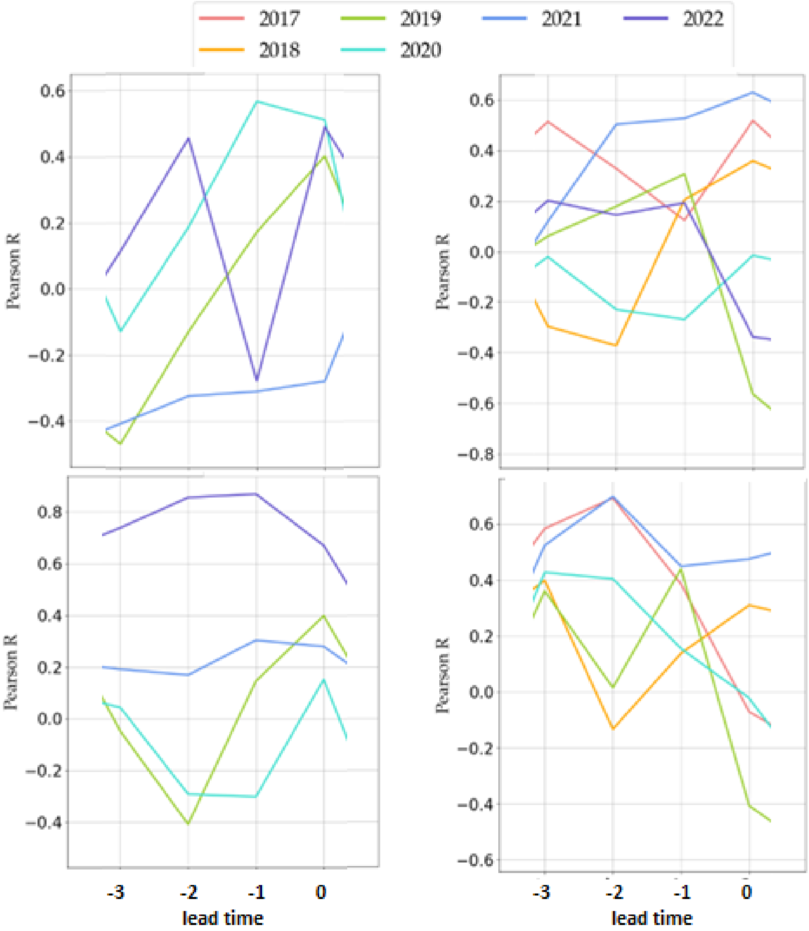

Fig. 9: Pearson R correlation for different years between the yield at the end of the season and CR (upper row) and EVI (lower row) for winter barley fields (left) and grain maize (right) for the 3 months before the harvest.

In a next step, the sensitivity of EO variables to vegetation under drought conditions was assessed. The goal here was to investigate if and which variables are able to observe the impact of droughts on the crops. To do so, in a first step the Pearson R correlation between the EO variables and the yield was compared for all years. Figure 9 shows these correlations for the crops winter barley and grain maize for the variables CR and EVI. As the figure indicates the correlations highly vary between the years, but also between the variables. The latter is especially the case for winter barley. CR has a high correlation for the year 2020, whereas EVI shows very low correlation for this year. The opposite is true for 2021. For this year CR has a strong negative correlation, whereas EVI has slight positive correlation. For grain maize the variables show more similar trends for the individual years. The years 2017 and 2021 have the highest correlation in the months –3 and –2. For the months at harvest time, 2019 has the lowest correlation and 2021 the highest value.

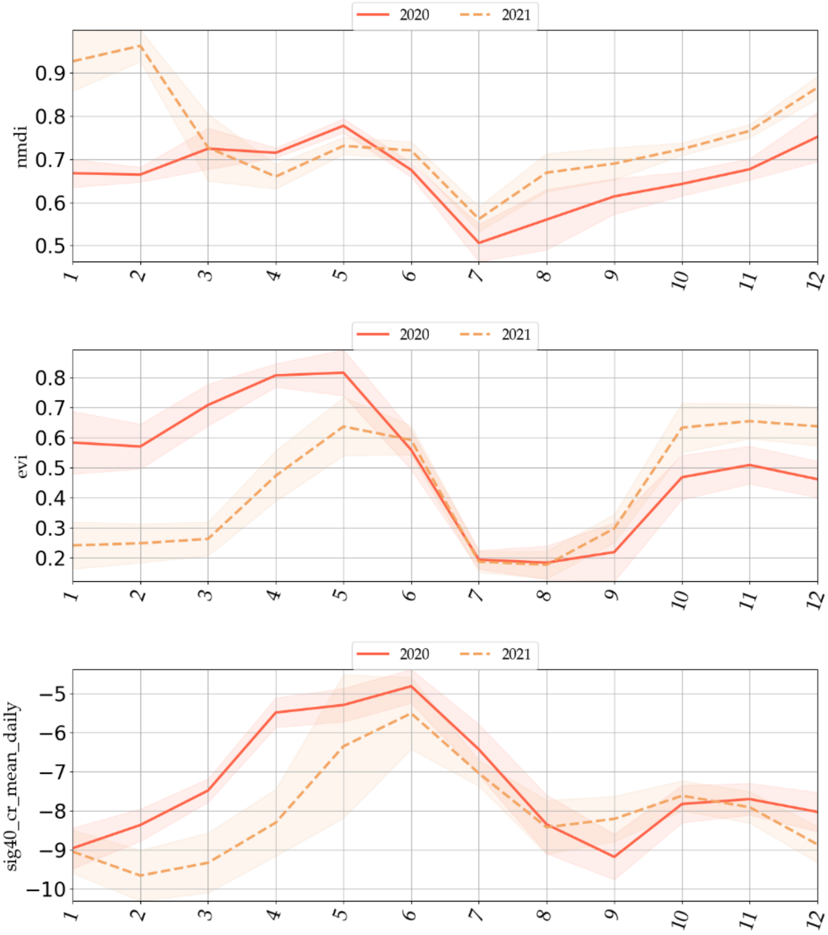

Fig. 10: Time series plots showing average NMDI, EVI and CR time series for winter barley fields in Czech Republic in the years 2020 and 2021. In addition to the mean value, +/- 0.25 standard deviation is plotted.

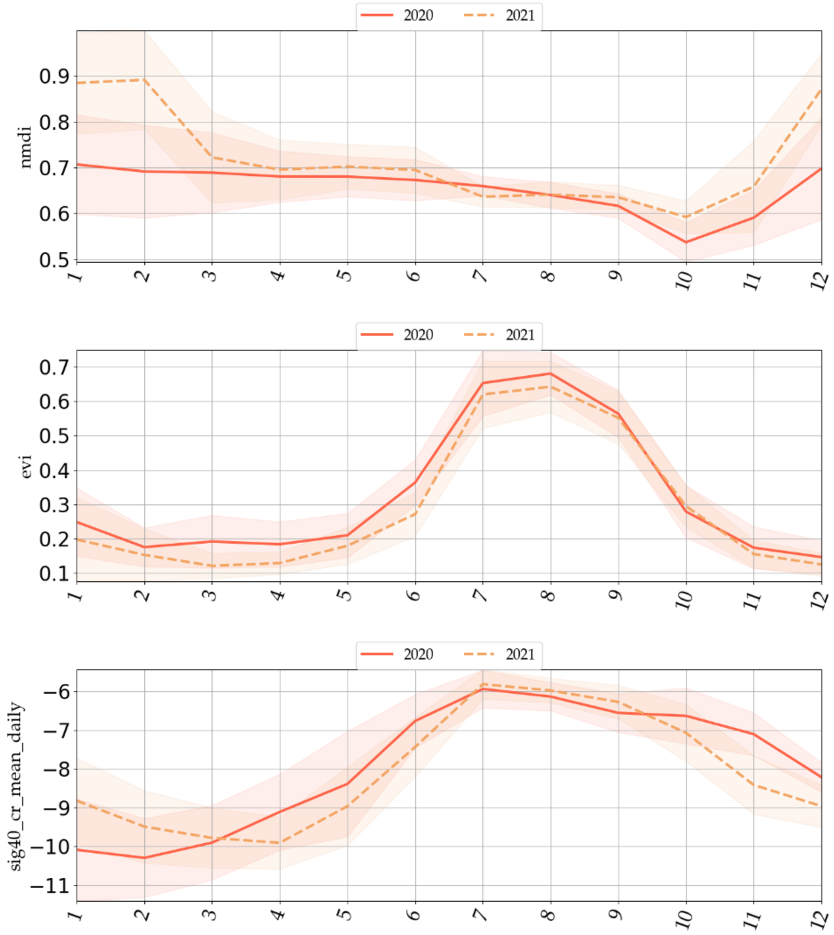

Fig. 11: Time series plots showing average NMDI, EVI and CR time series for grain maize fields in Czech Republic in the years 2020 and 2021. In addition to the mean value, +/- 0.25 standard deviation is plotted.

Afterwards, EO time series for years with and without drought conditions were investigated in detail. The years 2020 and 2021 were selected, as 2020 had normal to high precipitation in the winter months whereas 2021 observed above average precipitation in winter in the Czech Republic. And inverse trend was observed in summer with a severe drought in 2020 but normal to high precipitation in 2021. Figure 9 shows the comparison of winter barley time series for the year 2020 and 2021 for the CR, the NMDI and EVI. The NMDI time series for the year 2020 is higher in the first months of the year but decreases below the level of 2021 in April. This is most likely due to the fact, that the NMDI is sensitive to both the soil moisture and the vegetation water content but in an inverse behaviour. NMDI is high for low soil moisture, but low for high vegetation water content. With the growth of the winter barley plants, they start to dominate the signal in March/April leading to the decrease of the NMDI.

CR and the EVI both seems to capture the lower plant biomass in 2021 with the difference, that EVI is decreasing one month earlier than CR. This can be explained by the fact, that EVI is more sensitive to the chlorophyl content and CR to the plant structure. EVI thus already decreases before the harvest with ripening of the plant but CR after the harvest.

Figure 10 shows the same variables and years but for the crop grain maize. Despite the different precipitation conditions during spring/summer, the time series for the two years are very similar for both variables. If we look at the correlation for the two years in figure 9 we can see however that the correlation in the drought year 2020 is significantly lower compared to 2021. The reason for the lower correlation lies thus not in the EO variables but in the lower yield values for this year.

In summary, the results showed significant differences in the correlation of earth observation variables to crop yield and also in the visibility of droughts in EO time series. These can be partly explained by the sensitivity to different crop characteristics. This also leads to the differences between the EO variables in the correlation between lead times and the yield. Based on these facts, a combination of Sentinel-1 and Sentinel-2 promises the greatest accuracy in crop yield forecasts.

3 Crop yield forecast generation

3.1 Crop models

We used the Hermes model for yield prediction. The model is validated with the data observed in small-plot experiments at the Polkovice site. Model validation will be performed for winter wheat, spring barley, winter rape, and maize.

The HERMES model belongs among the widely used, easily accessible and well-documented crop growth simulation models (e.g. Palosuo et al., 2011). It is a process-oriented model. The benefit of using HERMES is the ability to work with a relatively small amount of input data sets that are ordinarily available at the farm level and that take into consideration plant growth, N-uptake, the process of net mineralization, the denitrification and transport of water and nitrate (Kersebaum et al., 2011).

The input data were divided into the following three parts: weather data, soil information and management data. Individual parameters entered into the model were obtained from soil and meteorological measurements including data about global solar radiation, air temperature (average, minimum and maximum), air humidity, wind speed, precipitation and tillage. Further, data of harvest, pre-crop and initial conditions were used to launch the model. These data were acquired from the Polkovice experimental station for the period 2018–2023.

3.2 Machine learning

The focus of YIPEEO is to generate crop yield forecasts on a field-level using Earth observation data and machine learning. As shown in D1.1: Requirement Baseline, there are several models that can be used for this task. Here, we focused on Random Forest (RF) and Extreme Gradient Boosting (XGB). The next step will be to also try deep learning models (Long-short term memory) with the Spain data.

3.2.1 Model setup

We established four crop yield forecasts per crop for different times before the harvest: four months to one month before the harvest (from now on called lead times 4 to 1: LT4 to LT1) (Fig. 12). For this, we temporally resampled the predictors to a monthly resolution. LT1 included the data of the last month proceeding harvest, LT2 the second to last month, and so on. Based on this temporal resampling we calculated the crop yield forecasts starting with LT4. This only included data of the fourth month before harvest of all predictors. The forecast of LT3 was then trained with the predictors of LT4 plus the ones of LT3. As predictors we used all Sentinel-1 data, i.e., Sig0 VV, VH, and CR, and the same for Sig40; from Sentinel-2 we used only the calculated indices (NDVI, EVI, NMDI, NDWI) and not the individual bands themselves, to not get a too large number of predictors. With these datasets from Sentinel-1 and 2, we already had 10 predictors per LT. As the information of the months before was always included, this results in 40 predictors for LT1 (e.g. NDVI of LT1, LT2, LT3, and LT4, respectively). For some tests, we included the additional predictors of ECOSTRESS LST and the precedent crop type.

The temporal resolution of one month is rather coarse. In practice, a higher temporal resolution could be of much benefit for farmers and decision-makers. This would lead to more regular updates and at the same time a first reliable forecast could be available earlier. Therefore, we tested a resampling to biweekly observations too. The model setup remained the same, but instead of four monthly timesteps, this led to 8 biweekly timesteps (LT2 to LT16 for 2 to 16 weeks). The disadvantage of this approach is that there are some regions with no cloudfree Sentinel-2 data with which to calculate the updated crop yield forecasts. Hence, we only did some test for biweekly forecasts, but did the main validation for a temporal resolution of one month.

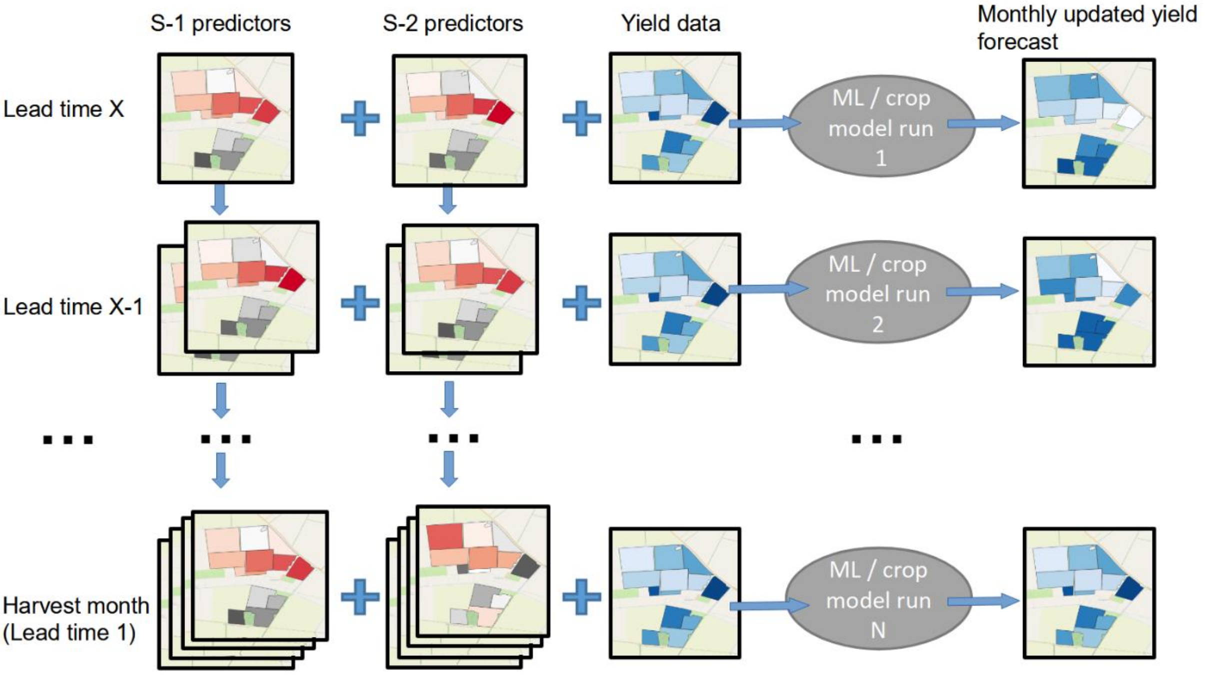

Fig. 12: Model setup of the machine learning algorithms.

Another important part of the model setup was to split the data into train, validation, and test data. As defined in D1.1 RB we will use approximately 60% for training, 20% for validation (i.e. for the optimization of the model for example using hyperparameter tuning) and the remaining 20% for testing only. Hence, all the results showed in this report are based on the training and validation data, and D3.2 PVR will show the testing using the last 20% of data that are not used here. The test data was selected in three ways:

1) Temporal leave-out: this will be done using a leave-one-year-out-cross-validation. Each year from 2016 to 2022 will be used once for testing only. The remaining years are then used for training and validation. This can be considered as the most realistic validation. An operational crop yield forecast, can as well only be trained with data from the previous years, and does not have any information about the crop yields of the forecasted year. This was only applied to data from Rostenice farm.
2) Spatial split: most crop yield data is available for the Rostenice farm in Czechia. For most other countries, there is not enough data to train and test a machine learning model. Hence, we tried to train and validate the model only using the data from

Rostenice, and test it on all the other countries and farms. The data from Polkovice (Czechia), Netherlands, and Ukraine were each used as test site separately.

3) Random split: the data from Rostenice is randomly split into test-train-validation data.

3.2.2 Predictor analysis

Before the modelling started, we performed an exploratory data analysis to compare the predictors to each other. The sensitivity of EO data to crop yields is one part of this (Chapter 2.2). In addition, it is as well important to know how the predictors correlate with each other, as predictors with multi-collinearity can negatively impact the performance of a machine learning model (Chan et al., 2022). In the used predictors, there are two possible ways of cross-correlations: between the predictors (e.g. NDVI to EVI) and temporal ones (e.g. predictors in LT3 and LT4). In Fig. 13, we show the cross-correlations of all predictors in June (LT1 for winter wheat). It shows an expected pattern. The predictors from Sentinel-1 have large positive cross-correlations to each other, as well as the ones from Sentinel-2. Other than that, the cross-correlations are from around –0.4 to 0.4 (Fig. 13). The temporal correlations are shown on the example of Sig40 VH and NDVI (Fig. 14). It shows that the temporal correlations are overall low. There are only some correlations around 0.6 for NDVI of LT3 to LT4 and LT2 to LT3, as well as for Sig40 VH between LT4 to LT2 and LT3.

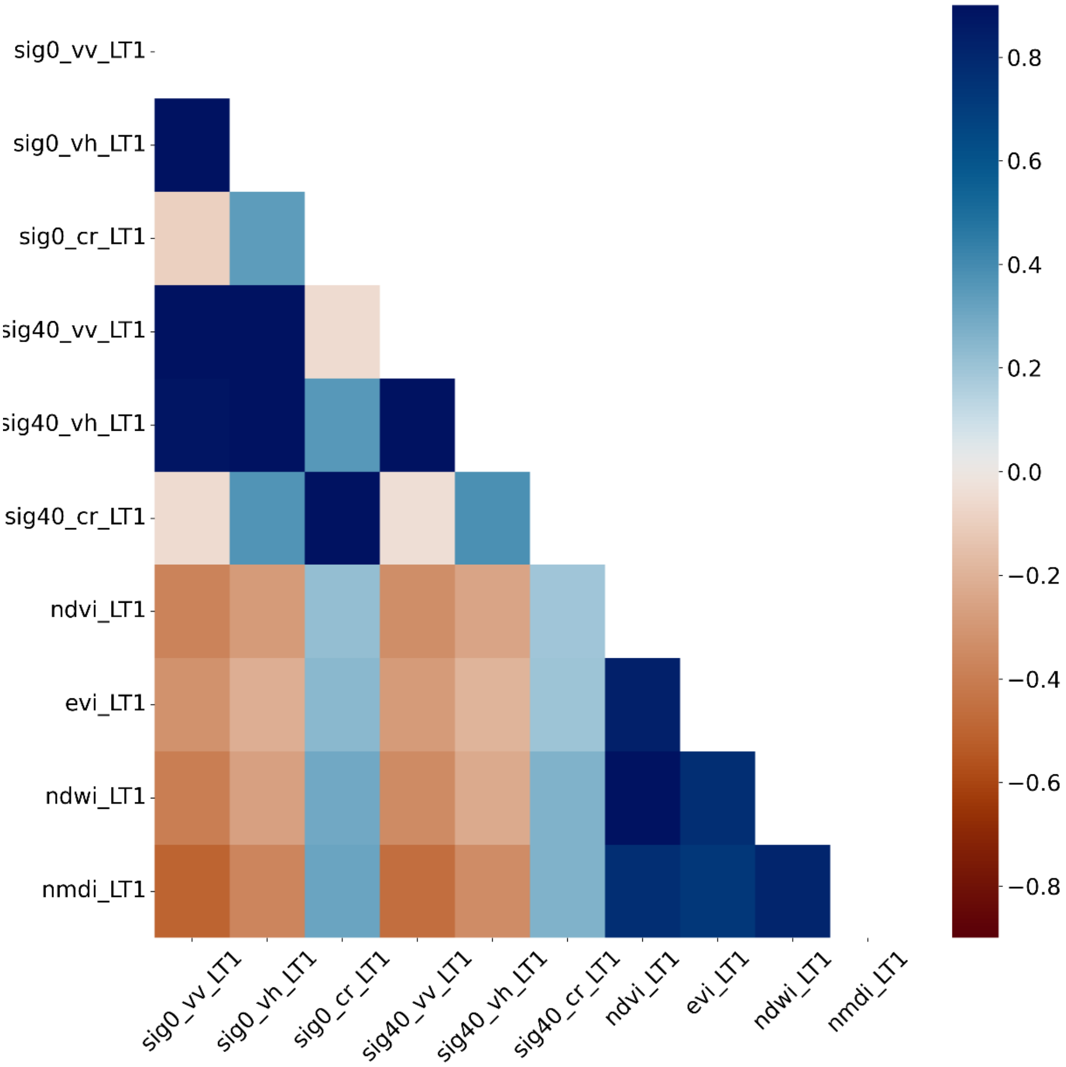

Fig. 13: Cross-correlations between the predictors at LT1.

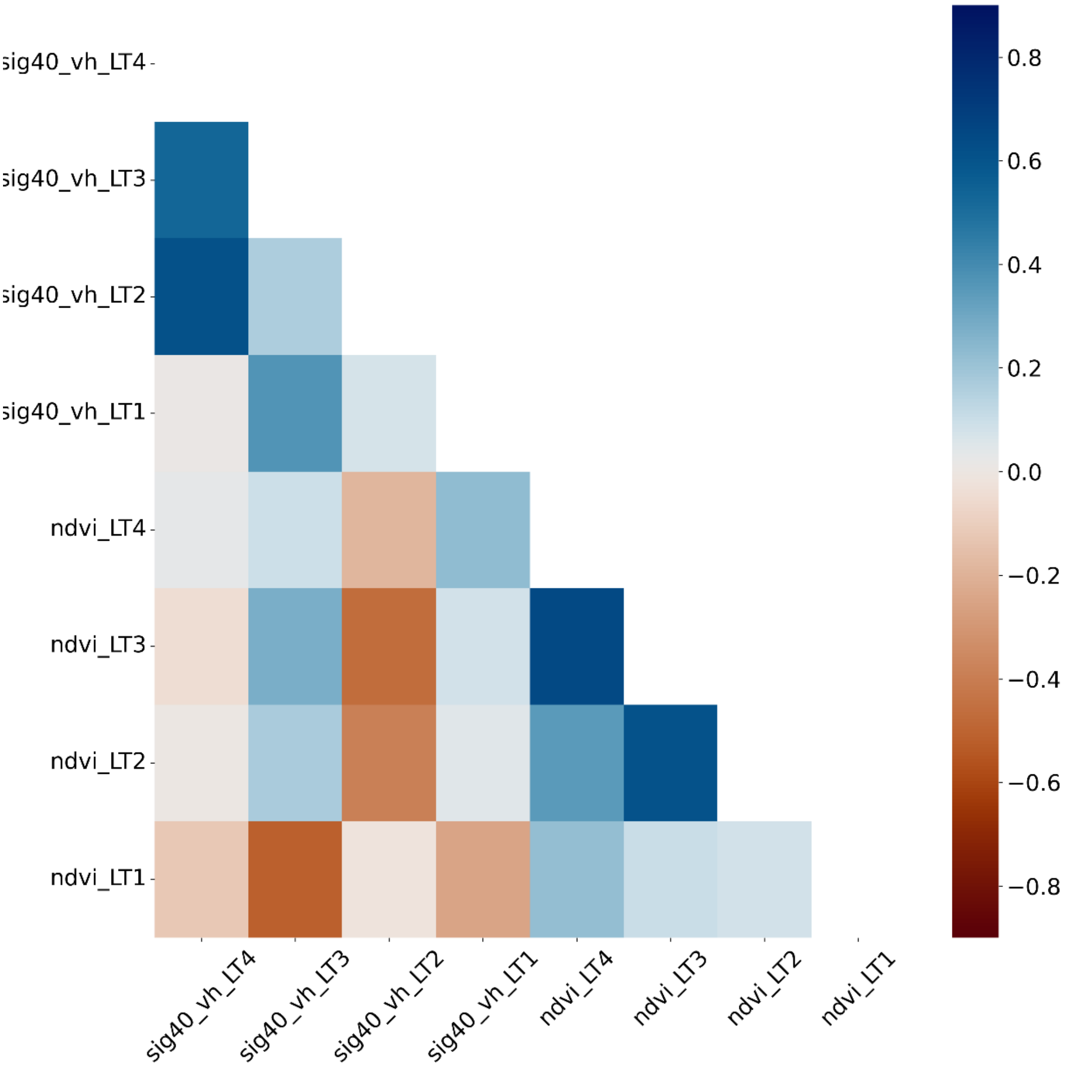

Fig. 14: Temporal cross-correlations between NDVI and Sig40 VH for the different lead times.

*3.2.3 **Model optimization***

Hyperparameter-tuning (HP-tuning) and feature selection are key for any machine learning application. The model performance can be improved using HP-tuning, while feature selection helps to reduce the number of used predictors. The latter is not only helpful to address the issue of multi-collinearity of predictors but also to obtain a sparse model that is more efficient and easier interpretable (Binder et al., 2020).

Features were selected firstly while only using default HP settings. The feature selection technique is based on recursive feature elimination. I.e., first all predictors are used, and then different predictors are removed to see how the performance changes. This was repeated until only 10 predictors were left. For this, we used the Python tool RFE from *sklearn* (Pedregosa et al., 2011). The feature selection could be further improved by testing different numbers of predictors. This has not yet been done, instead we just used a maximum number of 10 predictors.

After having obtained the most important predictors, the HPs were tuned. For this we used different values for the most common HP of XGB and Random Forests. These are the maximum tree depth, minimum sample split of the branches, and number of trees. (Probst et al., 2019).

A last step of model optimization we used, was to add further predictors. The model runs using only Sentinel-1 and Sentinel-2 data show an obvious deficit of representing temperature as a key variable of crop yield forecasting (Bueechi et al., 2023). Therefore, we tested, if the forecasts can be further improved using ECOSTRESS land surface temperature. The disadvantage of that is, that we lose 3 years of crop data since ECOSTRESS data is only available since mid-2018. At the same time, we tested the dataset of the antecedent crop type.

3.2.4 ***Model validation***

The model validation using Sentinel-1 and 2 data showed a clear advantage of combining the information of these two satellites. The model performance is almost always better when using both compared to only Sentinel-1 or 2 data (Fig. 15 to 17), except for LT1 for winter wheat (Fig. 15) and LT2 for spring barley (Fig. 17), where only Sentinel-2 data leads to the best results. The information from the two satellites complement each other well. While Sentinel-1 data adds much value in the long-term forecasts (LT3 and LT4) provides Sentinel-2 data good results in the last 2 to 3 months before harvest. This seems reasonable as the crops may not be evolved enough 4 months before harvest to give a clear signal in optical data. Sentinel-1 data which is more sensitive to soil moisture can already provide important information about the water availability of the crops at that stage. This is in agreement with the exploratory data analysis from Chapter 2, where we concluded that combining S1 and 2 data will most likely help to get more reliable forecasts.

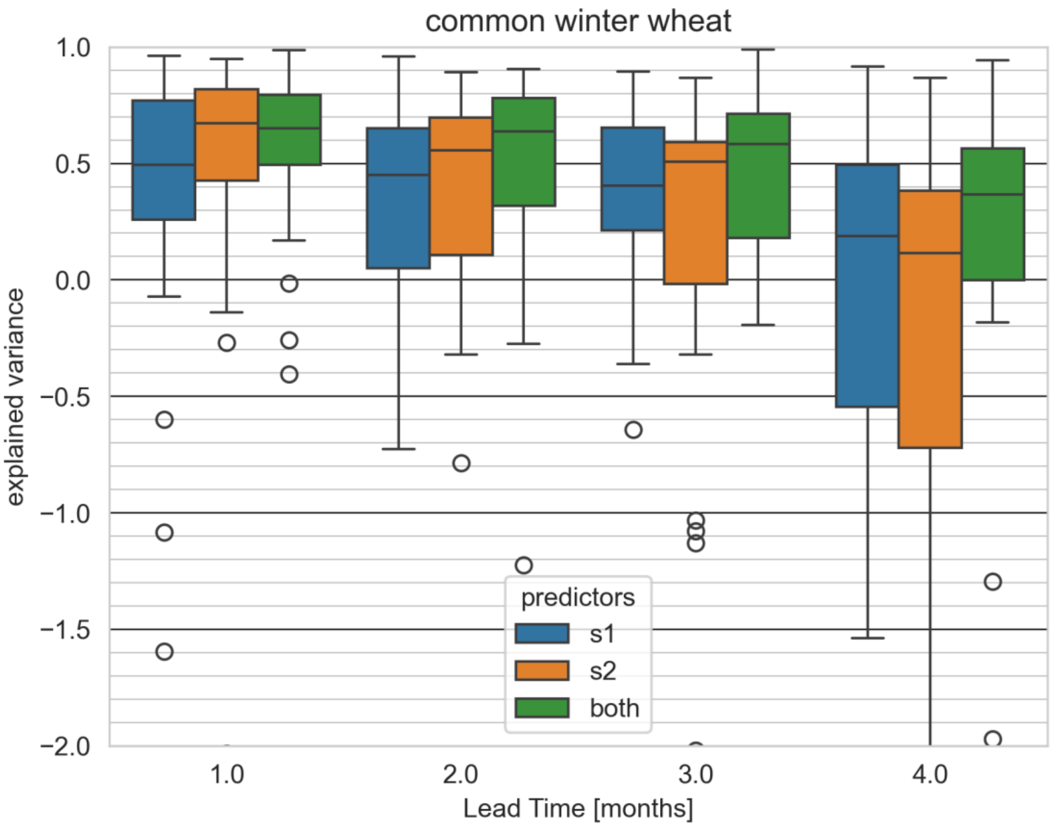

Fig. 15: Results of the cross-validation of the winter wheat forecast for using only Sentinel-1 and Sentinel-2 data compared to the performance when using all data as predictors.

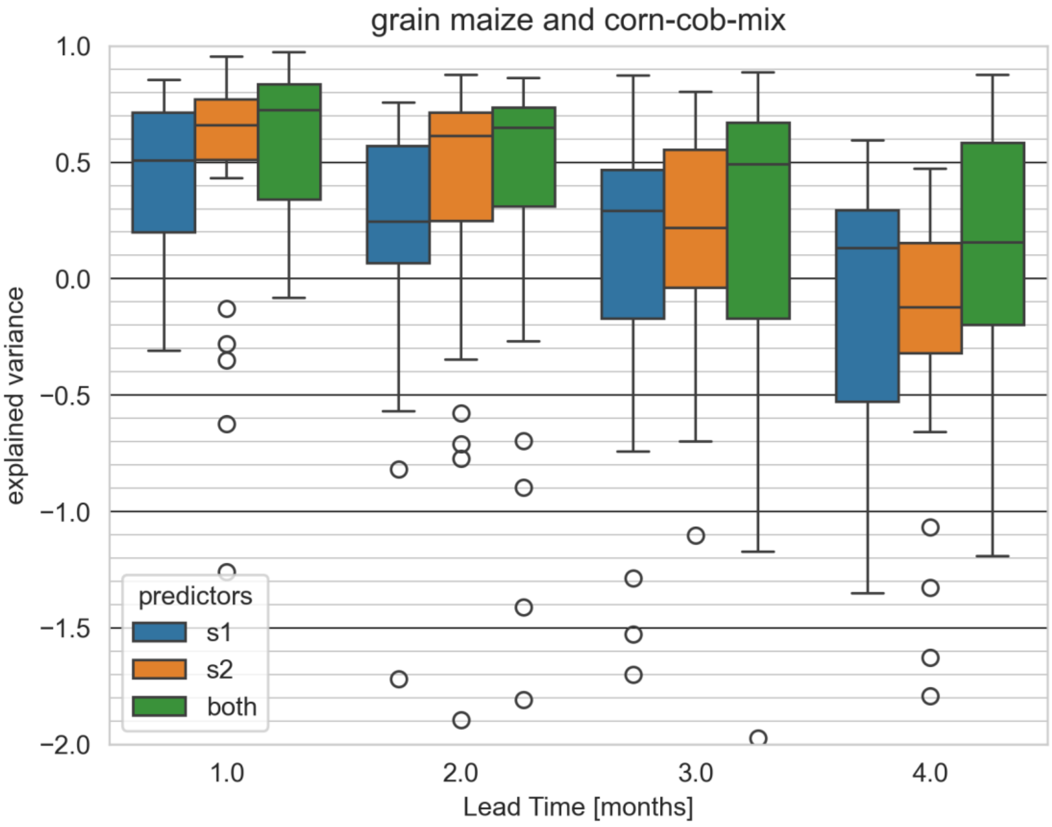

Fig. 16: Same as Fig. 15 for maize.

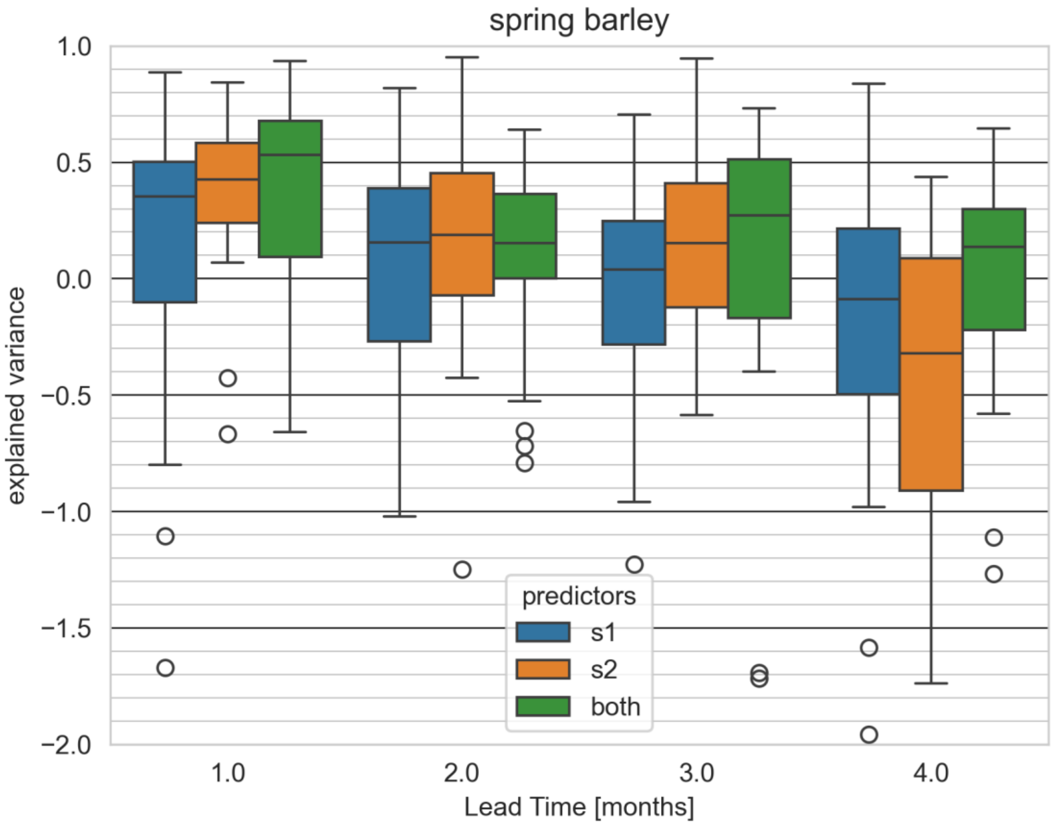

Fig. 17: same as Fig. 15 for spring barley.

Changing the temporal resolution of the predictors from monthly to biweekly leads to similar results. The overall performance is good starting from 14 weeks before the harvest. Hence, two weeks earlier than for the monthly forecasts with 3 months (i.e. 12 weeks). Regarding Sentinel-1 and 2 data, the conclusion is the same, that the winter wheat model based on S1 outperforms the S2 model in the early forecasts until around 12 weeks before the harvest and afterwards, the S2 model is better. A combination of S1 and S2 still shows the best results.

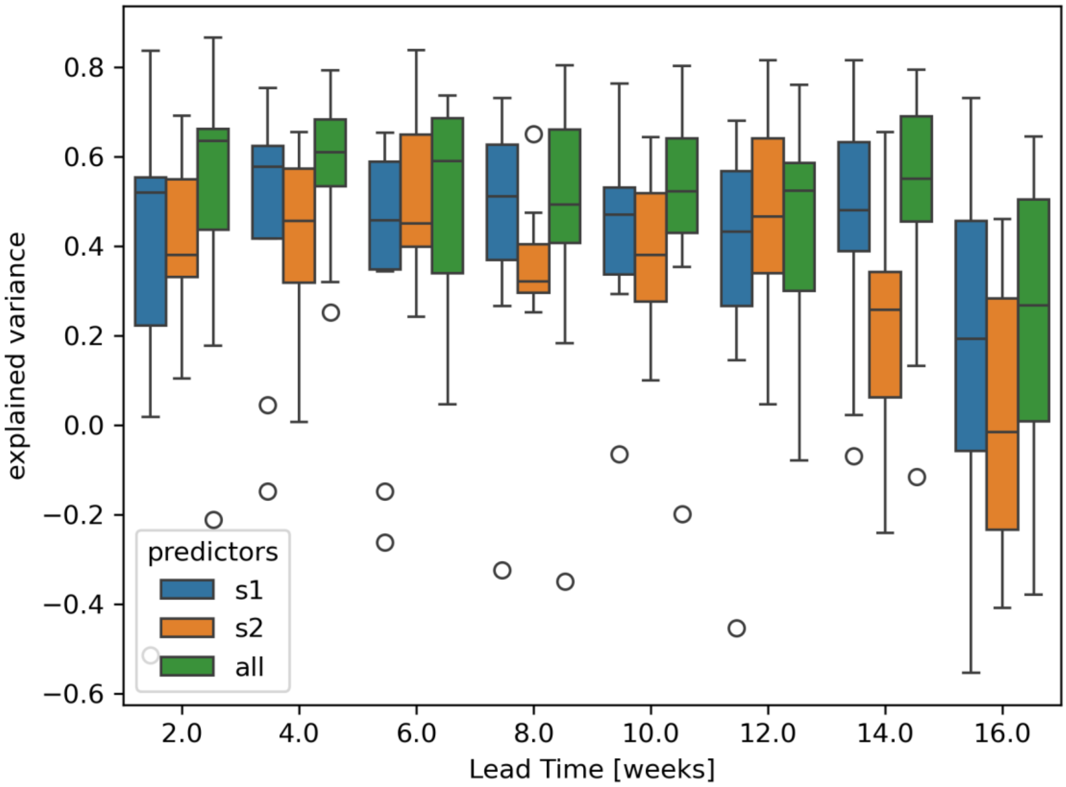

Fig. 18: same as Fig. 15 with using biweekly data instead of monthly values of the predictors

The forecast using ECOSTRESS land surface temperature data performs poorly ($R^2<0.2$). In combination with Sentinel-1 and 2 data, it improves the results significantly (Fig. 19). However, due to the reduced availability of training data (ECOSTRESS only available since mid-2018), the overall performance is lower than when using Sentinel-1 and Sentinel-2 data starting from 2016 (Fig. 15 and Fig. 19). Therefore, we refrained from using ECOSTRESS data for the further modelling in this task.

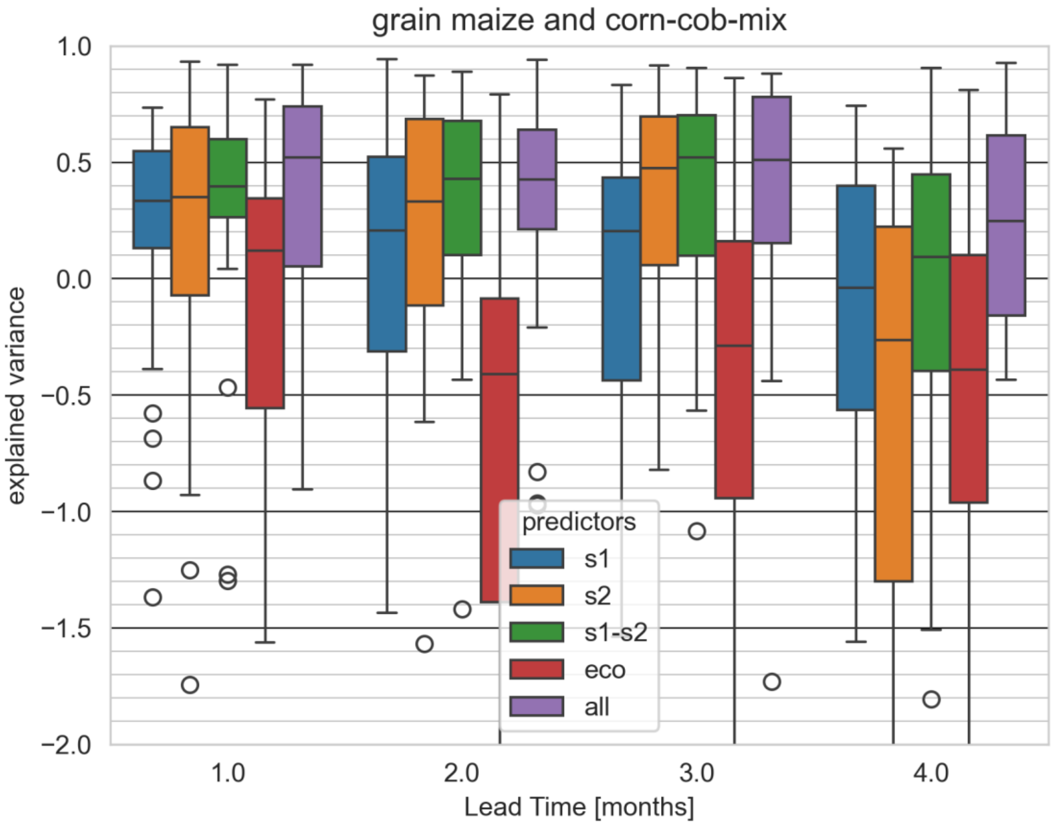

Fig. 19: same as Fig. 15 including ECOSTRESS LST data for maize.

The impact of feature elimination on the model performance is very small (Fig. 20). This is a good sign, as the same model performance can be obtained with fewer predictors. HP-tuning leads to a significantly better performance, especially for LT3. The combined effect of HP-tuning and feature elimination is overall a bit worse than when using only HP-tuning. Still, we decided to use both, due to the mentioned advantages of feature selection.

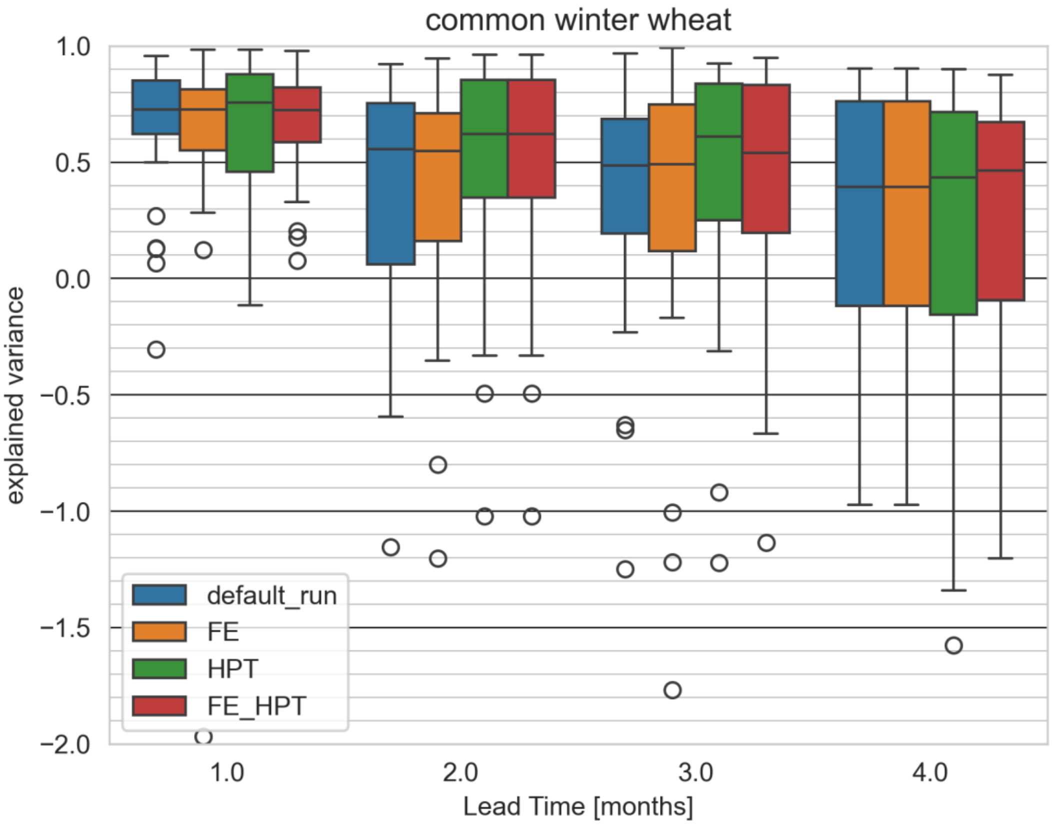

Fig. 20: Comparison of the performance of the default XGB winter wheat forecast to feature elimination, hyperparameter-tuning, and both.

Finally, we decided to use the model based on Sentinel-1 and 2 data only plus HP-tuning and feature elimination for the model runs validated in the PVR.

4 References

Bauer-Marschallinger, B., Cao, S., Navacchi, C. *et al.* The normalised Sentinel-1 Global Backscatter Model, mapping Earth's land surface with C-band microwaves. *Sci Data* **8**, 277 (2021).

Binder, M., Moosbauer, J., Thomas, J., & Bischl, B. (2020). Multi-objective hyperparameter tuning and feature selection using filter ensembles. GECCO 2020 - Proceedings of the 2020 Genetic and Evolutionary Computation Conference, 471–479. https://doi.org/10.1145/3377930.3389815

Bueechi, E., Fischer, M., Crocetti, L., Trnka, M., Grlj, A., Zappa, L., & Dorigo, W. (2023). Crop yield anomaly forecasting in the Pannonian basin using gradient boosting and its performance in years of severe drought. Agricultural and Forest Meteorology, 340, 109596. https://doi.org/10.1016/J.AGRFORMET.2023.109596

Chan, J. Y. Le, Leow, S. M. H., Bea, K. T., Cheng, W. K., Phoong, S. W., Hong, Z. W., & Chen, Y. L. (2022). Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review. Mathematics 2022, Vol. 10, Page 1283, 10(8), 1283. https://doi.org/10.3390/MATH10081283

Fisher, J. B., Lee, B., Purdy, A. J., Halverson, G. H., Dohlen, M. B., Cawse-Nicholson, K., Wang, A., Anderson, R. G., Aragon, B., Arain, M. A., Baldocchi, D. D., Baker, J. M., Barral, H., Bernacchi, C. J., Bernhofer, C., Biraud, S. C., Bohrer, G., Brunsell, N., Cappelaere, B., … Hook, S. (2020). ECOSTRESS: NASA's Next Generation Mission to Measure Evapotranspiration From the International Space Station. Water Resources Research, 56(4), e2019WR026058. https://doi.org/10.1029/2019WR026058

Gebeltová, Z., Malec, K., Maitah, M., Smutka, L., Appiah-Kubi, S. N. K., Maitah, K., Sahatqija, J., & Sirohi, J. (2020). The Impact of Crop Mix on Decreasing Soil Price and Soil Degradation: A Case Study of Selected Regions in Czechia (2002–2019). Sustainability 2020, Vol. 12, Page 444, 12(2), 444. https://doi.org/10.3390/SU12020444

Kersebaum, K. C. (2011). Special Features of the Hermes Model and Additional Procedures for Parametrization, Calibration, Validation, and Applications. In: AHUJA, L. R. and MA, L. (Eds.). Methods of Introducing System Models into Agricultural Research. Madison, Wisconsin, Adv.Agric. Syst. Model. 2. Madison: ASA, CSSA, SSSA, pp. 65–94.

Mahecha, M. D., Gans, F., Brandt, G., Christiansen, R., Cornell, S. E., Fomferra, N., Kraemer, G., Peters, J., Bodesheim, P., Camps-Valls, G., F. Donges, J., Dorigo, W., M. Estupinan-Suarez, L., H. Gutierrez-Velez, V., Gutwin, M., Jung, M., C. Londoño, M., G. Miralles, D., Papastefanou, P., & Reichstein, M. (2020). Earth system data cubes unravel global multivariate dynamics. Earth Sys-tem Dynamics, 11(1), 201–234. https://doi.org/10.5194/ESD-11-201-2020

Palosuo, T., Kersebaum, K. C., Angulo, C. et al. (2011). Simulation of winter wheat yield and its variability in different climates of Europe: A comparison of eight crop growth models. European Journal of Agron-omy, 35(3): 103–114.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., &

Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal OfMachine Learning Research, 12, 2825–2830. https://doi.org/10.1289/EHP4713

Perich, G., Turkoglu, M. O., Graf, L. V., Wegner, J. D., Aasen, H., Walter, A., & Liebisch, F. (2023). Pixel-based yield mapping and prediction from Sentinel-2 using spectral indices and neural networks. Field Crops Research, 292, 108824. https://doi.org/10.1016/J.FCR.2023.108824

Siesto (2022). Changes in band data after 25 Jan 2022 | Baseline 04.00 harmonizeValues | Sentinel 2 L2A | Snappy. https://forum.step.esa.int/t/changes-in-band-data-after-25-jan-2022-baseline-04-00-harmonizevalues-sentinel-2-l2a-snappy/36270

Son, N. T., Chen, C. F., Cheng, Y. S., Toscano, P., Chen, C. R., Chen, S. L., Tseng, K. H., Syu, C. H., Guo, H. Y., & Zhang, Y. T. (2022). Field-scale rice yield prediction from Sentinel-2 monthly image composites using machine learning algorithms. Ecological Informatics, 69, 11. https://doi.org/10.1016/J.ECOINF.2022.101618

Vreugdenhil, M.; Navacchi, C.; Bauer-Marschallinger, B.; Hahn, S.; Steele-Dunne, S.; Pfeil, I.; Dorigo, W.; Wagner, W. Sentinel-1 Cross Ratio and Vegetation Optical Depth: A Comparison over Europe. Remote Sens. 2020, 12, 3404. https://doi.org/10.3390/rs12203404

Zhen, Z., Chen, S., Yin, T., & Gastellu-Etchegorry, J. P. (2023). Globally quantitative analysis of the impact of atmosphere and spectral response function on 2-band enhanced vegetation index (EVI2) over Sentinel-2 and Landsat-8. ISPRS Journal of Photogrammetry and Remote Sensing, 205, 206–226. https://doi.org/10.1016/J.ISPRSJPRS.2023.09.024