YIPEEO: Yield Prediction and Estimation using Earth Observation

[Database Description v1.0]

[ID DDv1]

Version 1.0

[17/08/2023]

Submitted by:

Global Change Research Institute CAS (CzechGlobe)

in cooperation with:

EODC and TU Wien

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 Database Description of the project YIPEEO (DDv1).

Number of pages: 35

Authors:	CzechGlobe: Lucie Homolová (LH), Miroslav Pikl (MP), Petr Lukeš (PL) Milan Fischer (MF), EODC: Charis Chatzikyriakou (CC), Christoph Reimer (CR), TUW: Emanuel Büechi (EB)
Circulation (internal):	Project consortium
External:	ESA

Issue	Date	Details	Editor
0.1	12.6.2023	First draft	LH
1.0	17.8.2023	Final Revisions	LH

For any clarifications please contact Lucie Homolová (homolova.l@czechglobe.cz).

Table of Content

List of figures

List of tables

Acronyms

API	Application Programming Interface
C3S	Copernicus Climate Change Service
DIAS	Data and Information Access Services
ECMWF	European Centre for Medium-Range Weather Forecasts
EO	Earth Observation
HLS	Harmonized Landsat and Sentinel-2
LAI	Leaf area index
LST	Land Surface Temperature
NUTS	Nomenclature of Territorial Units for Statistics
OGC	Open Geospatial Consortium
S-1/2/3	Sentinel-1/2/3
SIF	Solar Induced Fluorescence
SPEI	Standardized Precipitation Evapotranspiration Index
STAC	SpatioTemporal Assest Catalog
VI	Vegetation Indices

Executive summary

This document is one of two results delivered within the Task 2 – Dataset Collection described in the proposal of the ESA YIPEEO project. The scope of this document - D2.2 Database Description is to provide a comprehensive summary of the input datasets, technical solution how various data are stored in a database, and access to it. Input data are organized into four groups: i) in-situ yield data, ii) EO data, iii) meteorological data, and iv) additional campaign data.

This task was led by CzechGlobe-RS team with the support of all project partners.

1 Introduction

The aim of Task 2 is to compile and describe input data for the consecutive Tasks of the YIPEEO project. The crop yield forecasts will rely on an extensive database that include: i) in-situ crop yield data at subfield, field and regional level, ii) various EO data, iii) meteorological data, and iv) additional campaign data. All data inputs (if available) were collected from 2016 onwards and are organized in a data pool (Figure 1). The data pool is composed from a PostGIS database, which stores in-situ crop yield data, and SpatioTemporal Asset Catalogs (STAC), which allow accessing raster data (EO and meteo data). The data pool enables effective queries to select data subsets for yield model training and validation. Following sections describe individual data sources, their metadata and organisation in the data pool.

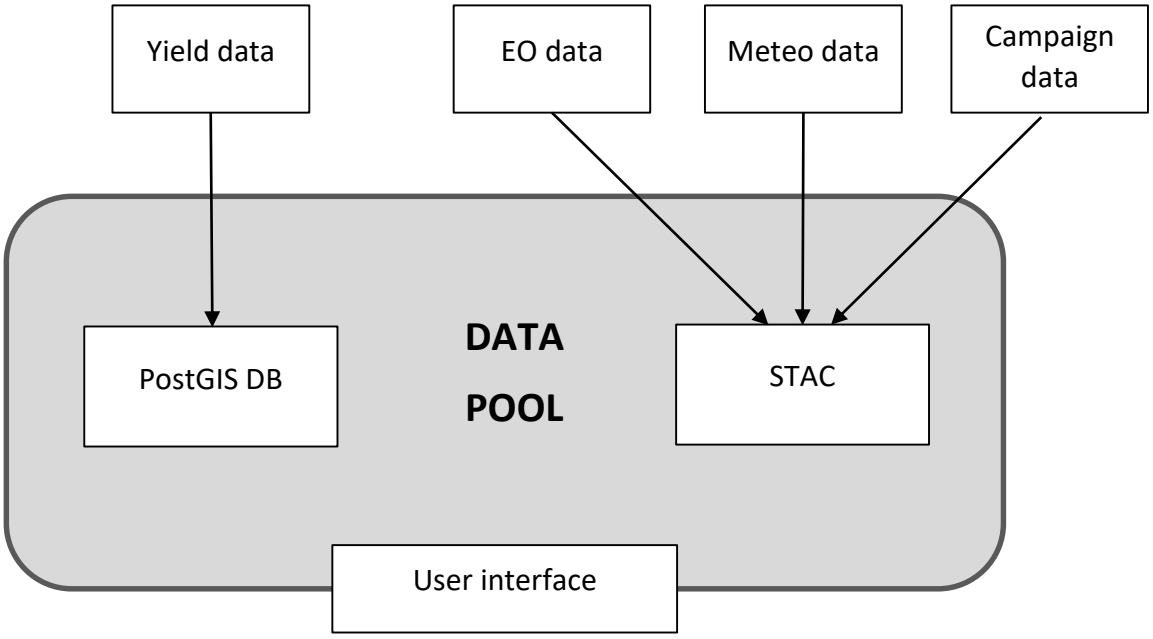

Figure 1: YIPEEO project data pool scheme (EO – Earth Observation, DB – database, STAC – SpatioTemporal Assest Catalogs).

2 In-situ crop yield data

The crop yield data are collected at three levels of detail:

- subfield (section 2.1),
- field (section 2.2),
- regional (section 2.3).

Yield data at the subfield and field levels were obtained through a network of collaborating farms and research contacts. Yield data at the regional level were obtained for selected countries (Czechia, Austria, France, Netherlands and Hungary) either through publicly available records of Eurostat for yield statistics at the NUTS2 level, or through national statistical offices or state agriculture intervention funds in case of the yield statistic at the NUTS3 and NUTS4 level. All yield data sets were harmonized in terms of crop nomenclature. We used the Eurostat crop classification names and codes (Eurostat 2023). Table 1 extracts the class names and codes for the dominant crops in our database. Yield data with associated geometry (polygons of fields and administrative units at NUTS2, NUTS3 and NUTS4 level) are stored in a PostGIS database (details in section 2.4).

Table 1: Crop classification according to Eurostat (Eurostat 2023) for the dominant crops in the project database.

Code	Class name	Code	Class name
C1111	Common winter wheat and spelt	P1100	Field peas
C1112	Common spring wheat and spelt	R1000	Potatoes
C1120	Durum wheat	R2000	Sugar beet
C1210	Rye	I1111	Winter rape and turnip rape seeds
C1310	Winter barley	I1112	Spring rape and turnip rape seeds
C1320	Spring barley	I1120	Sunflower seed
C1500	Grain maize and corn- cob- mix	I1130	Soya
C1410	Oats	I1190	Other oilseed crops
C1700	Sorghum	G3000	Green maize

2.1 Subfield level

Spatial scale	Spatial coverage	Temporal scale	Temporal coverage
1 m	Rostěnice farm (CZ) selected fields	yearly	2017 - 2022

Subfield level yield data are derived from harvest machines, and these are available only for the Rostěnice farm in Czechia. Point records require labour intense pre-processing, which includes filtering, recalibration and spatial interpolation (Řezník et al., 2019), therefore only several parcels were selected at this moment to represent the subfield-level yield data. The selected parcels represent variable soil conditions within the farm and distribution of crops between the seasons. These data were pre-processed by the project team and are delivered as raster GeoTIFFs with pixel size of 1 m (Figure 2). The spatial resolution of 1 m allows aggregating the subfield yield data into arbitrary pixel sizes based on further requirements. Subfield level yields will be used to test the scalability of the procedures developed at the field level and will help to assess the quality of field level estimates. They will also serve as reference data for process-based models which will help to study processes influencing final yield.

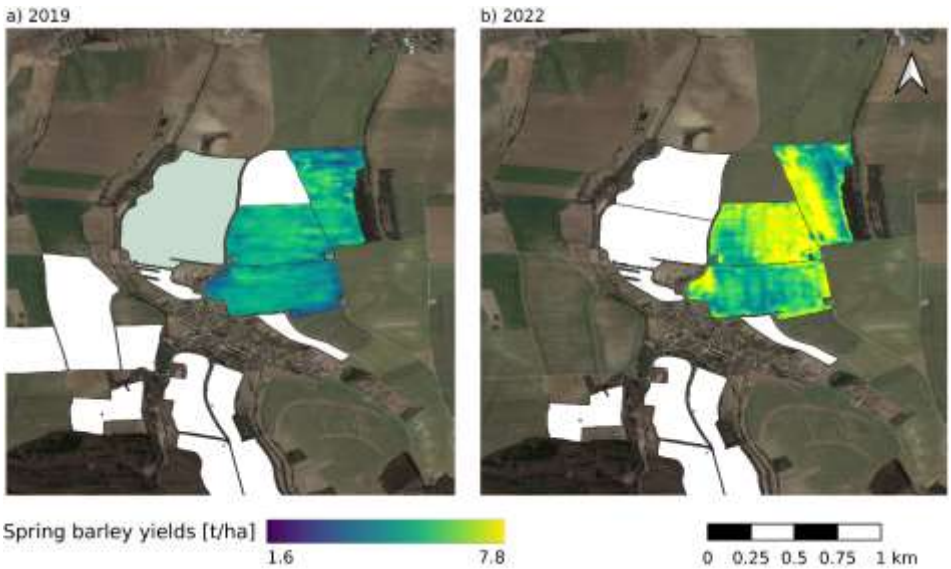

Figure 2: Example of subfield yield data variability for spring barley in 2019 (a) and 2022 (b) for selected fields at the Rostěnice farm (Czechia). Remaining white polygons represent fields with existing yield data at the field level in the particular year.

2.2 Field level

Spatial scale	Spatial coverage	Temporal scale	Temporal coverage
0.1 – 100 ha	Various farms across Europe	yearly	2016 - 2022

Yield data at the field level were obtained for the best candidate test areas, which were pre-selected to reflect various environmental conditions distributed over Europe by covering a large South-North geographical gradient (Figure 3, Table 3): sites are in southern Europe (Italy, Romania), central Europe (Austria, Czech Republic, Germany, Ukraine), and northern Europe (Netherlands, Denmark, Finland). This way, the models can be tested over different climate zones and we will get a representative impression of farms in Europe.

Yield data at the field and subfield levels were obtained through a network of collaborating farms and research contacts. Each contact was asked to provide additional information related to type of management, irrigation, nitrogen fertilisation, disturbances, etc. This additional information will support science and demonstration cases. Attributes requested together with the yield records are summarized in Table 2.

Field level yields are either computed as zonal statistics (means) from the subfield harvest machinery data (e.g. farms in Czechia) or extracted from a farm evidence statistic. Yield records at the field level were filtered for outliers, small fields with area < 0.1 ha or with inappropriate shape were removed from the database (too small or narrow fields will not allow to extract clear information from Sentinel-1 and –2 EO data). The current status of the database is summarized in Table 3. So far yield data have been collected from Czechia, Romania and Ukraine, and most of the remaining data are expected to be collected during September. Field data from Slovakia are likely not delivered. New contacts will probably make it possible to collect data from Spain. The database will be regularly updated with new data entries. Table 4 presents descriptive statistics of the available data for those crops with more than 20 records.

It is important to mention that all farmers consider their yield data as sensitive information, therefore all yield data at the subfield and field level can only be used for internal purposes of this project and cannot be disclosed outside the project consortium. The crop yield data may be shared as anomalies instead of absolute yield values, but this will need to be agreed with each farmer individually.

Table 2: Overview of the information requested from the cooperating farmers to describe field level yield data (this applies also for subfield level data).

Priority	Parameter	Comments
1	Location	preferably as field polygons (in vector format) or coordinates with a map
1	crop type	
1	harvest date	yyyy-mm-dd
1	yield	t/ha
1	moisture content	in percentage
1	data openness	indicate whether your yield data will be used only for the project or can be shared with ESA and become public
1	management type	select: conventional or organic
2	nitrogen fertilization load	N kg/ha
2	sowing date	yyyy-mm-dd
2	soil type	
2	crop damage	indicate if there has been any damage during the season such as drought, flooding, hail, pest, diseases or other
2	irrigation	indicate if you use irrigation or not
2	other comments	add other comment which might be relevant

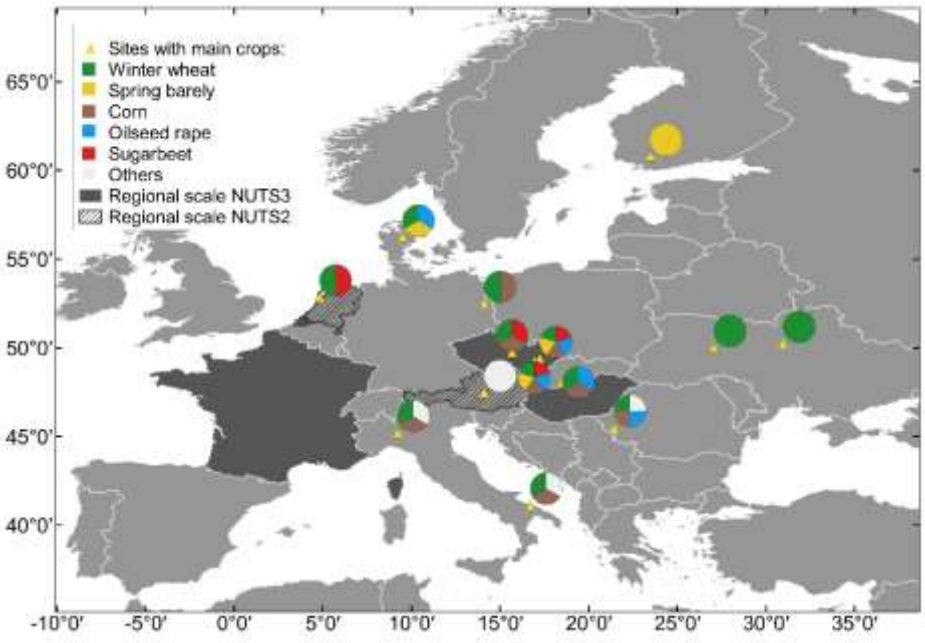

Figure 3: Distribution of test sites with crop yield data at the field level and countries with crop yield data at the regional level (NUTS).

Table 3: Summary and the actual status of the in-situ crop yield database at the field level. The last column indicates if or when a dataset is included in the database (irrig – irrigation, N fertil – Nitrogen fertilization, DB – database).

Country	Location / farm	Crops	Years	Irrig.	N fertil.	No of records	DB status
IT	Bari	Winter wheat, grain maize, rice, alfalfa	-	-	-	-	09/23
IT	Pavia	Winter wheat, grain maize, rice, alfalfa	-	-	-	-	09/23
RO	Tormac	Winter wheat, grain maize, soya, winter rape	2016-22	No	Yes	35	yes
AT	25 sites	Grass for hay	-	-	-	-	08/23
SK	various sites	Winter wheat, corn, oilseed rape	-	-	-	-	NA
CZ	Polkovice	Winter wheat, spring barley, winter rape, grain maize, sugar beet	-	-	-	-	09/23

Country	Location / farm	Crops	Years	Irrig.	N fertil.	No of records	DB status
CZ	Rostěnice	Winter wheat, spring and winter barley, winter rape, grain maize, green maize, sugar beet, soya	2017-22	No	No	1186	yes
CZ	Strážovice	Winter wheat, spring barley, oilseed rape, corn	-	-	-	-	09/23
UA	Chmelnyzkyj	winter wheat, spring wheat, sunflower, soya, sugar beet, peas	2014-16	No	No	246	yes
UA	Lviv area	winter wheat, soya	2021-22	No	No	8	yes
UA	Horodysche	Winter wheat	2015-21	No	No	25	yes
DE	Münchenberg, var. sites	winter wheat, corn	-	-	-	-	10/23
NL	Anna Paulowna	winter wheat, sugar beet, summer wheat, summer barley	-	-	-	-	08/23
DK	Various sites	winter wheat, spring barley, oilseed rape	-	-	-	-	09/23
FI	Various sites	Spring barley	-	-	-	-	09/23

Table 4: Descriptive statistics of yields at the field level for the main crops currently included in the database (crops with less than 20 records are not shown).

Crop name (crop code)	Count	Min - Max	Mean	Median	Stdev
Winter wheat (C1111)	347	2.5 - 10.8	6.8	7.2	1.6
Winter barley (C1310)	84	4.0 - 10.1	7.1	7.1	1.5
Spring barley (C1320)	241	1.5 - 10.9	5.7	5.6	1.3
Grain maize (C1500)	305	4.4 - 19.3	12.0	11.7	6.7
Soya (I1130)	112	0.6 - 3.4	2.2	2.2	0.6
Sunflower seed (I1120)	24	1.8 - 4.3	3.6	3.7	0.6
Winter rape (I1111)	191	1.5 - 5.1	3.6	3.5	0.8
Field peas (P1100)	32	0.9 - 4.9	3.0	2.8	1.2

2.3 Regional level

Spatial scale	Spatial coverage	Temporal scale	Temporal coverage
NUTS2	Austria, Czechia, Hungary, Netherlands, France	yearly	2016 - 2022
NUTS3	Czechia, Hungary, France	yearly	2016-2022
NUTS4	Austria, Czechia	yearly	2000 - 2022

Statistical yield data are collected at NUTS2, NUTS3 and NUTS4 regional levels for selected countries. NUTS2-level yield data are publicly available for Europe via Eurostat data browser (https://ec.europa.eu/eurostat/web/agriculture/data/database). The NUTS3 and NUTS4 level data are gathered by responsible project partners. Current status of the yield database at the regional level is summarized in Table 5. In addition, availability of NUTS3 level yield data for Slovakia, Poland, Austria, Slovenia, Croatia and Germany is currently under negotiation through the Clim4Cast project.

Table 5: Summary of the statistical crop yield data at the regional level for selected countries (SZIF – State Agriculture Intervention Fund of the Czech Republic).

Country	Level	Crops	Years	Source
CZ	NUTS2	Various crops	2016-2022	Eurostat
	NUTS3	Winter wheat, spring barley, winter rape, grain maize, green maize, oat, rye	2000-2022	SZIF
	NUTS4	Winter wheat, spring barley, winter rape	2000-2022	SZIF
FR	NUTS2	Various crops	2016-2022	Eurostat
	NUTS3	Winter wheat, spring wheat, durum wheat, sugar beet, winter rape, spring rape, sunflower, potatoes, winter oats, spring oats, winter barley, spring barley, grain maize, wine	1943-2018	Schauberger et al. (2022)

Country	Level	Crops	Years	Source
HU	NUTS 2	Various crops	2016-2022	Eurostat
	NUTS3	Winter wheat, spring barley, sugar beet, potatoes, grain maize	2001-2022	Not available yet
NL	NUTS 2	Various crops	2016-2022	Eurostat
AT	NUTS2	Various crops	2016-2022	Eurostat
	NUTS4	Winter wheat, grain maize, winter rape, soya, spring barely, sugar beet	2000-2022	Statistics Austria

2.4 PostGIS database

Combination of PostgreSQL database with PostGIS extension is used to store yield data at all levels and information on fertilization and/or irrigation activities.

After the input data were homogenised, the geometry was transformed to coordinate reference system common for the whole data pool (WGS84) and the data were uploaded into one of the following tables:

- yield_fl – yield data at the field level including field geometry
- nuts_geom – geometries of different NUTS levels 0 – 4
- nut_yields – yield data at the regional levels
- fertilizer_app – information about fertilisation (if available)
- irrigation_app – information about irrigiation (if available)

To ensure required functionality (spatial extraction of EO and meteorological data) and to ease the data query construction, selected tables were combined to “views” (virtual tables) based on unique attribute combinations.

The database tables and views can be divided into two blocks: field level block (Figure 4) and regional level block (Figure 5).

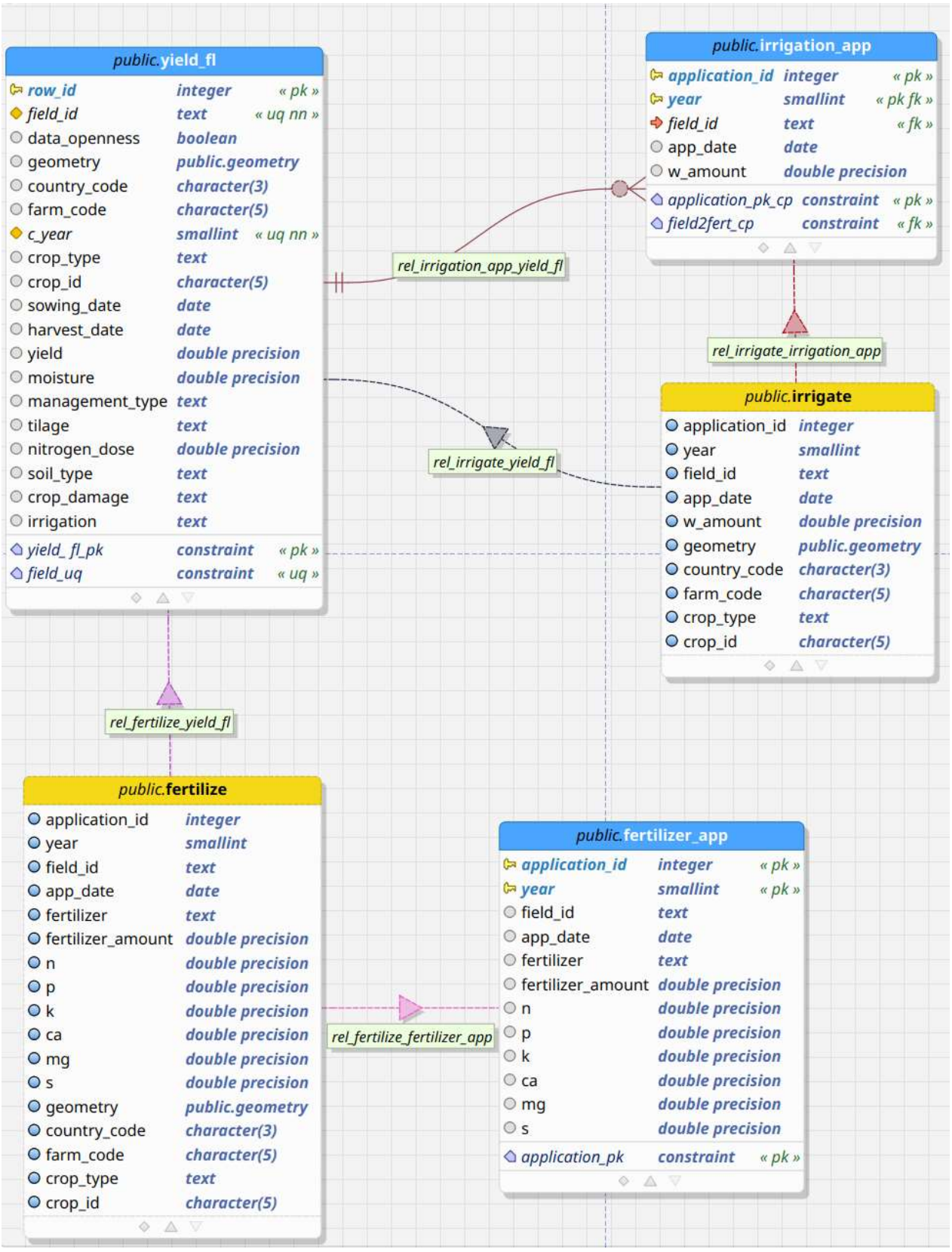

Figure 4: Database structure of data tables (blue) and views (yellow) at the field level.

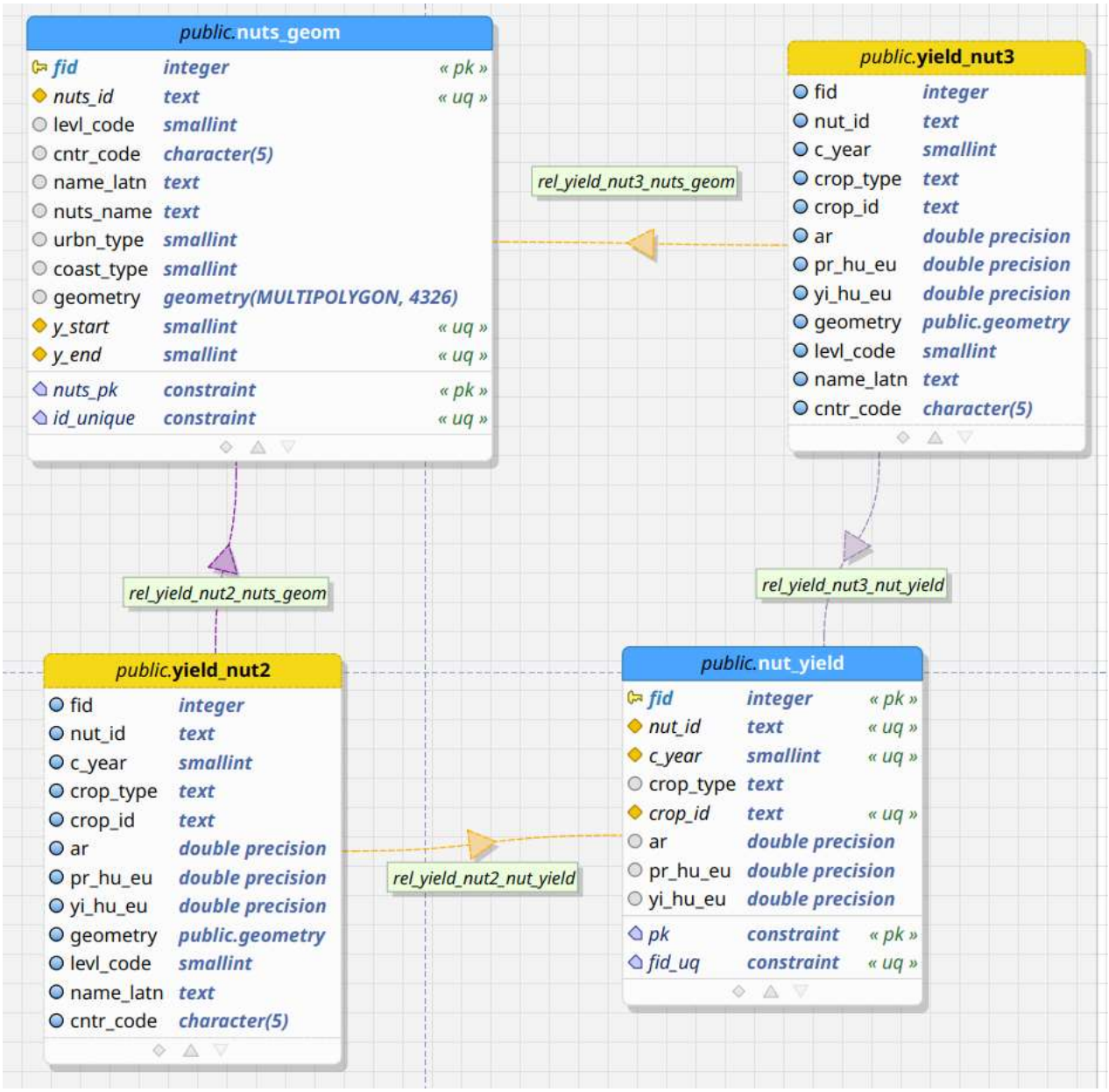

Figure 5: Database structure of data tables (blue) and views (yellow) at the regional level (NUTS).

2.4.1 ***Table description***

yield_fl

The table stores information on crop yield at the field level. Each row in the table represents crop yield from one parcel (field) in one year. The field level data can be queried directly from this table. The "*yield_fl*" attribute structure is described in Table 6.

Table 6: Attribute structure of the "yield_fl" table.

Column	data type	Description
row_id	integer	
field_id	text	Unique field geometry identifier. Composed from country code, farm code and local field ID.
data_openness	boolean	True – free to use, False – use only within the YIPEEO project
geometry	geometry(MultiPolygon,4326)	Parcel with single crop. Coordinate system WGS84 (EPSG:4326)
country_code	character(3)	
farm_code	character(5)	
c_year	smallint	crop year
crop_type	text	Eurostat crop name
crop_id	character(5)	Eurostat crop ID
sowing_date	date	YYYY-MM-DD
harvest_date	date	YYYY-MM-DD
yield	double precision	$t.ha^{-1}$
moisture	double precision	%
management_type	text	conventional/eco/na
tilage	text	yes/no/na
nitrogen_dose	double precision	total nitrogen dose N $kg.ha^{-1}$
soil_type	text	To be defined
crop_damage	text	yes/no/na
irrigation	text	yes/no/na

nuts_geom

The table stores all levels of NUTS geometries included in the YIPEEO project. Each entry contains attributes denoting geometry validity as considered within the project. The "*nuts_geom"* attribute structure is described in *Table 7*.

Table 7: Attribute structure of the "nuts_geom" table.

Column	data type	Description
fid	integer	feature ID
nuts_id	text	Eurostat NUT code
levl_code	smallint	NUT level
cntr_code	character(5)	country code
name_latn	text	
nuts_name	text	
urbn_type	smallint	1-Predominantly urban; 2-intermediate; 3 – predominantly rural
coast_type	smallint	TBD
geometry	geometry(MultiPolygon,4326)	Nut geometry downloaded from Eurostat (levels 0 – 3). Coordinate system WGS84 (EPSG:4326)
y_start	smallint	validity start (including)
y_end	smallint	validity end (including)

nut_yield

The table stores crop yield data at the different NUTS level only. The "*nut_yield*" attribute structure is described in Table 8.

Table 8: Attribute structure of the "nut_yield" table.

Column	Data type	Description
fid	integer	feature ID
nut_id	text	Eurostat NUT code
c_year	smallint	crop year
crop_type	text	Eurostat crop name
crop_id	text	Eurostat crop ID
ar	double precision	Area (cultivation / harvested / production) (1000 ha)
pr_hu_eu	double precision	Harvested production in EU standard humidity (1000 t)
yi_hu_eu	double precision	Yield in EU standard humidity (t.ha^{-1})

fertilizer_app

The table stores records of fertilizer application and nutrients supply. The "*fertilizer_app*" attribute structure is described in Table 9.

Table 9: Attribute structure of the "fertilizer_app" table.

Column	Data type	Description
application_id	integer	Feature ID
year	smallint	Harvest year
field_id	text	Unique field geometry identifier in space
app_date	date	Date when irrigation was applied
fertilizer	text	Name of fertilizer
fertilizer_amount	double precision	kg.ha^{-1}
n	double precision	nitrogen kg. ha^{-1}
p	double precision	phosphorus kg. ha^{-1}
k	double precision	potassium kg. ha^{-1}
ca	double precision	calcium kg. ha^{-1}
mg	double precision	magnesium kg. ha^{-1}
s	double precision	sulphur kg. ha^{-1}

irrigation_app

The table stores records of days when irrigation was applied and water supply. The "*irrigation_app*" attribute structure is described in Table 10.

Table 10: Attribute structure of the "irrigation_app" table.

Column	Data type	Description
application_id	integer	Feature ID
year	smallint	Harvest year
field_id	text	Unique field geometry identifier in space
app_date	date	Date when irrigation was applied
w_amount	double precision	Amount of water applied in $m^3.ha^{-1}.day^{-1}$

2.4.2 View description

yield_nut2, yield_nut3, yield_nut4

These views can be used to access NUTS yield information including geometry connected based on the harvest year. The view table contains all attributes from table "*nut_yield*" and geometry, levl_code, name_latn, cntr_code attributes from "*nuts_geom*" table.

Example of view definition for NUTS2 level:

```
CREATE VIEW yield_nut2 AS
SELECT
   nut_yield.*,
   nuts_geom.geometry,
   nuts_geom.levl_code,
   nuts_geom.name_latn,
   nuts_geom.cntr_code
FROM nut_yield JOIN nuts_geom
ON nut_yield.nut_id = nuts_geom.nuts_id AND nut_yield.c_year >= nuts_geom.y_start AND nut_yield.c_year <= nuts_geom.y_end
WHERE nuts_geom.levl_code = 2;
```

3 EO data

The main EO data sources are Copernicus operational satellite missions Sentinel-1, Sentinel-2 and Sentinel-3 that will be used for field and regional level testing. The other EO sources (including experimental hyperspectral missions PRISMA and EnMap, Harmonized Landsat and Sentinel-2 product (HLS), Sentinel-5p TROPOMI SIF and ECOSTRESS data) will be limited either to field or regional level and retrieved only for selected areas. In the analysis we will particularly focus on those datasets that have previously been identified for having a strong predictive skill, e.g., soil moisture, evapotranspiration, LAI, air temperature and precipitation from sources like e.g. Sentinel-1, Sentinel-2, EnMap and PRISMA, ERA5-Land, and C3S seasonal forecasts. These datasets have already been used to a large degree by the consortium. Sentinel-1 radar and Sentinel-2 surface reflectance data and derived soil moisture and NDVI are available on the EODC data cube. Other EO resources are either in the analysis ready form or are being pre-processed as such these can be relatively easily included into EODCs STAC.

3.1 Sentinel-1

Product	Spatial scale	Spatial coverage	Temporal scale	Temporal coverage
GRD	10 m	Selected countries (AT, CZ, HU, NL, FR, UA)	Several days	2016 - 2023
SSM	1 km	Selected countries (AT, CZ, HU, NL, FR, UA)	Daily	2016 - 2023
SWI	1 km	Selected countries (AT, CZ, HU, NL, FR, UA)	Daily	2016 - 2023

Sentinel-1 is a radar satellite mission operated by ESA using a C-band image sensor. The mission consisted of the satellites Sentinel-1A and Sentinel-1B. Since December 2021 Sentinel-1B is no longer in operation. For the two-satellite constellation period, the coverage frequency over Europe was 1-2 days. Currently with only one satellite in operation, it is reduced to 2-4 days. We will use analysis-ready Sentinel-1 backscatter data in VV and VH polarization as well

as their cross-ratio at a 20 m sampling provided by EODC. Several studies have proven the sensitivity of C-band microwaves for vegetation dynamics of various crop types (Vreugdenhil et al., 2018). Sentinel-1 data will be especially useful for time periods where no optical measurements are available due to cloud coverage and has already been successfully used in combination with Sentinel-2 for maize yield prediction in Kenya (Jin et al., 2019).

3.2 Sentinel-2

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

Sentinel-2 is an optical multispectral satellite mission operated by ESA that comprises a constellation of two satellites providing global data at 10-30 m resolution with revisit time of 5 days. For the project Level 2 surface reflectance product accompanied by cloud mask and additional quality layers is used. Sentinel-2 data are already ingested into the EODC cloud. We will explore a rich time series of Sentinel-2 and their derived products (e.g., vegetation indices, leaf area index) for timely and accurate prediction of crop yields at the field level (Hunt et al., 2019).

3.3 Harmonized Landsat and Sentinel-2 (HLS) product

Product	Spatial scale	Spatial coverage	Temporal scale	Temporal coverage
L30 reflectance S30 reflectance	30 m	Polkovice, Rostěnice farms (CZ)	Several days	2016 - 2023

The Harmonized Landsat Sentinel-2 (HLS) product combines the surface reflectance of Operational Land Imager (OLI) aboard Landsat 8 and 9 satellites and Multi-Spectral Instrument

(MSI) aboard Sentinel-2A and 2B satellites are combined in a single product (Claverie et al. 2018). The combined measurements allow global observations at a spatial resolution of 30 m, ideally every 2-3 days. It is worth noting the potential revisit period for cloud-free observations is approximately 8 days (Claverie et al. 2018). The HLS product tiling system is identical to that of Sentinel-2. The HLS version 1.4 can be accessed via https://hls.gsfc.nasa.gov/data/v1.4/ , while the latest version 2.0 is available on NASA's EarthData service (https://search.earthdata.nasa.gov/search?q=hls). The usefulness of the higher temporal frequency will be tested only at the field level for selected sites.

The HLS version 2.0 was downloaded as a single tile (33UXQ) covering the selected farms in the Czech Republic (Rostěnice, Polkovice). Multiband L30 and S30 images were created by stacking L30 and S30 products, which are distributed as separate files per each band. Only images with cloud cover less than 25%, as computed for the entire tile using Fmask algorithm, are considered for further use. A summary of these images is provided in Table 11. The data were transferred internally via FTP and incorporated into Data pool as STAC layer at EODC.

Table 11: Number of available Landsat (L30) and Sentinel-2 (S30) images in the HLS product for the Czech farms between 2016 and 2023 (June).

	L30	S30
Total number of images	329	676
Number of available images with Fmask < 25%	99	218
Number of available images with Fmask = 0	30	74

3.4 Sentinel-3

Product	Spatial scale	Spatial coverage	Temporal scale	Temporal coverage
OLCI L2	300 m	Selected countries	Several days	2016 - 2023
SLSTR L2 LST	1 km	Czechia	Several days	2016 - 2023

In case of Sentinel-3, level 2 products, surface reflectance from OLCI (21 bands at 300 m spatial resolution) and land surface temperature from SLSTR (500 m or 1 km spatial resolution) will be considered. Especially, the land surface temperature (not only from Sentinel-3, but also from ECOSTRESS and Landsat) will be tested for evapotranspiration modelling as one of the yield predictors. The evapotranspiration modelling using DisALEXI (Yang et al. 2017) will be tested over the well characterised experimental farm Polkovice.

3.5 Hyperspectral PRISMA and EnMAP

Product	Spatial scale	Spatial coverage	Temporal scale	Temporal coverage
PRISMA and EnMAP L2 reflectance	30 m	Rostěnice, Polkovice farms (Czechia)	A few observations per season	2023

PRISMA hyperspectral data are requested and collected via its dedicated PRISMA data portal (https://prisma.asi.it/) for selected sites (Czech and Dutch farms). The observation request is updated every month. EnMap hyperspectral data are accessed via the EnMap Instrument Planning Portal and the EOWEB® GeoPortal (https://planning.enmap.org/). A proposal for EnMap data acquisitions for this project was approved in May 2023 and data are currently requested for selected sites only (Czech farms). Both, PRISMA and EnMAP image tiles are 30 x 30 km and until now no cloud-free acquisition is available, therefore the added value of hyperspectral information will be tested only at the field level for selected sites.

To date, there has not been a successful acquisition of image data from either EnMAP or Prisma.

3.6 Sentinel-5 TROPOMI SIF

Product	Spatial scale	Spatial coverage	Temporal scale	Temporal coverage
L2b SIF	3.5 x 7.5 km	Europe	Daily	2018 - 2023

Sentinel-5 TROPOMI SIF product was developed by two research groups Köhler et al. (2018) and Guanter et al. (2021). Both products show overall good agreement with R2 = 0.96 (Guanter et al. 2021). In this project, the TROPOSIF product of Guanter et al. (2021) will be considered, as it covers the time period between May 2018 and April 2021 (data available in NetCDF format via a dedicated ftp access (https://s5p-troposif.noveltis.fr/data-access/) and from October 2022 the product is available through SP5-PAL STAC API (https://data-portal.s5p-pal.com/products/troposif.html). The coarse spatial resolution of TROPOSIF (3.5 x 7.5 km) will allow to test it for yield predictions at the regional level only.

TROPOSIF data are currently not included in the database, it is work in progress.

3.7 Ecostress

Product	Spatial scale	Spatial coverage	Temporal scale	Temporal coverage
LST	70 m	Rostěnice, Polkovice farms (Czechia)	A few observations per season	2018 - 2023

ECOSTRESS land surface temperature and evapotranspiration products at 70 m or 30 m spatial resolution are available from July 2018 and will be accessed via NASA's EarthData service (https://search.earthdata.nasa.gov/search?fi=ECOSTRESS).

ECOSTRESS LST data are currently not included in the database, it is work in progress.

4 Meteorological data

Product	Spatial scale	Spatial coverage	Temporal scale	Temporal coverage
ERA5-Land selected meteo parameters	0.1°	Europe	Daily	2016 - 2023
CS3 seasonal forecast for selected parameters	1°	Europe	Monthly	2016 (hindcast), 2017 – 2023 (forecast)

Meteorological variables for crop yield modelling were extracted from ERA5-Land reanalysis, which combines vast amounts of historical observations into global estimates using advanced modelling and data assimilation systems. A wide range of variables is available within the ERA5-Land database, but the most relevant meteorological parameters were identified by the project team (summarized in Table 12). Selected variables were extracted from ERA5-Land hourly data on single levels from 1950 to the present via Copernicus services (https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels).

Additional meteorological / climatological variables were computed for the purpose of yield forecasting. These include vapor pressure deficit, reference and potential evapotranspiration, surface net radiation, relative humidity (Allen et al. 1998) and Standardized Precipitation Evapotranspiration Index (SPEI, Vicente-Serrano et al., 2010). Reference evapotranspiration was computed according to standard methodology (Allen et al., 1998) with only two differences: instead of approximating surface pressure from the altitude, we used the values of surface pressure from ERA5-Land and the latent heat of vaporization was not considered as constant but as function of air temperature. SPEI was computed using an R-package SPEI (Vicente-Serrano et al., 2010) where the climatological water balance was computed as the difference between the precipitation and the computed reference evapotranspiration.

The hourly data of meteorological parameters (either retrieved directly from ERA5-Land or computed) were subsequently aggregated into daily values (averages, min, max, sum depending on a variable) and cropped from the global coverage to the spatial extent covering

Europe and Ukraine (10.65°W - 40.50°E, 71.50°N - 34.50°N). The meteorological variables are currently available for the period 2016 – 2022, data for 2023 will be added later. Files in GeoTIFF raster format were transferred internally via FTP and incorporated into Data pool as STAC layer at EODC.

The seasonal forecast data on monthly basis are extracted from the European Centre for Medium-Range Weather Forecasts (ECMWF) seasonal forecast product, which is part of a Copernicus Climate Change Service (C3S). We use ECMWF monthly mean data (system 5) in monthly step. Forecast data are available from 2017 – 2023 (2016 as hindcast). The seasonal forecast data are being pre-processed and are not included in the database yet.

Table 12: Overview of selected meteorological parameters for crop yield modelling. Meteorological parameters were aggregated into daily values (average, sum, minima or maxima) in case of seasonal forecasts these are monthly values.

Acronym	Parameter	Unit	Meteo (reanalysis, computed)	Seasonal forecast
WS10_avg	10m u- and w- component of wind (average)	$m\ s^{-1}$	ERA5-Land	C3S
T_avg, T_max, T_min	2m temperature (average, maximum, minimum)	°C	ERA5-Land	C3S
Td_avg	2m dewpoint temperature (average)	°C	ERA5-Land	C3S
VWC_1/2/3/4_avg	Volumetric soil water layers 1 – 4 (average)	$m^3\ m^{-3}$	ERA5-Land	Not available in C3S
AP_avg	Surface pressure	kPa	ERA5-Land	C3S
Rs_sum	Surface solar radiation downwards (sum)	$MJ\ m^{-2}$	ERA5-Land	C3S

Acronym	Parameter	Unit	Meteo (reanalysis, computed)	Seasonal forecast
P_sum	Total precipitation (sum)	mm	ERA5-Land	C3S
VPD_avg	Vapor pressure deficit (average)	kPa	Computed	Computed
ETo_sum	Reference evapotranspiration (sum)	mm	Computed	Computed
PET_sum	Potential evapotranspiration (sum)	mm	Computed	Will be computed
SPEI	Standardized Precipitation Evapotranspiration Index	-	Computed	Computed
RH_avg, RH_min	2 m relative humidity (average, minimum)	%	Computed	Will be computed
Rn_sum	Surface net radiation (sum)	MJ m^{-2}	Computed	Will be computed

5 Additional campaign data

Product	Spatial scale	Spatial coverage	Temporal scale	Temporal coverage
L2 reflectance	1 m	Polkovice farm (Czechia)	A few observations per season	2020 - 2022

The only additional campaign data that can be considered for this project are CzechGlobe's data for the Polkovice farm in Czech Republic. This campaign data includes airborne hyperspectral (VNIR covered by CASI-1500 and SWIR covered by SASI-600) and thermal images (covered by TASI-600) acquired by the CzechGlobe's Flying Laboratory of Imaging Systems (https://olc.czechglobe.cz/en/flis-2/) three to five times during the vegetation seasons 2020-2022 over the experimental fields at the Polkovice farm. The experiment studies the impact of till and no till management and application of biochar on growth conditions and yields of

winter wheat, spring barley, grain maize, sugar beet and soya. The yield data are available from harvest machines allowing to test subfield level predictions (Figure 6).

The spatial extent of the airborne data is very limited, covering the experimental fields only, and thus not compatible with other EO datasets. Therefore, it is not included in the project database at this moment but can be added anytime if the project team requests it. As agreed in the KO meeting, FLEXSense and SARSense campaign data will not be used due to a lack of available data at our field sites.

Figure 6: Example of the experimental campaign data for the Polkovice farm in Czechia. Crop yield data from harvest machines are displayed over a VNIR hyperspectral image displayed in true colour acquired on 7. 4. 2020.

6 SpatioTemporal Assest Catalogues for EO data

As part of the database generation process, the organisation and management of Earth Observation data, including data from Copernicus Sentinel-1, Sentinel-2, Sentinel-3, Sentinel-5p, hyperspectral missions like Prisma and EnMap, will be facilitated through the SpatioTemporal Asset Catalog (STAC) framework. This approach ensures efficient handling and accessibility of spatiotemporal resources within the project's context.

The fundamental building block of the STAC framework is the STAC Item. Representing an individual spatiotemporal resource, this unit is presented as a GeoJSON feature accompanied by datetime details and interconnected links. A spatiotemporal resource comprises a compilation of digital assets, including the primary data file, ancillary files, and associated metadata. Therefore, any data slated for processing should be prepared by incorporating STAC items in GeoJSON format, unless they are already in place. These STAC items can be stored adjacent to the primary data within a variety of storage systems, often denoted as static STAC collections or catalogues. The STAC items can be logically clustered within a STAC catalogue or collection. For an in-depth understanding of the STAC specifications, the official documentation available at https://stacspec.org/ should be consulted. The community has provided software tools for generating the necessary STAC metadata for specific Earth Observation datasets, accessible through https://github.com/stactools-packages.

The STAC specification offers a range of standardised sub-specifications, which incorporates a dynamic implementation of the SpatioTemporal Asset Catalog known as the STAC API. This API serves as a SpatioTemporal Asset Catalog, facilitating the retrieval of entities like STAC Catalog, Collection, Item, or STAC API ItemCollection from diverse endpoints. While STAC catalogue and collection entities are returned in JSON format, Item and ItemCollection entities adhere to the GeoJSON standards, including foreign members. Typically, a single Feature is employed when conveying a singular Item object, whereas a FeatureCollection structure is used for multiple Item objects, rather than a JSON array of Item entities. Establishing a STAC API for users necessitates a server implementation. The STAC community maintains several implementations based on various technology stacks, including Java, Python, Node.js,

PostgreSQL, and OpenSearch, which can all be accessed through https://github.com/stac-utils.

A notable implementation of the STAC API is stac-fastapi, accessible at https://stac-utils.github.io/stac-fastapi/. Developed as a mature Python FastAPI application, this implementation enables the provision of STAC objects in JSON format. These objects can be stored within a PostgreSQL (PostGIS) database, interconnected via one of the supported backend drivers. Indexing STAC objects, represented as JSON, into a designated PostgreSQL backend is achieved through client libraries, tailored to specific backends, such as pyPgSTAC (https://stac-utils.github.io/pgstac/pypgstac/), or via standard SQL insert commands.

7 Database user interface

The Database User Interface serves as a vital link, effortlessly connecting users with the extensive array of EO data and vector datasets used within the project. It integrates with both lTAC and the "OGC API - Features" specification (often abbreviated as OAFeat), enabled by an innovative tool known as pygeoapi (https://pygeoapi.io/).

It empowers users with the capabilities of a robust HTTP API, featuring endpoints such as /search and /collections designed to cater to common client needs. These endpoints efficiently facilitate the retrieval of specific datasets and collections, promoting dynamic interactions. The API's versatile design accommodates various query parameters, tailoring results to specific requirements and enhancing user engagement.

pygeoapi lies at the heart of this interface, representing a versatile Python server implementation of the OGC API standards. It operates as an open-source software under the MIT license, allowing organisations to create user-friendly RESTful OGC API endpoints using OpenAPI, GeoJSON, and HTML. Notably, the STAC API specification directly aligns with OAFeat, allowing for the flawless harmonisation of shared API endpoints (e.g., /collections). Beyond providing access, the user interface will simplify interactions too. Empowered by pygeoapi, users can navigate datasets, perform queries, and gain insights without needing to tackle data complexities.

Assets, including EO data, linked within the STAC items can be accessed in two distinct ways. The first method involves local access to the storage system on a dedicated virtual machine (VM), which is the current solution. The second approach, currently in development, allows immediate asset download through a direct URL that points to an HTTP Server. To ensure secure access, data retrieval is being implemented using an Apache web server in combination with Keycloak for robust Authentication and Authorization mechanisms.

Various clients, such as the pystac-client, a QGIS plugin, or the python requests module, provide pathways for users to access these assets efficiently. Metadata and preview images are available without requiring authentication and can be readily viewed through a user-friendly STAC Browser that interfaces seamlessly with the STAC API.

Similarly, vector data stored within the PostGIS database, exposed via pygeoapi, is accessible as a vector layer. For example, users can effortlessly access and interact with vector data through applications like QGIS. This functionality enhances the Database User Interface by enabling straightforward visualization and analysis of vector datasets.

Access to existing EODC data is https://stac.eodc.eu/api/v1 and implementation of the OGC vector features is work in progress.

8 References

Allen, R.G., Pereira, L.S., Raes, D. and Smith, M. (1998). Crop evapotranspiration - Guidelines for computing crop water requirements - FAO Irrigation and drainage paper 56, Rome, Italy, 290 pp.

Claverie, M., Ju, J., Masek, J.G., Dungan, J.L., Vermote, E.F., Roger, J.-C., Skakun, S.V., Justice, C. (2018). The Harmonized Landsat and Sentinel-2 surface reflectance data set. Remote Sensing of Environment 219, 145–161. https://doi.org/10.1016/j.rse.2018.09.002

Eurostat (2023). Annual Crop Statistics Handbook, 2023 Edition. Eurostat. Accessed online on 26.6.2023 https://ec.europa.eu/eurostat/cache/metadata/Annexes/apro_cp_esms_an1.pdf

Guanter, L., Bacour, C., Schneider, A., Aben, I., van Kempen, T.A., Maignan, F., Retscher, C., Köhler, P., Frankenberg, C., Joiner, J., Zhang, Y. (2021). The TROPOSIF global sun-induced fluorescence dataset from the Sentinel-5P TROPOMI mission. Earth System Science Data 13, 5423–5440. https://doi.org/10.5194/essd-13-5423-2021

Hunt, M. L., Blackburn, G. A., Carrasco, L., Redhead, J. W., & Rowland, C. S. (2019). High resolution wheat yield mapping using Sentinel-2. Remote Sensing of Environment, 233, 111410. https://doi.org/10.1016/j.rse.2019.111410

Jin, Z., Azzari, G., You, C., Di Tommaso, S., Aston, S., Burke, M., Lobell, D.B. (2019). Smallholder maize area and yield mapping at national scales with Google Earth Engine. Remote Sensing of Environment 228, 115–128. https://doi.org/10.1016/j.rse.2019.04.016

Köhler, P., Frankenberg, C., Magney, T.S., Guanter, L., Joiner, J., Landgraf, J. (2018). Global Retrievals of Solar-Induced Chlorophyll Fluorescence With TROPOMI: First Results and Intersensor Comparison to OCO-2. Geophysical Research Letters 45, 10,456-10,463. https://doi.org/10.1029/2018GL079031

Řezník, T., Pavelka, T., Herman, L., Leitgeb, Š., Lukas, V., Širůček, P. (2019). Deployment and Verifications of the Spatial Filtering of Data Measured by Field Harvesters and Methods of Their Interpolation: Czech Cereal Fields between 2014 and 2018. Sensors 19, 4879. https://doi.org/10.3390/s19224879

Schauberger, B., Kato, H., Kato, T., Watanabe, D., Ciais, P. (2022). French crop yield, area and production data for ten staple crops from 1900 to 2018 at county resolution. Sci Data 9, 38. https://doi.org/10.1038/s41597-022-01145-4

Vicente-Serrano, S.M., Beguería, S., López-Moreno, J.I. (2010). A Multiscalar Drought Index Sensitive to Global Warming: The Standardized Precipitation Evapotranspiration Index. Journal of Climate 23, 1696–1718. https://doi.org/10.1175/2009JCLI2909.1

Vreugdenhil, M., Wagner, W., Bauer-Marschallinger, B., Pfeil, I., Teubner, I., Rüdiger, C., Strauss, P. (2018). Sensitivity of Sentinel-1 Backscatter to Vegetation Dynamics: An Austrian Case Study. Remote Sensing 10, 1396. https://doi.org/10.3390/rs10091396

Yang, Y., Anderson, M.C., Gao, F., Hain, C.R., Semmens, K.A., Kustas, W.P., Noormets, A., Wynne, R.H., Thomas, V.A., Sun, G. (2017). Daily Landsat-scale evapotranspiration estimation over a forested landscape in North Carolina, USA, using multi-satellite data fusion. Hydrology and Earth System Sciences 21, 1017–1037. https://doi.org/10.5194/hess-21-1017-2017