



This project has received funding from the European Union's Horizon Europe Research & Innovation program Project. 101086355 – HORIZON-CL6-2022-GOVERNANCE-01



Deliverable D4.3 “Sensor-integrated data products, V1”

Revised version, 27 January 2025



Author		
Name	Organization	Signature
Isabelle Piccard	VITO	
Nuno Grosso	Deimos	
Hanna Huitu	LUKE	
Matti Pastell	LUKE	
Annimari Hartikainen	LUKE	
Rado Guzinski	DHI	
Sam Oswald	VITO	
Giorgia Milli	VITO	
Kristof Van Tricht	VITO	
Nick Berkvens	ILVO	
Dainis Jakovels	IES	
Idan Kopler	MIGAL	
Panagiota Louka	Neuropublic	
Sarah Verbeke	UGent	
Michal Blaszcak	PSNC	
Adam Fojud	WODR	
Valentina Manstretta	HORTA	
Mariapina Castelli	EURAC	
Paolo Cosmo Silvestro	Deimos	
Maria Pat Gonzalez	IFAPA	
Gunnar Grosse Hovest	ATB	
Nikos Tsakiridis	AUTH	
Nikiforos Samarinas	AUTH	
Eleni Kalopesa	AUTH	

Review on behalf of the Executive Board		
Name	Organization	Review date
Marios Vlachos	ICCS	25/06/2024

Revision records			
Version	Date	Changes	Authors
1.0	30/06/2024	Original document	VITO
1.1	27/01/2025	Revised version that takes the comments of the EC and external reviewers into account	VITO

Acronyms and Abbreviations

Acronyms and Abbreviations	
ANN	Artificial Neural Network
API	Application Programming Interface
APSIM	Agricultural Production Systems sIMulator
AGINS	AgroInsurance International
AI	Artificial Intelligence
ATB	Institut für angewandte Systemtechnik Bremen GmbH
AUTH	Aristotle University of Thessaloniki
AVR	AVR BVBA (Belgium)
CA	Consortium Agreement
CNHi	CNH Industrial Belgium
CNN	Convolutional Neural Network
DA	Data Assimilation
DEM	Digital Elevation Model
DES	Deimos Spain
DHI	DHI A/S (Denmark)
DL	Deep Learning
DME	DEIMOS ENGENHARIA SA
DMK	DMK Deutsches Milchkontor GmbH
DSS	Decision Support System
DVC	Data Versioning Control
EC	European Commission
EC	Electrical Conductivity
EEAB	External Expert Advisory Board
EGM	Easy Global Market SAS
EO	Earth Observation
EOD	Earth Observation Data
ET	Evapotranspiration
EURAC	Accademia Europea di Bolzano (Eurac Research)
EV ILVO	Eigen Vermogen van het Instituut voor Landbouw en Visserij Onderzoek
ExBo	Executive Board
fAPAR	Fraction of Absorbed Photosynthetically Active Radiation
FEU	Farm Europe
GA	General Assembly
GDPR	General Data Protection Regulation

GPI	Grassland Production Index
GPP	Gross Primary Productivity
GRD	Ground Range Detected
HE	Homomorphic Encryption
HPC	High Performance Computing
ICCS	Institute of Communication and Computer Systems
IFAPA	Instituto Andaluz de Investigación y Formación Agraria, Pesquera y Alimentaria
ILVO	Flanders Research Institute for Agriculture, Fisheries and Food
IPR	Intellectual Property Rights
KMI	Royal Meteorological Institute of Belgium
KUVA	Kuva Space Oy
LAI	Leaf Area Index
LAU	Local Administrative Unit
LSTM	Long Short-Term Memory
LUKE	Natural Resources Institute Finland
MIGAL	MIGAL Galilee Research Institute
ML	Machine Learning
MPC	Multi-Party Computation
MSI	Multispectral Instrument
MST	Management Support Team
NP	Neuropublic SA
NPP	Net Primary Productivity
OC	Organic Carbon
OHB DS	OHB Digital Services GmbH, Bremen, Germany
PEF	Product Environmental Footprint
PET	Privacy Enhancing Technologies
PO	Project Officer
PPI	Plant Phenology Index
PRESTO	Pretrained Remote Sensing Transformer
PSNC	Instytut Chemii Bioorganicznej Polskiej Akademii Nauk
R&D	Research and Development
RECO	Ecosystem respiration
RF	Random Forest
RGB	Red Green Blue
RH	Relative Humidity
RIE	Research and Innovation Environment

RIL	Research and Innovation Lab
RZSM	Root Zone Soil Moisture
SM	Soil Moisture
SME	Small and Mid-size Enterprise
SOC	Soil Organic Carbon
SSL	Secure Sockets Layer
T	Temperature
UAV	Unmanned Aerial Vehicle
UGent	Universiteit Gent
VI	Vegetation Index
VITO	Vlaamse Instelling voor Technologische Onderzoek
VNIR	Visual Near Infrared
VPD	Vapor Pressure Deficit
VRI IES	Foundation "Institute for Environmental Solutions"
VTT	Technical Research Centre of Finland Ltd.
WODR	Wielkopolski Ośrodek Doradztwa Rolniczego w Poznaniu
WP	Work Package

Table of Contents

1.	Introduction	12
1.1.	Project overview	12
1.2.	Scope of the document.....	12
1.3.	Document structure.....	13
1.4.	Evolution of the document	13
2.	Methodological frameworks.....	14
2.1.	Privacy-preserving technologies	14
2.2.	Data assimilation methodologies.....	16
2.3.	Data integration methodologies.....	17
2.3.1	Solutions for limited availability of labeled data for model training	17
2.3.2	Modeling: development of improved data products using sensor data	19
3.	Agri-environmental data products.....	22
3.1.	RIL Water management	23
3.1.1	Field water status and predicted yield for target crops	23
3.1.2	Satellite based field water status and predicted yield for target crop.....	27
3.2.	RIL Crop management.....	28
3.2.1	Calculated indicators (aggregates) based only on the ground truth evidence.....	28
3.2.2	Soil moisture & evapotranspiration aggregation at LAU / Commune Level.....	34
3.2.3	Calculated indicators (LUKE) based on data assimilation mechanisms along with the respective annotations	35
3.2.4	Aggregated pesticide use for policy makers	35
3.2.5	Calculated sustainability indicators	36
3.2.6	DSS model outputs.....	38
3.2.7	Statistical data on the accuracy of observations of the occurrence of agrophages.....	40
3.2.8	Improved predicted agrophage occurrence data based on geolocation.....	42
3.2.9	Predicted overall level of agrophage occurrence risk for the selected region	44
3.3.	RIL Yield monitoring.....	45
3.3.1	Potato yield estimates	45
3.3.2	Improved tare yield estimates for potatoes	52
3.3.3	Winter wheat yield estimates (LUKE)	54
3.3.4	Winter wheat yield estimates (VITO).....	57
3.4.	RIL Soil Health	57
3.4.1	EO based regional soil organic carbon map.....	57
3.4.2	Soil health indicator estimates (field level).....	59
3.5.	RIL Grassland.....	61
3.5.1	Gap-filled grasslands LAI maps at parcel level	61

3.5.2	Estimated grassland yield at parcel level	64
3.5.3	Improved grassland GPP maps based on flux tower sensors	67
3.6.	RIL Dairy	71
3.6.1	Regional productivity of dairy farms.....	71
3.6.2	Deviation of milk quality & quantity.....	74
3.6.3	Assessment of grass yield at regional level.....	76
4.	Availability of results on the RI environment	78
5.	References	80

List of tables

Table 1: Requirements and priority for the development of the Federated Learning module.....	15
Table 2: Overview of agri-environmental data products developed by the RI Labs, incl. timeline.....	22
Table 3: Pilot parcel characteristics	29

List of figures

Figure 1: General overview of the proposed pipeline for implementing few-shot learning using a robust Foundation Model pre-trained on multi-sensor data time-series. The model can be fine-tuned with a limited number of application-specific annotated examples to generate informative embeddings, which can be used to train an application-specific downstream ML model for regression, multiclass classification, or binary classification tasks.....	18
Figure 2: Local meteorological station near peppermint fields in Latvia	25
Figure 3: Node with soil moisture and temperature sensors in peppermint field in Latvia.....	25
Figure 4: Young Quinoa in Israel test field.....	25
Figure 5: Aggregation levels for the two iterations of the project	29
Figure 6: Location of pilot parcels T1 - T3, C1 - C3 and their respective commune boundaries	30
Figure 7: Location of pilot parcels P1 – P2 and their respective commune boundaries.....	30
Figure 8: Methodological framework	33
Figure 9: Visualisation of sustainability indicators in the DSS grano.net(R)	38
Figure 10: Location of the test fields in Wielkopolska district.....	41
Figure 11: On the left image rye test fields and on the right image sugar beet test fields	42
Figure 12: Histogram showing frequency of yield values within Belgium and Netherlands subfields. Data is cut to 120000 to remove impact of large outliers.	46
Figure 13: Model framework for yield estimation.....	48
Figure 14: (Top Left) Distribution of actual vs. Predicted yield (kg/Ha) for test set of BE/NL, using 1D-CNN model on full dataset (Top Right) Residual plot of yield values using 1D-CNN model on test set (Bottom) Comparison of prediction models in terms of the number of predictions within 5% and 10% of the dataset range (98047) of ground truth median yield (in Kg/Ha), showing which combinations of models correctly predict which percent of values.....	50
Figure 15: Histograms of median yield values per field for each region across the dataset.....	51
Figure 16: Fields monitored in 2023	55
Figure 17: Fields monitored in 2024	55
Figure 18: 210 sample points for topsoil SOC throughout Flanders, taken in 2021 and 2022.	58
Figure 19: 2208 sample points for topsoil SOC throughout the region of Central Macedonia, and their corresponding topsoil SOC content in the form of a histogram; samples collected between 2014 and 2024.	58
Figure 20: Location of the study area and the field sites. Imagery: Google, © TerraMetrics.	62
Figure 21: Location of the five pilot farms selected for section 3.5.2 and the flux towers (initially ECT2 and 3) used for section 3.5.3 in the Pedroches region (Spain).	65
Figure 22: Schematic representation of the adapted Monteith model that is used to calculate NPP. 66	
Figure 23: Validation of grassland biomass – results from previous campaigns.....	67
Figure 24: Sensors installed in grasslands in the Pedroches area	68
Figure 25: From field measurements to GPP and RECO variables. Data processing scheme.....	69
Figure 26: EO and in Situ data for GPP estimations. Algorithm flow chart.	70
Figure 27: Spatial distribution of dairy farm locations in the sample region in Northern Germany....	71
Figure 28: Annual time series for the milk quality parameters a) fat and b) protein percentage, as well as c) mean milk quantity per farm. Values are daily means from all farms.	73
Figure 29: Comparison of milk quality and quantity parameters between counties of Cuxhaven and Stade.	73

Figure 30: Distribution of Pearson coefficients for the correlation between time series of fat content in milk of individual farms to the time series of mean fat content over all farms. 75

Figure 31: Example for timeseries of fat percentage in milk of individual farms with diverse values of Pearson correlation coefficients to the mean values for year 2018..... 75

1. Introduction

1.1. Project overview

ScaleAgData is a response to the call HORIZON-CL6-2022-GOVERNANCE-01-11 Upscaling (real-time) sensor data for EU-wide monitoring of production and agri-environmental conditions. The ScaleAgData project will run from January 2023 till December 2026 and consists of a consortium of twenty-six partners from fourteen countries. The vision of ScaleAgData is two-fold. On one hand it wants to obtain insights in how the complex data streams should be governed and organised (governance call). On the other hand, it aims to develop the data technology needed to scale data collected at the farm level to regional datasets, agri-environmental monitoring and the management of agricultural production.

To do so, ScaleAgData has five objectives:

- Developing innovative approaches for collecting in-situ data and applying data technologies.
- Enabling and promoting data sharing along the entire data value chain.
- Demonstrating how the sensor data can be scaled to agri-environmental data products at the national, regional or European level.
- Demonstrating the benefit of the improved monitoring capacities in a precision farming context.
- Demonstrating the benefit of upscaled regional datasets for the agricultural sector in general.

During its lifecycle, the project will explore seven innovation areas: innovative sensor technology, edge processing, data sharing architecture and data governance, satellite data augmentation, from data assimilation to service development, privacy-preserving technology, and data integration methodologies.

Six Research and Innovation Labs (RIL) have been identified within the project, across various biogeographical regions of Europe, where different data upscaling and integration models or approaches will be evaluated and demonstrated. The six RILs are: water productivity, crop management, yield monitoring, soil health, grasslands and sustain dairy. Recommendations will be formulated on how such integrated datasets can be capitalized to help national and regional policy making to strengthen both the competitiveness and sustainability of European agriculture.

1.2. Scope of the document

ScaleAgData Task 4.2 is focusing on providing the methodological frameworks to integrate sensor data into agri-environmental data products. This aims to enhance monitoring capacities both at the local scale, by providing improved information and services to farmers, and at regional scale, by supporting public authorities and private companies operating within the agricultural sector with relevant information. This integration of sensor data can be done at several stages of the monitoring workflows. Sensor data can be used (i) to directly improve the models, (ii) adapt/correct the model outputs, or enable the development of a new data product based on only a limited number of sample data.

The ScaleAgData methodological frameworks and prototypes are evaluated by the RI Labs in Task 4.4. The methodological frameworks (code and notebooks) and the resulting data products are made available on the Research & Innovation Environment (RIE), set up in Task 4.3, together with the already existing data products. An iteration approach will be followed aiming to reach a TRL5. Based on these

validations, the RI Labs will have a clear understanding of which data and methods can be implemented in their RI Lab for their specific applications.

The present document provides a description of the methodological frameworks that have been developed in Task 4.2 and that are evaluated by the RI Labs in Task 4.4. It presents the resulting data products that have been developed during the first iteration round (M7-M18) and that will be made available on the RIE that has been set up in Task 4.3.

1.3. Document structure

This document is structured as follows:

- Chapter 1 provides a project overview and then goes on to describe the scope, responsibilities, and structure of this deliverable.
- Chapter 2 provides an overview of the methodological frameworks that are being developed to facilitate and promote the use of sensor data to improve agri-environmental data products
- Chapter 3 describes the agri-environmental data products that are being developed in the RI Labs during the first iteration round
- Chapter 4 explains describes how and where the methodological framework and data products will be made available on the Research & Innovation Environment (RIE)

1.4. Evolution of the document

Version 1.0 of this document, submitted on 30 June 2024, provides a description of the methodological frameworks and the resulting sensor-integrated data products that have been developed during the first iteration round of the ScaleAgData project (M7-M18).

The present version of this document, version 1.1, submitted on 27 January 2025, includes minor changes, taking the comments of the EC and external reviewers on this deliverable into account. References to ScaleAgData Tasks 4.2, 4.3 and 4.4 have been added to section 1.2 (scope of the document) and to the introduction of Chapter 2 and Chapter 3. A clarification about the technology validation outside the RIE has been added to Chapter 4.

An updated version of this deliverable, version 2.0, with the methodological frameworks and sensor-integrated data products that will be developed during the second iteration round is foreseen for December 2025.

Additional updates will take place if necessary due to changed circumstances that require alterations to the approaches presented herein.

2. Methodological frameworks

ScaleAgData aims to provide methodological frameworks to support the wider use and integration of sensor data in agri-environmental data products, to provide better monitoring capacities both at the local and regional scale.

The following sections describe the methodological frameworks that have been developed during the first iteration round of ScaleAgData (status at M18), based on the needs and requirements gathered during the co-design process (WP2) by the RI Labs and their stakeholders, and include:

- Federated learning, a privacy-preserving technology which enables the use of private sensor data for model training,
- Data assimilation methodologies, combining sensor data and EO data products with crop simulation models as part of Digital Twins, and
- Data integration methodologies, including
 - solutions for training models with a limited amount of training data, such as few-shot learning, and
 - methods to integrate local sensor data into spatially explicit data products to improve the local accuracy of the model or update it in near real time.

During the first 18 months, the methodological frameworks have been defined, and collaborations have been established with several RI Labs. Methods have been developed for a number of specific use cases (Task 4.2). These methods have been tested within the labs as reported in Chapter 3 (Task 4.4). In the coming months (M18-M24), and especially during the second iteration, the methods will be further developed and refined. Additionally, it is expected that more collaborations with labs will be initiated.

2.1. Privacy-preserving technologies

In the context of Big Data EO analytics, privacy is increasingly becoming a concern. The advanced development of EO sensors and the proliferation of satellites in orbit today means there is a significantly large amount of recent imagery accessible today. Coupled with the rise of EO data (EOD) fusion with other information sources, allowing for the generation of new links between statistical or qualitative data and geographic information systems, concerns over data privacy and ethics arise. This combination of EOD with micro-data which holds information collected on individual units such as people or households, has produced many useful maps of metrics such as population (e.g. WorldPop - <https://hub.worldpop.org/>).

To try and facilitate this potential while mitigating privacy risks, there has been a drive in privacy preserving techniques that can be used to ensure privacy when accessing data for a range of analytics, i.e. Privacy Enhancing Technologies (PETs). There are three main groups:

- model-based: aimed at ensuring privacy through the use of model handling or training techniques (Federated Learning, Continuous Learning and Neural Network Encoding methods);
- data-based: techniques that, without hiding data, remove or change its features, such as Differential Privacy or Differentially Private Continual Learning, and
- encryption-based: techniques that use encryption to achieve privacy, such as, Multi-Party Computation (MPC) or Homomorphic Encryption (HE).

ScaleAgData, throughout the project, will gather data sharing privacy preserving user requirements in the co-design workshops (organized in WP2) to understand which techniques are more relevant to

develop in this activity. One of the most promising of those methods is Federated Learning. It can be defined as a setting where several machines (clients) have data that cannot be shared, and a central entity (a server) coordinates the updates of the models that are trained individually in each client and aggregated in a central server. In the opposite direction, this setup also allows the global model to be partly retrained for a specific region with local data. All these steps are done with no data sharing between the different clients or between clients and server. Based on the user requirement analysis, which will be continuously updated throughout the project, the project aims to develop a Privacy Preserving software Python based library or modules based on the requirements coming from the RI Labs that could support them if they require to implement a PET related technique.

The first development iteration of this module was based on requirements coming from the Soil Health RI Lab that are implementing a Federated learning experiment to train an AI based soil carbon content estimation model. The requirements are listed in the table below. As a result of dedicated meetings with that RI Lab team it was possible to understand that their requirements were associated with the extension of the capabilities of the Federated learning framework Flower.ai. More information on this library can be found here: <https://flower.ai/>.

Table 1: Requirements and priority for the development of the Federated Learning module

Category	Requirement	Description Subtasks	Priority 1=High 3=Low
Development	Network	To write a class/function that helps to deploy and configure Flowers.ai servers and clients within the scaleAgData platform when a cluster is created in the virtual lab. The function will facilitate the establishment of all the connections (IP, SSL keys...) so that the users don't have to specify anything except what instances to use for the training. Write a Unit Test to validate Document this function in a Jupyter Notebook	1
Development	Metrics	To add additional metrics for the training / evaluation / validation Write a Unit Test to validate Document this in a Jupyter Notebook	(done)
Development	Validation data	To add the possibility of specifying local datasets for the validation of the model during the training phase (after each update of the global model). This could be done by specifying it, when a client join a server. Write a Unit Test to validate Document this in a Jupyter Notebook	2
Development	Balanced datasets	To add the possibility of specifying the weight of each dataset in the contribution of the model, during the training. When you start the training	2

		<p>on the server, a weighted metric will be selected (to be added to the list of metrics)</p> <p>Write a Unit Test to validate</p> <p>Document this in a Jupyter Notebook</p>	
Study	Pretraining	Analyse/document the benefits of pretraining the model locally, before to train the model using federated learning, depending on the size of the different datasets	2
Models	Classic ML	To identify if it is possible to use classic Machine Learning models such as Random Forest, Support Vector Machine, Boost (...) with Flower.ai and to write some tutorial with Jupyter Notebook.	3
	Advance ML	To explain, how to add new/custom models within the framework by writing a tutorial with Jupyter Notebook	3

The outcome for this first iteration round of ScaleAgData was focused on the implementation of the first requirement: the setup of client and servers using the Flower.ai platform. The first version of this module is available in the project Github repository - <https://github.com/spatialops-team/ScaleAgData>. During the development phase, the access to the GitHub repository is restricted to the project partners. However, access to other parties can be provided upon request during this phase.

In the coming months, the team will continue to implement the identified requirements, following the defined priority list. In parallel, there will be a collection of additional requirements for the second iteration round coming from all RI Labs to complement the table above.

2.2. Data assimilation methodologies

Data assimilation is the process of combining models with observations. In ScaleAgData data assimilation will be used as part of Digital Twins (see also deliverable D4.2) to combine remote sensing data products with crop simulation models with the aim to deliver yield and yield quality forecasts, precision nitrogen management decisions based on those forecasts and irrigation scheduling.

Data-assimilation of remote sensing or proximal measurement data to crop growth models is a common practice in crop modeling. The most common approach is to assimilate observed crop leaf area index (LAI) to the crop model by adjusting modeled crop biomass at the time of the observation using e.g. Ensemble Kalman Filter. The biomass pools of the crop model need to be rebalanced between different organs (stems, leaves, grains, roots) based on estimation of their relative contribution to the LAI. This approach is based on the theoretical assumption that the crop model calibration, soil water holding parameters and soil nutrient status are correctly initialized and the difference between observed and modeled values is due to unknown factors such as pests or unobserved local rainfall.

When working with on-farm data, these assumptions don't necessarily hold. For instance, there can be a shortage of data from each cultivar to reliably identify cultivar parameters and detailed soil profiles are usually not available. The data assimilation methodology in ScaleAgData will assume that there is also uncertainty in the crop model parameters and initial conditions. Crop model ensembles

considering these uncertainties will be developed and most likely models under the observations will be used for forecasting and decision making. The most likely model will be chosen with a modification of the particle filter algorithm. The use of multiple new EO data products, such as canopy chlorophyll content, canopy water content and soil evapotranspiration, simultaneously as observation in the assimilation will be studied. We expect that the inclusion of new data products will allow more accurate modeling of crop nitrogen and water stress effects during the growing season.

During the first iteration round, a python library and C# extension has been developed to run model ensembles using the APSIM crop simulation model and as well as an example Jupyter notebook for running the code. Past experimental data and Terrascope EO data products from wheat fields from Finland and Belgium have been collected for development and testing of the algorithms. The performance of the method is being explored in collaboration with the Labs (see also Chapter 3).

2.3. Data integration methodologies

2.3.1 Solutions for limited availability of labeled data for model training

Remote sensing offers a large availability and variety of data that can be used as inputs to train Machine Learning (ML) and Deep Learning (DL) models. The most typical learning paradigm to obtain powerful models is supervised training, which expects the utilization of large and fully labeled datasets. However, extensive data acquisitions, privacy concerns and insufficient annotations often restrict the amount of available labeled data. Yet, this is crucial to obtain reliable and robust models that can be exploited in real-world applications.

Recent approaches to overcome these limitations involve leveraging pretrained models that have learned either in a supervised manner from large, annotated datasets or in an unsupervised or self-supervised manner from unlabeled datasets. In the latter scenario, when the training dataset is large and diverse enough to comprehensively cover the variations in the input distribution, the resulting model exhibits strong generalization capabilities which can be exploited for wide range of applications dealing with similar types of input data. These models are known as Foundation Models. In remote sensing, for instance, a model trained on extensive unlabeled satellite data which learned to capture and interpret the variations of different type of sensor signals, could be adapted for multiple downstream tasks that make use of similar input signals.

Given a dataset with only a few annotated examples, a pretrained model can either be used to create compressed versions of the inputs for supervised ML training or provided with a classification/regression layer and retrained for the specific task.

Few-shot learning aims to develop models that can learn from a small number of labelled instances while enhancing generalization and performance on new, unseen examples. In ScaleAgData, we intend to implement few-shot learning using a robust Foundation Model pre-trained on multi-sensor data time-series. This model will be fine-tuned with a limited number of application-specific annotated examples. The resulting network should generate informative compressed versions of the inputs, known as embeddings, which can be used to train an ML model tailored to specific tasks. The figure below offers a general overview of the proposed pipeline.

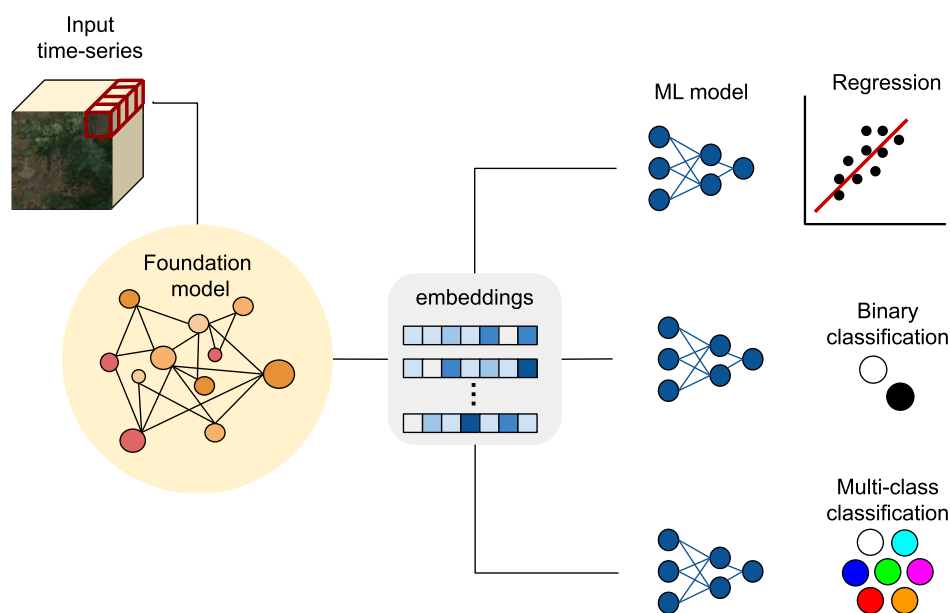


Figure 1: General overview of the proposed pipeline for implementing few-shot learning using a robust Foundation Model pre-trained on multi-sensor data time-series. The model can be fine-tuned with a limited number of application-specific annotated examples to generate informative embeddings, which can be used to train an application-specific downstream ML model for regression, multiclass classification, or binary classification tasks.

To achieve our objectives, we build upon the Presto (**P**retrained **R**emote **S**ensing **T**ransformer) framework as our foundation model (Tsjeng et al., 2023). Presto was originally trained on a large unlabeled dataset (~21M examples) of Sentinel-2, Sentinel-1, Meteorological (precipitation and temperature) and Topography (slope and elevation) pixel-timeseries, with each time-step corresponding to the monthly aggregation of the sensor observations (Tsjeng et al., 2021). The model integrates such multi-sensor data and, for each pixel-timeseries, compresses the information into 1D embeddings, effectively capturing long-range relationships across the temporal and radiometric & sensor dimensions. Utilizing Presto allows us to eliminate redundant details from the raw input data, to significantly increase the signal-to-noise ratio, and to produce a concise yet highly informative version of the inputs. Making use of these rich embeddings rather than raw data should make it easier to effectively train a ML model for a given downstream task and achieve higher performance even in case of limited amount of data available. In essence, the label-free self-supervised learning step leads to a significant head start for a downstream supervised learning task compared to any other approach that would have to start learning from scratch.

For the first iteration round, we selected the version of Presto developed in collaboration with WorldCereal¹. This updated architecture was enhanced to deal with the sensor input data that will be used within the ScaleAgData project and to manage scenarios with missing time steps or sensor data points, an updated strategy compared to the original Presto architecture. We further refined the Presto architecture in the framework of this project by enabling Presto to ingest and accurately interpret data at a 10-day temporal resolution instead of only monthly resolution, as well as to be fine-tuned for regression and multi-label classification tasks. This ensures that the few-shot learning

¹ <https://github.com/WorldCereal/presto-worldcereal/>

methodology developed within this task is able to maximally support the multitude of use cases outlined within the different research labs.

2.3.2 Modeling: development of improved data products using sensor data

Earth Observation based data products provide a broad spatio-temporal overview of Earth's surface while sometimes being limited when it comes to localized accuracies. On the other hand, in-situ measurements are usually characterized by superior accuracy while providing only a limited snapshot in space and time. By synergistically utilizing the EO methods and in-situ data it should be possible to obtain a product which combines the advantages of both.

In this methodological framework we focus on refining and improving the Root Zone Soil Moisture (RZSM) data product produced by DHI. The current operational workflow to estimate RZSM is based on the integration of remote sensing data and physics-based methods. Physics-based methods rely on physical relationships that connect remotely sensed data with the estimated variable and, in addition to EO observations, require other input data, such as local soil properties, meteorological forcings and information on agricultural practices (such as irrigation). By using local measurements in addition to potentially less accurate global or regional datasets, we can improve the accuracy of models and the quality of data products.

Integrating in-situ measurements with the existing modeling framework can be done in different ways, described below.

Machine Learning / Deep Learning

Extensive literature documents the use of Machine Learning (ML) methods as a way of integrating EO and in-situ observations while addressing the non-linear relationships between remotely sensed data and RZSM (Li et al., 2023; Kisekka et al., 2022; Guo et al., 2023; Yinglan et al., 2022; Souissi et al., 2022). These methods can simplify the model, handle different and discontinuous data sources, and predict soil moisture at deeper soil layers.

In-situ observations of soil moisture are essential for model training and validation. Some predictor variables, which are also used as input to physical models of RZSM, that can be used to construct the ML model include:

- Meteorological forcings
- EO-based evapotranspiration
- EO-based vegetation indices (e.g. NDVI) and biophysical parameters (e.g. LAI)
- Soil parameters and characteristics (e.g. sand, silt and clay content, wilting point, field capacity)

Within the model construction process, input variables undergo a feature selection phase, which filters them according to their statistical importance and influence on the model. Interestingly, the influence of meteorological variables was shown to be relatively small in some conditions (Li et al., 2023).

The most commonly used models to estimate soil moisture are Random Forest (RF) and Artificial Neural Networks (ANN) (Kisekka et al., 2022). Long Short-Term Memory (LSTM) models have also been investigated due to their ability to learn time-dependencies, such as the influence of soil moisture history on present soil moisture (Yinglan et al., 2022).

An important aspect to consider is transferability of a model trained on in-situ SM measurements to other parts of the same field or even the region. One limiting factor could be the variability of soil

parameters, which strongly influence the water storage capacity of the root-zone. Previous studies have shown that for accurate results, the model should be trained with in-situ SM measurements taken in all soil types existing within the area of interest (Kisekka et al., 2022). This implies that the collection of in-situ SM data and location of the sensors could be as important as ML model design. Therefore, a related study will be conducted during which the number and placement of in-situ sensors will be determined, based also on inter- and intra-field variability observed within EO datasets, including EO-modelled RZSM.

Additionally, literature suggests that the blending of physical-based and machine learning models could address transferability and achieve greater accuracy (Li et al, 2023). Deep learning models such as LSTM require extensive training data, stressing once again the importance of data collection, including sensor placement and number. In this framework, sensor data could be used both to train the LSTM model directly, and to create a physically modelled input to it.

Parameter Tuning

Another possibility of using in-situ observations is either to fine-tune physical model parameters or as direct input to such models. The input parameters with probably largest uncertainty are soil properties. A limited number of global databases of soil properties exist (e.g. OpenLandMap - <https://openeohub.org/about-openlandmap/>), but they are usually based on augmented extrapolation of sparse in-situ observations. By analyzing a long time-series of in-situ SM measurements, and combining this with other datasets, it is possible to determine critical soil properties such as wilting point (when soil is the driest), saturation (when soil is the wettest) or field capacity (steady-state obtained after large wetting event in the absence of large evapotranspiration). Under the assumption that the local soil parameterization, from one or more SM sensors, is more representative of local and regional conditions than the global datasets, this information can then be used within the EO-based RZSM model.

Similarly, data from local agro-meteorological station could be used as meteorological forcings for local-extent or regional-extent runs of the EO-based model, instead of estimates from global or regional weather models. This is expected to have most impact on parameters which can be highly spatially variable, such as rainfall and solar irradiance (i.e. cloudiness). Finally, irrigation timings and amounts can be also used as inputs into a field-extent model runs, replacing less accurate irrigation detection based on satellite data.

Data Assimilation

Data assimilation (DA) methods have been proven to yield good results for estimation of RZSM (Li et al., 2023) and are frequently used in meteorological or hydrological models. In DA, the state of the model is periodically updated based on observations while taking into account the uncertainties of both the model and the observations. DA works best when the modelling domain is connected and changes in one area have an impact on other areas (e.g. updating the state of upstream river flow will influence the downstream river flow). However, in the current DHI implementation of the RZSM model, each pixel is modelled independent of its neighboring pixels. At the same time, it cannot be assumed that RZSM of pixels within a given area have the same value or even shown the same temporal patterns. Therefore, further investigation into the applicability of DA will be undertaken, but at the moment it does not seem like a feasible solution for combining in-situ and EO data in the context of current RZSM modelling.

Current status, timeline and outcomes

As part of initial stages of the development of this methodological framework, timeseries of ET and SM maps have been provided to six ScaleAgData partners from three RI Labs (Water Management and Crop Management, Yield Monitoring). The provision of those data has a triple purpose:

- To enable the Labs to integrate it into their own models of e.g. yield or irrigation requirements.
- To enable the partners to get familiar with the data before sharing their in-situ observations.
- To assess the accuracy and limitations of the current EO products.

In all cases, the already provided data serves as a benchmark against which improvements derived from integrating in-situ and EO data can be evaluated by the Labs and the partners.

At the same time, an experiment is being conducted, together with WP3, to assess the utility of EO data (in particular SM and ET products) in determining the optimal placement and number of in-situ SM sensors to capture the spatial variability of a field or a region. This is directly related to the aims of the development of this methodological framework since it will allow for a following workflow:

- EO data products are used to determine the placement of in-situ sensors.
- The minimum required number of sensors are placed in the ground at specifically selected locations.
- The data from those sensors is integrated with EO products to provide detailed spatio-temporal maps of SM at field or regional scale.

The timeline for the completion of this task follows an agile approach with two development cycles as show below:

- July 2024 – December 2024 – Prototyping of the different methods (machine learning / deep learning, parameter tuning) based on existing EO products and in-situ data received from the partners.
- January 2025 – June 2025 – Deployment of the developed methodological framework and the improved EO products on the RIL environment for testing and feedback by Labs and partners.
- July 2025 – September 2025 – Refinement of the methodological framework based on feedback from the Labs and partners.
- October 2025 – December 2025 – Redeployment on the RIL and final assessment by Labs and partners.

The outcome of this development will be knowledge and tools / models to enhance the integration of EO products and in-situ data both within the project but also in the general community. While SM is used during the development of the methodological framework it is expected that large parts of it will have a broader applicability to also other EO products and types of in-situ data.

3. Agri-environmental data products

This chapter describes the various agri-environmental data products that are being developed in the ScaleAgData RI Labs. By providing sensor-integrated data products, ScaleAgData aims to improve monitoring capacities both at the local scale, by assisting farmers and agricultural advisors with actionable information and tools, and at regional scale, by supporting public authorities and private companies operating within the agricultural sector with relevant information.

During the first iteration round (M7-M18), the RI Labs mainly focused in Task 4.2 on developing basic models to generate agri-environmental data products. Sensor data, EO data and other reference data were collected. Collaborations were set up with technology providers in the context of WP3, related to new sensor developments and sensor data standardization, and in WP4, to co-develop and test the various methodological frameworks (see Chapter 2). In some Labs, this already resulted in initial data products that were validated in the Labs as part of Task 4.4. For most Labs, however, model development is still ongoing. Consolidated data products are expected to become available on the RIE that has been set up in Task 4.3 (see Chapter 4) after the 2024 growing season, and will be used to set up demonstrators in WP5.

In the second iteration round, the developed models will be further refined in WP4. The focus will shift on the one hand towards the integration of new and improved sensor and/or EO data, in close collaboration with WP3, and on the other hand towards the provision of demonstrators that can be tested by end-users, as part of WP5.

The table below provides an overview of the data products that are being developed in the different RI Labs during the first iteration round and when these products will become available.

In the sections below, the data products that are produced in each of the RI Labs and the underlying models, including the required input data, are described in more detail.

Table 2: Overview of agri-environmental data products developed by the RI Labs, incl. timeline

RIL	Data product	Timeline
Water productivity	field water status and predicted yield for target crops	Dec 2024
	satellite based field water status and predicted yield for target crops	Jan 2025
Crop management / Agri-environmental monitoring for policy makers	soil moisture and evapotranspiration aggregation at LAU / Commune Level	Dec 2024
	calculated indicators (aggregates) based only on the ground truth evidence	Dec 2024
	calculated indicators based on data assimilation mechanisms along with the respective annotations	Dec 2024
	aggregated pesticide uses for policy makers	Dec 2024
Crop management / Sustainability performance	Calculated sustainability indicators	Dec 2024
	DSS model outputs	Dec 2024
Crop management / Early pest detection	statistical data on the accuracy of observations of the occurrence of agrophages	Dec 2024
	improved predicted agrophage occurrence data based on geolocation	Dec 2024

	predicted overall level of agrophage occurrence risk for the selected region	June 2025
Yield monitoring	Potato yield estimates (subfield level)	June 2024
	Improved tare yield estimates for potatoes	Dec 2024
	Winter wheat yield estimates (LUKE)	Dec 2024
	Winter wheat yield estimates (VITO)	Dec 2024
Soil health	EO based regional soil organic carbon map	Dec 2024
	Soil health indicator estimates	March 2025
Grasslands	Gap-filled grasslands LAI maps at parcel level	Dec 2024
	Estimated grassland yield at parcel level	Dec 2025
	Improved grassland yield maps based on flux tower sensors	Dec 2025
Dairy	Regional productivity of dairy farms	June 2024
	Deviation of milk quality & quantity	June 2024
	Assessment of grass yield at regional level	June 2024

3.1. RIL Water management

3.1.1 Field water status and predicted yield for target crops

This product, developed by IES and MIGAL, includes field level target crop water status and potential yield estimates from in situ meteorological and soil sensor data in combination with high resolution (<1 m/pix) airborne VNIR spectral and thermal data for more detailed distribution insight into the field level. Two target crops requiring intensive irrigation are chosen – Peppermint in Latvia and Quinoa in Israel. IES is responsible for Peppermint model development, MIGAL for Quinoa. Both models will be based on the same principles and will use similar input data, but they will be adapted and demonstrated for specific target crops.

Sensor input data:

In situ sensor data is collected from local meteorological stations and soil sensor probes, additional reference data is collected manually. Initial data acquisition was organized during the vegetation season 2023, however, it failed in the case of peppermint in Latvia due to severe weather conditions (drought followed by heavy rain with hail) resulting in the total loss of yield. According to the project timeline, model development and testing will be based on data from the vegetation seasons of 2024 and 2025. Currently, data acquisition of the vegetation season of 2024 has started.

Sensor data	Source	Data provider	AOI/test sites	Nr. Fields/season	Meas. frequency	Season(s)	Nr. training data
Precipitation (mm) Air temperature (C) Solar radiation (W/m2) Air humidity (%) Air pressure (mm Hg) Wind speed (m/s) Wind direction	Local meteorological station	IES and MIGAL	Peppermint test fields in Latvia, Quinoa test fields in Israel	1 station for 4 Peppermint fields in Latvia, 1 station for 3 Quinoa fields in Israel	Once every 30 min	2023*, 2024, 2025	1440 per month

Evapotranspiration (ET, mm day)							
Soil temperature (C) Soil moisture (cB or %)	Soil sensors (~10 cm depth)	IES and MIGAL		4 sensors in Peppermint fields in Latvia, 12 sensors in Quinoa fields in Israel	Once every 30 min	2023*, 2024, 2025	5760 per month
Irrigation water supply (mm)	Irrigation monitors	IES and MIGAL		3-4 different irrigation regimes	Once per irrigation action	2023*, 2024, 2025	TBD
Yield (t/ha)	Weighting	IES and MIGAL		4 fields	Once per season	2023*, 2024, 2025	4
Plant status (healthy / stressed)	Visual assessment	IES and MIGAL		4 fields	Once per week	2023*, 2024, 2025	16 per month

**only in Israel*

Test fields are located in two distant regions, Peppermint fields in Latvia and Quinoa fields in Israel. In the case of peppermint, there are four test fields (~1 ha each) with different irrigation regimes located in one place. The local meteorological station is located in close proximity of the fields (see image below) and has a mobile data transfer module. Soil sensors are located within the fields and are wirelessly connected to the meteorological station (see image below). Meteorological and soil sensor data is acquired continuously (once every 30 minutes). Irrigation water supply is measured manually during the watering process using manual rain gauges. Plant status (healthy / stressed) is assessed visually once a week during on-site visits. Yield assessment is planned once at the end of the season.



Figure 2: Local meteorological station near peppermint fields in Latvia



Figure 3: Node with soil moisture and temperature sensors in peppermint field in Latvia

In the case of Quinoa, there are three test fields (1.44 ha each) with different irrigation regimes located within the perimeters of MIGAL’s “Mataim” experimental farm. The local meteorological station is located 5 km from the test site and its data is automatically stored in a designated, yet open to public site (<http://www.mop-zafon.net/>). Similar to Latvia, meteorological and soil sensor data is acquired continuously (once every 30 minutes). Yet the soil moisture sensors are rented for the duration of the growing season from an AgTech company – Phytex.com, which provides installment and maintenance of the sensors and storage of their data on a designated cloud. The access to the data is admitted only to specific users (MIGAL crew). Irrigation water supply is measured automatically via an irrigation computer. The amount of irrigation is concluded each morning according to the Penman–Monteith ET measure supplied by the weather station. The irrigation method is drip irrigation. As the site is located in MIGAL’s experimental farm Plant status (healthy / stressed) is assessed visually daily. Yield assessment is planned once at the end of the season.



Figure 4: Young Quinoa in Israel test field

Airborne input data:

Airborne remote sensing data is gathered using drone or small manned aircraft.

EO products	EO data provider	AOI/test sites	Time period	Temporal frequency	Spatial resolution
Thermal image	IES and MIGAL	Peppermint test fields in	Jun-Sep 2024 and 2025 for	Monthly	1 m
VNIR hyperspectral	IES and MIGAL	Latvia, Quinoa test fields in Israel	Peppermint, June-Oct 2024 and 2025 for Quinoa	Monthly	1 m

Despite different data acquisition platforms and spatial resolutions, it is expected to acquire comparable airborne data which could be used for the assessment of vegetation, evapotranspiration indices and plant status distribution mapping within the field of interest. Monthly airborne data acquisition is planned during two vegetation seasons – 2024 and 2025. Currently, the first airborne data acquisitions have been performed.

In the case of Peppermint, airborne data is acquired using a small manned aircraft with broadband thermal sensor (1 spectral band) and VNIR hyperspectral (24 spectral bands in 400-1000 nm spectral range) ensuring 1 m/pix or higher spatial resolution.

In the case of Quinoa, airborne data is acquired using drones equipped with sensors in the field of visible light, near infrared, and in the field of thermal radiation.

Methodology and validation results:

Setting up test fields and ensuring data collection has been the main focus till now. It is expected to start model development, when necessary, reference data on yields at the end of the season will be available (October 2024).

Data preparation is the first data processing step to ensure standardized inputs for models. Currently, in situ sensor data should be downloaded manually. It is planned to explore possibilities for automated data access through the API which is crucial for digital twin concept implementation. Both meteorological and soil sensor data have the same data acquisition frequency, therefore, minimizing the need for data preprocessing. The output of preprocessed in situ sensor data will be a CSV or similar dataset for each field of interest as a point containing data from local meteorological station as well as soil sensor. In the case of airborne data, a field shape will be used to define the extent for airborne raster data. VNIR spectral data will be resampled to thermal data spatial resolution. Soil sensor locations will be used to extract airborne sensor data values for in situ sensor data upscaling within the field of interest.

Availability of reference limits the possibility to use different machine learning models. Therefore, it is planned to start model development based on meteorological and soil sensor data as input features and weekly plant status observations as reference data for the development of a model for field water status assessment. Peppermint and quinoa are relatively different crops; however, both require irrigation which affects the potential yield. It is planned to explore similarities in two different models by extracting similar features to be used as a basis for generic model development which can be easily adapted to a new crop type. Airborne data will be used for model output spatial resolution enhancement from one point to 1 m/pix within field boundary.

The amount of available reference data might not be sufficient for correct validation of the performance; therefore, the second year (2025) data acquisition will focus primarily on validation of data products.

Sensor-integrated data products:

Two sensor-integrated data products are expected as an output – field water status as a vegetation health status assessment and predicted yield as a forecast. It is expected to ensure daily temporal frequency based on meteorological and soil sensor data, but 1 m/pix spatial resolution based on monthly airborne data updates. It is expected that first data products will be ready in Oct 2024.

Sensor-integrated data product	AOI/test sites	Time period	Temporal frequency	Spatial resolution	Accuracy
Field water status (normal / stressed)	Peppermint in Latvia, Quinoa in Israel	Jun – Aug for Peppermint in Latvia, June – Oct for Quinoa in Israel	Daily	1 m/pix	TBD
Predicted yield (t/ha)			Daily	1 m/pix	TBD

Use case(s):

The pilot sites are implemented in collaboration with local farmers who will evaluate the result and provide feedback. In Latvia, Peppermint test fields are implemented in collaboration with the farming company SIA Field and Forest – regional leader in production of medical and aromatical plants. In Israel, the chief cereal guide in the Ministry of Agriculture is deeply involved in designing the experiment and sees it as a pivotal test for advancing Quinoa cultivation in Israel. Several Quinoa growers, mainly in the northern part of Israel, have been informed on the conduction of the experiment and with some we exchange knowledge. The Galilee Agriculture Company is a main stakeholder which is involved in the erection of the experiment and puts in favor of the experiment some of its own resources.

3.1.2 Satellite based field water status and predicted yield for target crop

This product, developed by IES and MIGAL with EO data inputs from DHI and VITO, includes field level target crop water status and potential yield estimates from satellite data (20 m/pix). Two target crops requiring intensive irrigation are chosen – peppermint in Latvia and quinoa in Israel. IES is responsible for peppermint model development but MIGAL for quinoa. Both models will use in situ and airborne sensor data-based model outputs as a reference for training of satellite data-based models which can be further used to screen fields of interest outside the pilot territory.

EO input data:

EO input data is provided by partners DHI and VITO:

EO products	EO data provider	AOI/test sites	Time period	Temporal frequency	Spatial resolution
Evapotranspiration	DHI	Peppermint test fields in Latvia, Quinoa test fields in Israel	Jun-Sep 2024-2025 for Peppermint, June-Oct 2024-2025 for Quinoa	Daily	20 m/pix
Soil moisture	DHI			Daily	20 m/pix
NDVI	VITO			Weekly	20 m/pix
LAI	VITO			Weekly	20 m/pix
PPI	VITO			Weekly	20 m/pix
FAPAR	VITO			Weekly	20 m/pix

Methodology and validation results:

Above mentioned EO data products will be used as an input feature for model development but outputs from in situ and airborne sensor data products as a reference. Reference data will be resampled to EO data spatial resolution – 20 m/pix.

It is planned to test different machine learning approaches for the development of empirical models. Feature significance evaluation will be performed to optimize the necessary input and exclude the less valuable EO data products from the input, thus, also simplifying the model.

The amount of available reference data might not be sufficient for correct validation of the performance; therefore, the second year (2025) data acquisition will focus primarily on validation of data products.

Sensor-integrated data products:

Two sensor-integrated data products are expected as an output – field water status as a vegetation health status assessment and predicted yield as a forecast. It is expected that the first data products will be ready in Nov 2024.

Sensor-integrated data product	AOI/test sites	Time period	Temporal frequency	Spatial resolution	Accuracy
Field water status (normal / stressed)	Peppermint in Latvia, Quinoa in Israel	Jun – Aug for Peppermint	Weekly	20 m/pix	TBD
Predicted yield (t/ha)		in Latvia, June – Oct for Quinoa in Israel	Weekly	20 m/pix	TBD

Use case(s):

The pilot sites are implemented in collaboration with local farmers who will evaluate the result and provide feedback. In Latvia, Peppermint test fields are implemented in collaboration with the farming company SIA Field and Forest – regional leader in production of medical and aromatical plants. In Israel, the Galilee Agriculture Company is collaborating in the implementation of the experiment.

3.2. RIL Crop management

3.2.1 Calculated indicators (aggregates) based only on the ground truth evidence

This product will be produced by Neupublic and will provide an aggregated estimation of the applied agrochemical and water quantities per hectare based on farm-log data that are collected from the pilot parcels. During the two iterations of the project, the aggregation will be implemented in two different scales, as presented in the following figure.

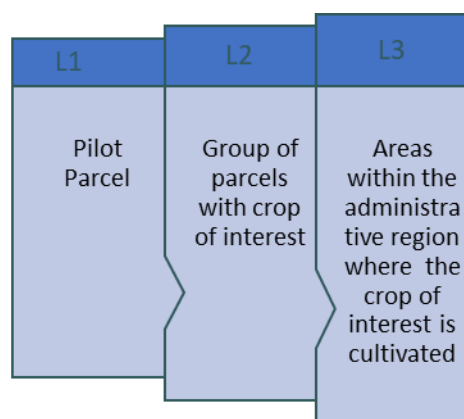


Figure 5: Aggregation levels for the two iterations of the project

In the first iteration, the pilot parcel’s measurements (L1) will be aggregated to the level of group of parcels (L2). During the second iteration, L1 data will be aggregated to the total area within the administrative region where the crop of interest is being cultivated (L3).

In the following table, the 8 selected pilot parcels for 2023 are summarized. They encompass three crop types (potato, tomato and cotton) and are distributed across various Greek communes.

Table 3: Pilot parcel characteristics

Parcel code	Crop type	Area (ha)	Commune
T1	Tomato	7.354	Sofadon
T2	Tomato	5.025	Domokou
T3	Tomato	5.902	Farsalon
C1	Cotton	5.044	Kileler
C2	Cotton	6.743	Kileler
C3	Cotton	2.400	Kileler
P1	Potato	0.203	Oropediou Lasithiou
P2	Potato	0.318	Oropediou Lasithiou

The parcels T1, T2, T3, C1, C2, C3 are located in the broader Thessaly area and their respective commune boundaries are visible in the following map.



Figure 6: Location of pilot parcels T1 - T3, C1 - C3 and their respective commune boundaries

The parcels P1, P2 are located in Crete and their respective commune boundaries are visible in the following map.



Figure 7: Location of pilot parcels P1 – P2 and their respective commune boundaries

Sensor input data:

To develop this product, hourly data measurements were collected from multiple sensors across the various pilot parcels and seasons.

- Meteorological parameters including air temperature, wind speed, solar radiation, rainfall, humidity
- Crop/vegetation properties (e.g. type, height, root-depth)
- Soil properties derived from available soil analysis data (e.g. clay/sand/organic content fraction)
- Irrigation measurements derived from available farm calendars
- Soil moisture at profile from 10cm up to 90cm

In detail, the sensor data are described in the following table:

Sensor data	Source	Data provider	AOI/test sites	Nr. fields/season	Meas. frequency	Season(s)
Leaf Relative Humidity (%) Relative Humidity (%) Soil Moisture 10cm – 90 cm (%) UV Atmospheric Pressure (mbar) Soil Salinity 10cm – 90 cm Average Wind Speed (km/h) Wind Direction Rainfall (mm) Leaf Temperature (°C) External Temperature (°C) Soil Temperature 10cm – 90cm (°C) Leaf Wetness (h) Solar Radiation (wh/m ²)	IoT stations	NP	Thessaly and Crete	8	hourly	2023

EO input data:

EO products	EO data provider	AOI/test sites	Time period	Temporal frequency	Spatial resolution
Sentinel 2 - Level-1C, Level-2A		Four communes in Thessaly, (Tiles 34SEJ , 34SFJ)	Early 2023 – Early 2024	~5 days	10 meters
Sentinel 3 - Level-1B: Top-of-atmosphere radiances		Four communes in Thessaly	Early 2023 – Early 2024	1-2 days	300 meters

ESA WorldCover 10m	ESA - https://viewer.esa-worldcover.org/worldcover/	8 pilot parcel Communes	2020 - 2021		10m resolution
Binary crop classification based on Sentinel 2 Level-2A products	Google Earth Engine	8 pilot parcel Communes	Early 2023 - Early 2024	~5 days	10 meters
Natura 2000 ecological network of protected areas	European Environment Agency	Greece , 8 pilot parcel Communes	2021		Vector polygons

Methodological framework

Our methodological approach uses as inputs in-situ farm data from the pilot parcels to generate related regional agricultural statistics. The workflow also utilizes additional groups of parcels for creating an intermediate aggregate level that will be used for evaluation of the quality of pilot farm data to the regional level. The collection of farm-level input data within a region of interest (ROI), utilizes a Local Administrative Unit (LAU) area (<https://ec.europa.eu/eurostat/web/gisco/geodata/statistical-units/local-administrative-units>) or Commune area (<https://ec.europa.eu/eurostat/web/gisco/geodata/administrative-units/communes>) including the farm calendar data and parcel polygons that fall within the region of interest.

Both pilot parcel data and group parcel data are used to calculate specific indicators per hectare, such as the amount of pesticides used or the yield produced. Additional, regional-level input data for the ROI, including crop type classification binary mask and Earth Observation (EO) data, or other data from external data sources (e.g. Eurostat) are gathered. The calculated farm-level indicators are then aggregated to the regional level based on the regional crop type area, resulting in comprehensive regional indicators such as pesticides used per hectare, irrigation used per hectare, yield per hectare, and financial inputs per hectare. Natura 2000 areas and regional landuse data are also used as additional source of information to provide context to the end users regarding the environmental importance and rural overview of the region of interest.

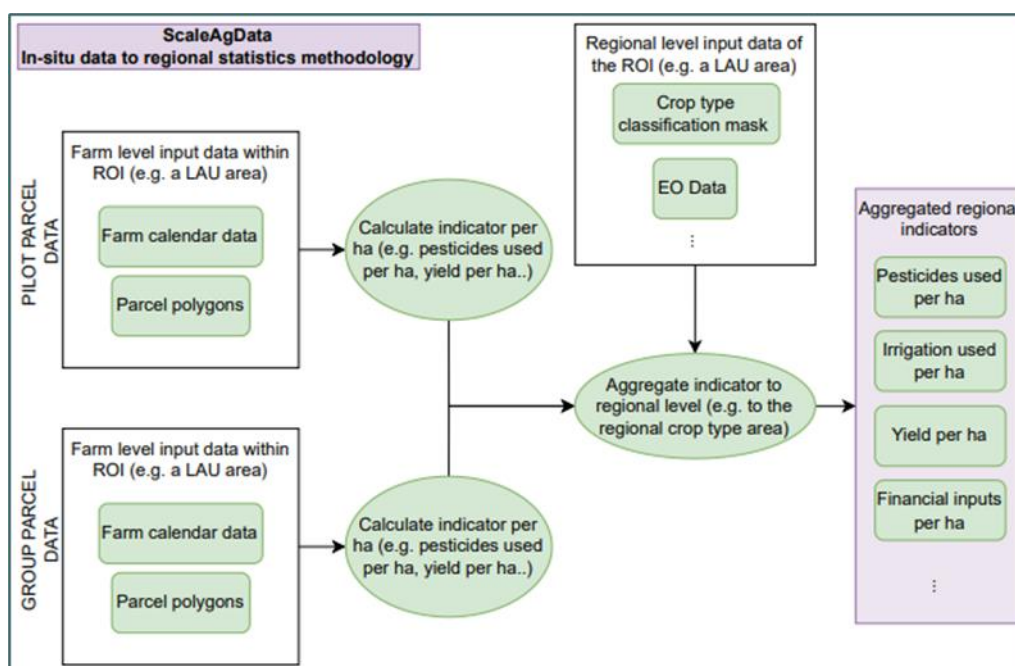


Figure 8: Methodological framework

Sensor-integrated data products:

The different data products that will be generated are shown in the table below. These products are expected to become available at the end of 2024.

Sensor-integrated data product	AOI/test sites	Time period	Temporal frequency	Spatial resolution	Accuracy
Estimation of yield per hectare	8 pilot parcel Communes	Early 2023 – Early 2024	Once, after the end of the growing period (end of 2024 for 1st iteration)	Commune level	Qualitative metrics to be provided after the final outputs
Estimation of pesticide use water quantities per hectare	8 pilot parcel Communes	Early 2023 – Early 2024	Once, after the end of the growing period (end of 2024 for 1st iteration)	Commune level	Qualitative metrics to be provided after the final outputs
Estimation of water use per hectare	8 pilot parcel Communes	Early 2023 – Early 2024	Once, after the end of the growing period (end of 2024 for 1st iteration)	Commune level	Qualitative metrics to be provided after the final outputs
Estimation of financial yield per hectare	8 pilot parcel Communes	Early 2023 – Early 2024	Once, after the end of the growing period (end	Commune level	Qualitative metrics to be provided after the final outputs

			of 2024 for 1st iteration)		
--	--	--	-------------------------------	--	--

Use case(s):

The products will be assessed mainly by farmers, agricultural cooperatives and policy makers and their main goal is to promote the adoption of more efficient and sustainable agricultural practices.

3.2.2 Soil moisture & evapotranspiration aggregation at LAU / Commune Level

These products, developed by Neupublic in collaboration with DHI, include evapotranspiration and soil moisture that are derived from in-situ sensors in selected pilot parcels. The products are designed to use as inputs localized soil moisture information (probe that samples moisture at multiple depth profile from 10cm up to 90m) and weather parameters from IoT sensors and calculate aggregation statistics at the LAU/Commune level, crucial for precision agriculture and land management.

Sensor input data:

The sensor data are similar to Section 3.2.1

EO input data:

The EO data are similar to Section 3.2.1

Methodology and validation results:

The soil moisture mapping product is in development and will use satellite images and precipitation data from early 2023 to early 2024. It is based on the Root Zone Soil Moisture (RZSM) data product produced by DHI and described in detail in section 2.3.2. Time series analyses of Sentinel-2 and Sentinel-3 satellite data are processed with advanced machine learning and physical modeling techniques. The outcome will be a national time series with 20-meter resolution depicting absolute soil moisture in the root zone, covering agricultural land and light-open natural areas. The mapping takes advantage of the frequent satellite overpasses over **four Greek communes in Thessaly**, enabling the calculation of monthly and annual average volumetric soil moisture values, expressed as % water/volume soil.

The methodology maps soil moisture in both the topsoil and the root zone, based on time series data for current evapotranspiration, precipitation, and soil type. The depth of the root zone varies throughout the growing season, so the soil moisture data in the delivered product does not correspond to a fixed depth.

Estimating evapotranspiration at fine high spatial resolution is based on synergistic use of Sentinel 2 and Sentinel 3 satellites’ observations. The actual EvapoTranspiration (ET) data component (dekadal, in mm/day) is the sum of the soil evaporation (E) and canopy transpiration (T). The value of each pixel represents the average daily actual evapotranspiration for that specific dekad.

Sensor-integrated data products:

Sensor-integrated data product	AOI/test sites	Time period	Temporal frequency	Spatial resolution	Accuracy
Estimation evapotranspiration	Thessaly & Crete, Greece	2023	Daily, provided after the end of the growing	20 m resolution	(to be provided after

			period (late 2024 for 1st iteration)		the final outputs)
Estimation of soil moisture	Thessaly & Crete, Greece	1 April – 01 October 2023	Daily, provided after the end of the growing period (late 2024 for 1st iteration)	>20 m resolution	(to be provided after the final outputs)

Use case(s):

The products will be assessed mainly by farmers, agricultural cooperatives and policy makers and their main goal is to promote the adoption of more efficient and sustainable agricultural practices. By providing detailed, evidence-based insights into agricultural inputs and outputs at both the farm and regional levels, the methodology aims to inform decision-making, optimize resource use, monitor and enhance overall agricultural productivity and sustainability at multiple levels.

3.2.3 Calculated indicators (LUKE) based on data assimilation mechanisms along with the respective annotations

The specific products will be developed by Neupublic in collaboration with LUKE aiming at analysis of the variation of nutrients and water within the fields and at the estimation of the crops’ health status and predicted yield.

The pilot parcels that will be utilized for this research are 4 wheat fields located in Northern Greece, in the municipality of Kilkis, with an average field size of 2.0 ha.

Sensor input data:

The sensor data are similar to Section 3.2.1

EO input data:

The EO data are similar to Section 3.2.1

Methodology and validation results:

Yet to be defined, is under discussion.

Use case(s):

The products will be assessed mainly by farmers, agricultural cooperatives and policy makers and their main goal is to promote the adoption of more efficient and sustainable agricultural practices.

3.2.4 Aggregated pesticide use for policy makers

This product is under development and will be produced by Neupublic to provide an aggregated estimation of the applied pesticides based on data that is collected from the pesticide station that is installed within a selected parcel.

Sensor data	Source	Data provider	AOI/test sites	Nr. fields/season	Meas. frequency	Season(s)
Estimation of pesticide use in fields	Pesticide stations	NP	Thessaly	1	Depending on the spraying applications	2024

EO input data:

The EO data are similar to Section 3.2.1

Methodology

Similar to methodology described in Section 3.2.1.

Sensor-integrated data products:

Sensor-integrated data product	AOI/test sites	Time period	Temporal frequency	Spatial resolution	Accuracy
Aggregated pesticide use in fields	Thessaly Greece	2024	during the growing period		

A first version of this product is expected to become available at the end of 2024.

Use case(s):

The products will be assessed mainly by farmers, agricultural cooperatives and policy makers and their main goal is to promote the adoption of more efficient and sustainable agricultural practices.

3.2.5 Calculated sustainability indicators

This product, developed by HORTA and improved in previous research projects, includes several sustainability indicators, which are calculated for the wheat parcels monitored in the ScaleAgData project.

Sensor input data:

Weather data, measured by in-situ weather sensors, as well as soil data and farm logs data entered by the user in the Decision Support System grano.net® (<https://www.horta-srl.it/en/grano-net/>) are used as input for the calculation of sustainability indicators.

Sensor data	Source	Data provider	AOI/test sites	Nr. fields/season	Meas. frequency	Season(s)
Weather data (Temperature, Relative Humidity, Rain)	Weather stations (different models)	Horta	North Italy	10	Hourly	Cropping season 2023-24 is considered in the project

EO input data:

No EO data are currently used to calculate sustainability indicators.

Methodology and validation results:

Weather data measured from in-situ sensors need to be checked for time series completeness and correctness, in order to have data ready to be used as input for the models in the grano.net® Decision Support System (DSS), including sustainability indicators calculation. Horta has an internal quality control procedure, that is automatically applied to data retrieved from the weather stations installed in field, with the main aim to fill in gaps in the data and exclude anomalous values. In order to identify anomalous values, the procedure operates several comparisons among the registered data and weather forecast data, climatological data, historical data of the same weather station, data of nearby weather stations. The comparison is operated for each hourly value retrieved from the weather station. If the data passes all the checks, it is considered validated, and can be part of the time series used as models input. If the data does not pass the checks, the data is considered anomalous, and it is discarded. Discarded, or anomalous data are substituted with forecast data, which are retrieved from an external provider for each of the weather stations.

Sensor data products:

The Decision Support System grano.net® includes a functionality for calculating sustainability indicators, useful to monitor environmental impact of the activities carried out in the fields. The initial set of 20 indicators was developed by HORTA capitalising results from the EU-funded projects PURE and INNOVINE, which were then adapted to wheat crop. The set of indicators available on the DSS was then expanded, and presently includes more indicators, such as the Product Environmental Footprint (PEF) ones.

The initial set of indicators is divided in six domains: Air, Biodiversity, Energy, Human Health, Soil, Water, with group indicators measuring the impact of crop management on each domain. Indicators in the Air compartments are related to the measurement of the amount of greenhouse gases emitted in connection to human activities, and the estimation of the amount of carbon sequestered by plant tissues during the growing season. In the Biodiversity compartment, the indicators evaluate the farm's biodiversity, on the base of the different types of land use and assess the chemical ecosystem hazard score. In the Energy compartment, indicators consider the amount of fuel used for the mechanized operations carried out in the field, the use of fuel from renewable sources, and the farm's waste management. In the Human Health compartment, indicators evaluate the chemical's hazard to humans, the exposure of individuals to chemical products and the surface treated with pesticides. In the Soil compartment, indicators estimate the risk of soil compaction; the loss of soil due to water-caused erosion; the duration of soil coverage; the percentage of organic matter contained in the soil; evaluate land surface necessary to provide the resources. In the Water compartment, indicators assess the water footprint; evaluates the type of irrigation system and water used for crop irrigation; the emissions of compounds causing acid rains and the effect of excess of nutrients on water ecosystems. The Product Environmental Footprint indicators assess in a multi-criteria perspective the environmental performance of agricultural activities, throughout the life cycle approach (https://green-business.ec.europa.eu/environmental-footprint-methods-0_en). The indicators assess the global warming potential; the ozone depletion potential; the impact on human health and ecosystem; the effect on freshwater and marine; impact on soil quality; and it evaluates resource depletion. For each compartment, the recommendation for the farmer to decrease the environmental impact of the cultivation phase, choosing among the available cultivation techniques, the ones that lead to a lower impact on environment.

Sustainability indicators are presented to the user in both a graphical way, using a radar graph (Figure 10) and in the form of a table. The graph allows the user to visually understand his sustainability performances and to individuate the compartments in which crop management need to be improved.



Figure 9: Visualisation of sustainability indicators in the DSS grano.net(R)

Sustainability indicators are calculated for each of the parcels monitored in the RI Sub Lab 2b. The indicators address the environmental sustainability, especially the domains of Air, Water, Soil Energy, Biodiversity and Health.

Sensor-integrated data product	AOI/test sites	Time period	Temporal frequency	Spatial resolution
Sustainability indicators	10 parcels cropped with wheat in North Italy	Wheat cropping season 2023-24	Calculation is done once, at the end of the cropping season.	Each monitored parcel relies on data of a nearby weather station

Use case(s):

In the RI Sub Lab 2b, sustainability indexes will be calculated for the parcels monitored in the SubLab activities. Field managers (farmers, technicians, agronomists, can visualise the sustainability performance for each parcel, and then decide to take action to adapt the crop management in order to improve sustainability.

3.2.6 DSS model outputs

The Decision Support System (DSS) grano.net[®] was developed by HORTA in previous years. The DSS helps farmers in the sustainable management of their wheat crops, allowing them to take informed

decisions for all the main steps in the crop management. The DSS is used for the monitoring of wheat parcels in the ScaleAgData project.

The DSS has available several models and algorithms providing, among the others, indications on fertilization, water balance, disease risk, and crop yield and quality prediction. The models in the DSS are run for the parcels selected for monitoring in the RI Sub Lab 2b. On the base of the disease risk prediction, and crop yield and quality model outputs for the monitored parcels, an upscaling will be done for a larger area.

EO data for selected parcels are developed by project partners DHI and VITO.

Sensor input data:

Weather data, measured by in-situ weather sensors, as well as soil data and farm logs data entered by the user of the Decision Support System grano.net® are used as input for the calculation of the models and algorithms in the DSS.

Sensor data	Source	Data provider	AOI/test sites	Nr. fields/season	Meas. frequency	Season(s)
Weather data (Temperature, Relative Humidity, Rain, Leaf wetness)	Weather stations (different models)	Horta	North Italy	10	Hourly	Cropping season 2023-24 is considered in the project

EO input data:

EO products	EO data provider	AOI/test sites	Time period	Temporal frequency	Spatial resolution
NDVI	External data provider	10 parcels cropped with wheat in North Italy	Cropping season 2023-24 is considered in the project	3-7 days	20m
ET, SM (yet to be retrieved for current season)	DHI	10 parcels cropped with wheat in North Italy	Cropping season 2023-24 is considered in the project	Daily	20m
Yield potential index (yet to be retrieved for current season)	VITO	10 parcels cropped with wheat in North Italy	Cropping season 2023-24 is considered in the project	Once in the season	20m

Methodology and validation results:

Weather data measured from in-situ sensors are checked according to the procedure described in section 3.2.5.

Models and algorithms in the DSS use weather data, soil data and farm log data for their running. The work needed to include EO information (i.e. NDVI indexes) as model input is ongoing, taking into

consideration the existing wheat yield model, which is already implemented in the DSS. EO vegetation indexes can provide the model with near time vegetation information, allowing to improve the model outputs.

Sensor-integrated data products:

DSS models are calculated for each of the parcels monitored in the RI Sub Lab 2b. The models and algorithms are aimed to support the farmer in making informed decisions about wheat crop management.

Sensor-integrated data product	AOI/test sites	Time period	Temporal frequency	Spatial resolution
DSS models output for yield, diseases, fertilisation, water balance.	10 parcels cropped with wheat in North Italy	Wheat cropping season 2023-24	Calculation is done during the cropping season, depending on the type of model or algorithm.	Each monitored parcel relies on data of a nearby weather station, soil data and farm log input from the

Use case(s):

In the RI Sub Lab 2b, DSS will run for the parcels monitored in the SubLab activities. Field managers (farmers, technicians, agronomists) can receive support from the DSS for the management of their wheat crops. The DSS support their decisions regarding the main operations to be performed in field (i.e. fertilisation, plant protection interventions against diseases), and allows to have information about the water content in soil and forecasts on the crop yield.

3.2.7 Statistical data on the accuracy of observations of the occurrence of agrophages

The objectives of data processing services involve enhancing situational awareness through the organization and acquisition of plant inspection data related to diseases and pests. The primary recipient of this information is the agricultural advisor, who, in the role of the monitoring coordinator, receives a daily risk review, ensuring they remain well-informed about potential threats.

Sensor input data:

The main source of data, which is also the validation set, is the data set regularly acquired from the mobile expert application that is directly documenting the occurrence of pests. This process is coordinated and takes place in a structured form. Data are collected periodically (at least once a week) at points designated by specialists. The points are designated in the web application (eDWIN eODR), where the specialist can conveniently indicate a point e.g. based on the plot number, while the data collection itself, thanks to data exchange via API is possible in the dedicated eDWIN Agrophages mobile application (<https://play.google.com/store/apps/details?id=pl.edwin.agrofagi&hl=pl>)

The set of data collected by the mobile expert application, enriched with other measurement data from physical sensors, is the basis for statistical analysis.

The map below shows the test fields located in Wielkopolska district that are managed by WODR.

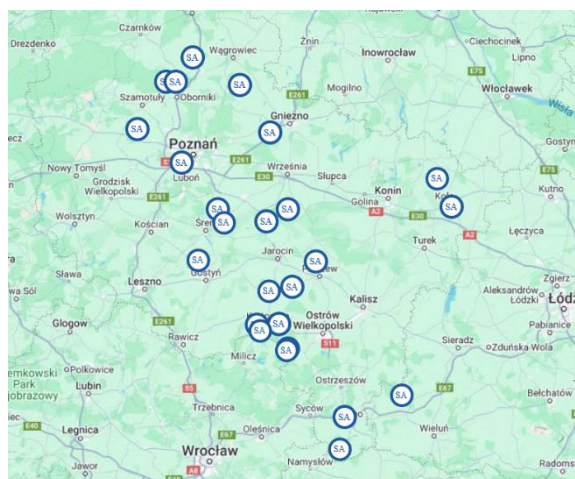


Figure 10: Location of the test fields in Wielkopolska district

In selected seasons these parcels were cultivated with Corn. Observation data of Corn plants on test-fields is used as reference data.

Sensor data	Source	Data provider	AOI/test sites	Nr. fields/season	Meas. frequency	Season(s)
Meteorological data	eDWIN platform	PSNC/WODR	~100	3	hourly	2021/22/23

EO input data:

No EO data is being used in this case.

Methodology and validation results:

The methodology will be consistent with previously carried out implementations and launches of DSS models.

That is, a compilation of data collected locally in the fields and the results of disease models will be used. If these two data sets coincide, it means that a monitoring point is valuable and should be the focus of further observations in support of the pest model indications.

Each statistical tool will be based on physical validation of the result, broadly describing the measured phenomenon. The data from statistical tools will indicate areas that do not show a discrepancy between the locally measured data set and the disease model results set.

Data differ over the available seasons due to different strategies in local plant production market.

The field observation report should include the following parameters:

- name of the pest or disease,
- BBCH development phase of the plant
- development phase of the pest
- percentage of colonization by the pest
- the severity of the pest's occurrence
- optional notes
- report date
- information whether chemical treatment is required.

Sensor data products:

The data product will combine meteorological statistics with physical observations, synthesize it spatially and create simple and useful DSS module with relevant representation on the user interface, which can indicate the accuracy of local indications of pest or disease occurrence compared to previously validated pest or disease models.

Sensor-integrated data product	AOI/test sites	Time period	Temporal frequency	Spatial resolution	Accuracy
Accuracy of observations	20 parcels	1 season	Weekly/daily		

The expected results should be ready to test in release at the end of the first iteration round (M24).

Use case(s):

The data products will be available as micro services ready to be deployed locally. Each test is organized in expert's applications, presenting “experimental data” in experts module concerning pests monitoring. Application administrators or granted with proper rights experts are accessing and accepting results of implemented tools.

3.2.8 Improved predicted aphorage occurrence data based on geolocation

The main issue with the solution presented in the previous section is that the results are calculated for the nearest meteorological station, not the specific chosen field. In this improved approach, sensor data is interpolated, allowing for data retrieval from any chosen point. Additionally, user reports and treatments are incorporated, along with NDVI (Normalized Difference Vegetation Index), to achieve more precise results.



Figure 11: On the left image rye test fields and on the right image sugar beet test fields

Data acquired from test fields are diverse over the available seasons due to different strategies in local plant production context.

Sensor input data:

The data product will use two types of sensor-based information with a common timespan and located in a limited radius (area of interest):

Sensor data	Source	Data provider	AOI/test sites	Nr. season	Meas. frequency	Season(s)
Meteorological data	eDWIN platform	PSNC/WODR	~100	3	hourly	2021/22/23
Phenological observations	eDWIN platform	PSNC/WODR	8	1	daily	2024

EO input data:

The data product will use two types of EO data. The NDVI is calculated from Sentinel-2A and Sentinel-2B images using data from Band 4 and Band 8. The EO product named RainGRS is a long-term multi-source precipitation estimation with high resolution (reference link: <https://amt.copernicus.org/articles/16/4067/2023/amt-16-4067-2023.pdf>) creating a possibility to upscale calculations from selected points to complete areas of interest.

EO products	EO data provider	AOI/test sites	Time period	Temporal frequency	Spatial resolution
NDVI	PSNC	29 parcels cropped with wheat, beet and corn in West Poland	From seasons 2021/22/23	5 days	10m
RainGRS	Open Data (IMGW-PIB)	selected area	From seasons 2021/22/23	hourly	1km

System data:

Users have the possibility to report agrophages observed on the fields and to create a registry of performed protection treatments. Collecting the closest reports and treatments for an observation point gives information about an agrophage that could potentially be a threat to the point. This data is added by users and is not verified, so it may not be correct. The range is determined by the type of the agrophage.

System data	Source	Data provider	AOI/test sites	Seasons
Spraying treatments	eDWIN platform	PSNC	depends on the type of agrophage	2021/22/23
Agrophage reports	eDWIN platform	PSNC	depends on the type of agrophage	2021/22/23

Methodology and validation results:

The process begins with the interpolation of the input data, followed by data fusion. Machine learning is then used to interpret the results, obtaining a percentage of an agrophage occurrence probability.

The Kriging algorithm with a trend is used to interpolate temperature and humidity. Agrophages reports and protective treatments added in the system by users are point data, so they cannot be interpolated. The fusion uses the range and the conducive conditions to the development and spread of the agrophage on which the notification originated, or the protective treatment was performed. The validation set are the physical observations made by the advisors on the selected 29 fields.

Sensor-integrated data products:

A risk evaluation model will be created to anticipate risks, which will make it possible to get ahead of the validation stage.

Sensor-integrated data product	AOI/test sites	Time period	Temporal frequency	Spatial resolution	Accuracy
Probability of an agrophage occurrence	29 parcels cropped with wheat, beet and corn in West Poland	In growth season	daily	4km	-

The expected results should be ready to test in release at the end of the first iteration round (M24).

Use case(s):

The data products will be available as micro services ready to be deployed locally. Tests are organized in experts applications, presenting “experimental data” in experts module concerning pests monitoring. Application administrator or granted with proper rights expert are accessing and accepting results of implemented tools. Advisors receive daily risk assessments with a specific focus on pest occurrences within their designated work areas, facilitating targeted responses. Farmers benefit from this system by receiving precise information about threats in their locality, enabling timely and effective action.

3.2.9 Predicted overall level of agrophage occurrence risk for the selected region

This case represents the highest level of abstracting available data into a single conclusion.

Input data

Case will combine results of (3.2.7) Statistical data on the accuracy of observations of the occurrence of agrophages and (3.2.8) Improved predicted agrophage occurrence data based on geolocation.

Sensor-integrated data products:

The given product will be the estimated level of risk for the entire voivodeship (the highest-level administrative division in Poland) or powiat (one level lower) in the scope of a selected disease in a specific crop.

It will be streamed into an alert that can be published on any interface or distributed between systems. Alerts will be created with a minimum weekly frequency.

The expected results should be ready to test in release during the second iteration round (M29).

Use case(s):

The first use case will be the user interface of the farm management application. Desirably this data product would be presented in UI applications in the form of a map where each high-level administrative unit has data regarding current estimated conditions. Managing and funding entities gain access to statistical data and reports, which are crucial for making inferences and supervisory decisions. This tool has comprehensive approach ensures that all stakeholders have minimal information to manage plants health decisions.

3.3. RIL Yield monitoring

3.3.1 Potato yield estimates

This product, developed by VITO in collaboration with AVR, currently includes subfield-level (50-70m resolution) potato yield estimates, but will be extended to yield estimates at pixel (10m), field and regional level, at harvest time (no predictions yet).

Sensor input data:

Gross yield sensor data from is collected using AVR’s Puma 4.0 harvesters.

Sensor data	Source	Data provider	AOI/test sites	Nr. fields/season	Meas. frequency	Nr. training data
Potato yield (ton/ha)*	Puma 4.0 harvester	AVR	Belgium, Netherlands	+/- 2000	Once, at harvest	1135 Fields, 14075 subfields

Currently, field boundary data is not available across all jurisdictions. As such, fields are automatically calculated based on the harvester data, which is received in CSV format and contains info on yield collection time and location. The harvesters use an internal calculation to return a yield in kg/ha, based on the area of each collection point. The algorithm for creating field boundaries does so by iterating through the data points, grouping them until no further points are located within 250m of the current grouping. It will then add this grouping as a field and move on to the next points. These fields are divided into 70m x 70m subfields, with areas within 10m of an infrastructure/river intersection removed, ensuring polygons still maintain an internal area of 50m x 50m.

In this first stage, we focused on Belgium and the Netherlands. We used harvest data collected in 2022. In future stages, data will be extended to include sites from several other European nations, as well as different years. The inclusion is dependent on provision by suppliers of both harvester and EO input data for these various regions and timeframes. The dataset is constrained to only values between 10000 and 120000, as anything outside of this is considered an outlier. The below figure shows the distribution of values across the dataset.

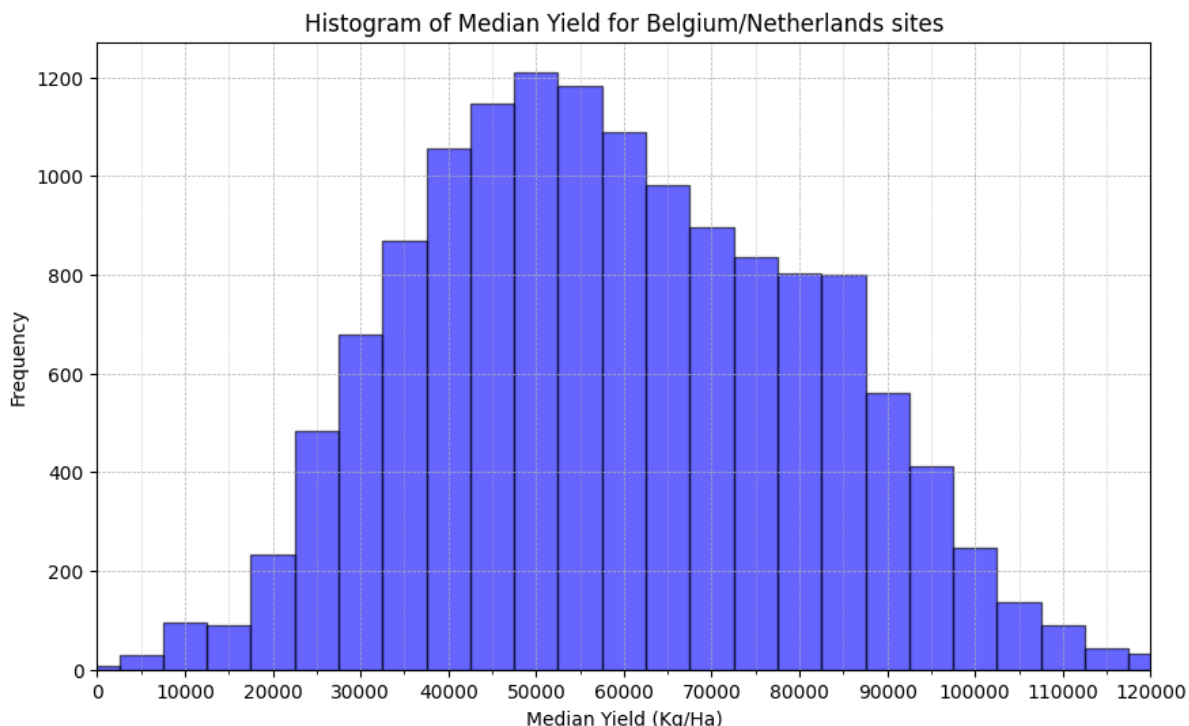


Figure 12: Histogram showing frequency of yield values within Belgium and Netherlands subfields. Data is cut to 120000 to remove impact of large outliers.

Current Study Region

Nation	Fields	Sub Fields	Harvester Points (Average)	Year
Belgium	782	9413	800 per subfield	2022
Netherlands	355	4662	1192 per subfield	2022

More details on the sensor data can be found in deliverable D3.2.

EO input data:

The following EO data are used as input for potato yield modeling:

EO products	EO data provider	AOI/test sites	Time period	Temporal frequency	Spatial resolution resolution
Sentinel-1 and Sentinel-2 imagery	VITO	Belgium, Netherlands	2022	10-Daily composite	10m (subfield aggregated)
AgERA5 meteo (temperature, rainfall)	VITO	Belgium, Netherlands	2022	10-Daily composite	10m (subfield aggregated)
Sentinel-1 and Sentinel-2 fused products (CropSAR Fapar)	VITO	Belgium, Netherlands	2022	Daily	Field level
Actual evapotranspiration	DHI	Sentinel Tile 31UFS	2022	Daily	15m x 25m (subfield aggregated)

Soil moisture	DHI	Sentinel Tile 31UFS	2022	Daily	15m x 25m (subfield aggregated)
---------------	-----	------------------------	------	-------	---------------------------------------

More details on EO data, source and processing can be found in deliverable D3.3.

Methodology and validation results

Data Preparation

As data sources are collected at different temporal and spatial resolutions, they must be processed to a common timeseries format suitable for modelling. The actual timeseries can be user-defined, which will both restrict the start and end of the timeseries based on the desired inference period, and define the temporal resolution. Timeseries variables are first adjusted to fit the correct temporal resolution by compositing higher resolution data into a standard interval (e.g., 10-day periods). This ensures that all variables align to the same temporal scale. Subsequently, the timeseries data are interpolated using nearest-neighbour matching to synchronize the exact temporal points across all variables. For example, even if one dataset originally covers a period from day 1 to day 10 and another from day 2 to day 11, the interpolation process adjusts these periods so that all datasets match precisely to cover consistent intervals, such as from day 1 to day 10. For static variables (e.g. elevation, latitude, longitude), these are simply extrapolated across the length of the timeseries.

These datasets are individually saved and can then be used for testing with several model architectures, or used to create feature embeddings for ML models. Datasets are to be stored and versioned in a Data Versioning Control (DVC) system.

Feature Extraction

A challenge of using multiple sensors across a timeseries, and having relatively limited sample sizes, is that the total input features can become too large for model architectures to handle well. Whilst the architecture will create its own embeddings before outputting to the final regression or classification layer, these may be suboptimal. “Feature embeddings” aim to first extract useful features, which can then be used as lower-dimensionality model inputs. In the case of yield estimation, PRESTO (Section 2.3.1) is used to extract usable features from the sentinel and meteo timeseries, transforming them into 128 feature embeddings. For other input data sources not handled within embeddings, such as soil moisture and evapotranspiration, other feature extraction methods should be adopted to similarly compress this data to key features.

Model Framework

As several data sources and feature extraction methods are combined in the context of yield estimation, the model framework is versatile to either timeseries inputs, or static features (embeddings), with the user being provided options to develop and compare different models and their parametrization. For time series models, a 1D-CNN model is selected to use as a baseline. This model has been found to better handle the dimensionality issues of the dataset when used without feature engineering (many timeseries), maintaining performance even when using reduced samples (e.g. focusing on the 31UFS tile with soil moisture/evapotranspiration). It is also suitable for multiple tasks, including regression and categorical classification, allowing users to also model ordinal categories.

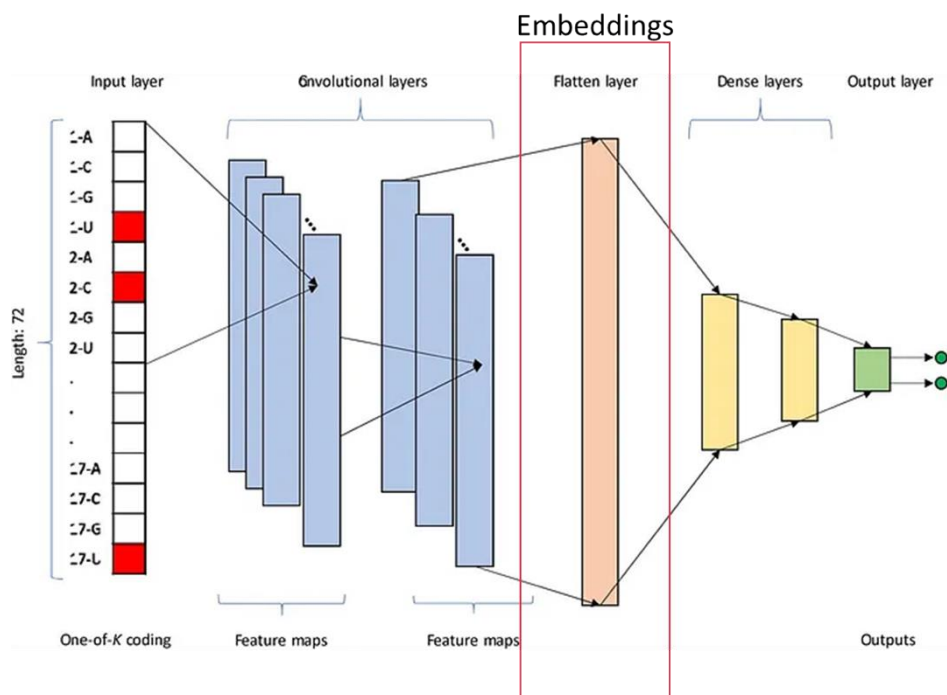


Figure 13: Model framework for yield estimation

Essentially, the initial 1D-CNN convolutions will, once flattened, produce embeddings, similar to the embeddings created during feature extraction. From this point it is possible to attach different model heads, depending on user requirements.

Initial models have been trained from scratch for each dataset. Moving forward, an ensemble of models trained on these datasets will be used. Subfields where the model ensemble is in agreement in most cases, surpassing a voting threshold, will be selected and used as high-confidence predictions. These confident predictions will be used as the basis for estimations at field and regional levels. To understand the within-field differences and account for less-confident subfields, second-stage models will be utilized, incorporating products such as cropSAR 2D, which measures pixel-level vegetation indices (such as NDVI, FAPAR, and plant cover) across growing seasons. In addition to estimation at field or higher level, predictions can also be upscaled to the highest available data resolution, through spatial interpolation of lower-resolution values.

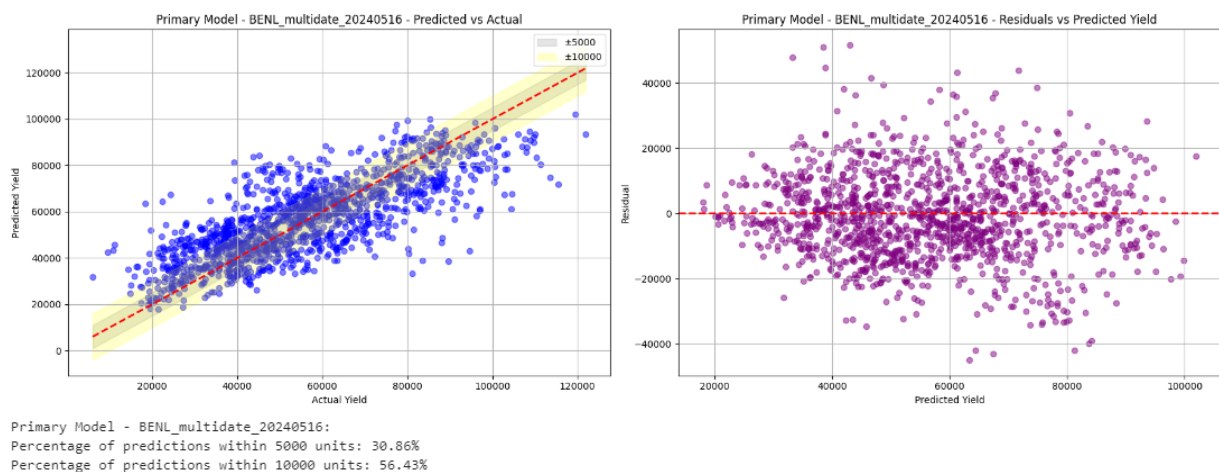
Results: Subfield-Level Monitoring

Model	Dataset	Model Embeddings	RMSE Full Dataset (kg/Ha)	R2 Full Dataset	RMSE 31UFS dataset (Kg/Ha)	R2 31UFS Dataset
1D-CNN	Belgium/Netherlands Field	Sentinel-1/Sentinel-2 Imagery, Meteo Data	13382.34	0.59	11715.17	0.47

1D-CNN	31UFS Tile	Sentinel-1/Sentinel-2 Imagery, Meteo Data, Soil Moisture, Evapotranspiration	-	-	11396.07	0.50
CatBoost	Belgium/Netherlands Fields	PRESTO (Sentinel-1/Sentinel-2 Imagery, Meteo data)	12111.90	0.66	11530.97	0.48

**NB: Pixel-Level monitoring utilises 50-70m2 fields, to account for differing data resolutions.*

Several models were trained, utilizing both the 1D-CNN embeddings, and generated PRESTO features. Models were trained using the same training/test split in each instance, making it possible to compare performance in several instances. For the full dataset, including all sites in the Netherlands and Belgium, the model utilising PRESTO embeddings with a CatBoost head outperformed a 1D-CNN model incorporating the raw input. For the reduced dataset, where soil moisture and evapotranspiration data were available, RMSE was reduced as compared to models trained without it. An investigation of the poor performing samples showed that the effectiveness of prediction of various samples can change with different models. Focusing on the 31UFS tile, 17% of samples were predicted at residuals greater than 10% (9804.77 kg/Ha) of the dataset range in all the models, increasing to 35.5% when lowering the threshold to 5% (4902.38 Kg/Ha). In other cases, at least one of the two 1D-CNN models, or the PRESTO-embeddings with CatBoost head were able to predict within the threshold. Due to the inherent uncertainty in yield estimation, an ensemble model in future could allow for more accurate predictions, and an estimate of the uncertainty produced.



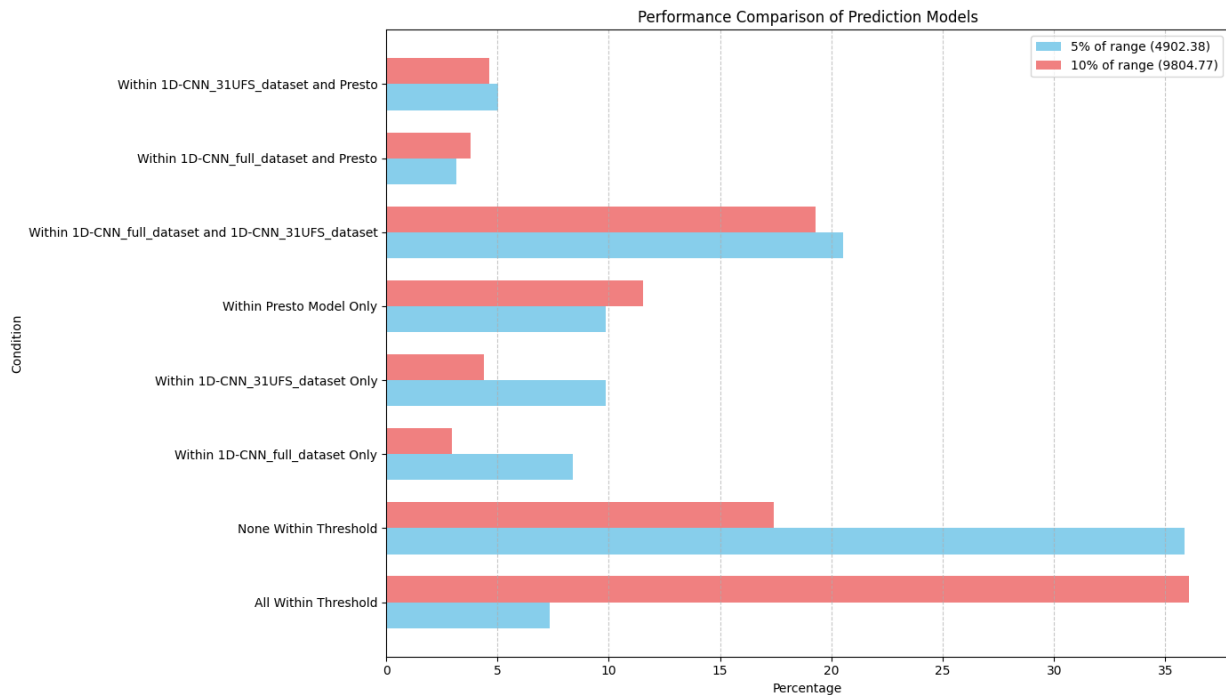


Figure 14: (Top Left) Distribution of actual vs. Predicted yield (kg/Ha) for test set of BE/NL, using 1D-CNN model on full dataset (Top Right) Residual plot of yield values using 1D-CNN model on test set (Bottom) Comparison of prediction models in terms of the number of predictions within 5% and 10% of the dataset range (98047) of ground truth median yield (in Kg/Ha), showing which combinations of models correctly predict which percent of values.

Limitations and Future Work

Initial models show ability towards identifying and differentiating different levels of yield based on input sensor data, though the applicability of the model is limited due to high levels of uncertainty. Future model iterations will need to account for this, so that yield estimates can be provided, as well as confidence in these estimates on a pixel, field, and regional basis. Using PRESTO feature extraction improves the results compared to the raw satellite inputs. Such methods should also be applied to other sensor data to allow for further model improvements.

Initial assessment of the model and data distribution shows that there is a strong spatial autocorrelation of data. Typically, pixels within a field show similar values, as do fields within a region (shown in the graph below, where histograms typically have different peaks, as do median values, which are similar in a local area). Future iterations and up-sampled predictions should account for this, using the location of a sample to improve estimation. The below image displays the histograms of median yield values per field for each region across the dataset, demonstrating that nearby sites typically have similar values. Such information is useful to both improve pixel-level estimation, as well as stabilize model expectations when predicting at field and regional level.

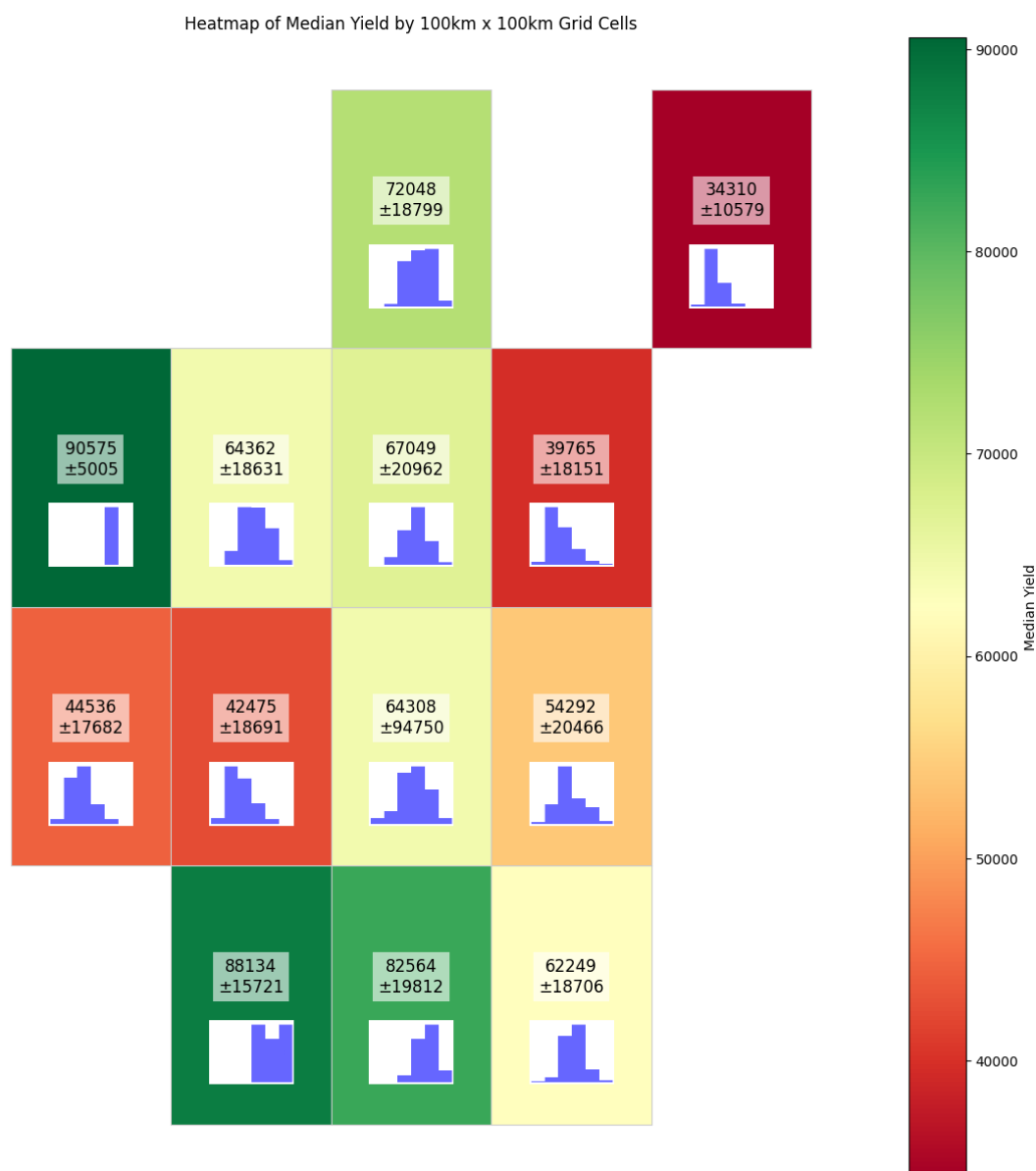


Figure 15: Histograms of median yield values per field for each region across the dataset

Sensor-integrated data products:

Sensor-integrated data product	AOI/test sites	Time period	Temporal frequency	Spatial resolution	Accuracy
Potato yield at subfield level	Belgium, Netherlands	2022	once, at harvest	50-70m	N/A*

*Confidence metric required to provide users with accurate estimations

Use case(s):

Subfield (and future pixel)-level yield estimates are showing variations of crop yield within a field. After diagnosing the causes of the observed yield variations, farmers (whether or not in collaboration with advisors, researchers...) can use the maps to optimize their field practices, e.g., by applying variable rate fertilization.

The maps will also be useful for insurance companies, for damage assessment or for assessing historical performance of their customers’ fields. For this reason, ScaleAgData partner AgroInsurance will also test and evaluate the yield maps.

Further, the maps could also be of interest to input suppliers, such as seed companies, to select homogeneous fields for trials or seed multiplication.

3.3.2 Improved tare yield estimates for potatoes

This product, developed by Ugent in collaboration with AVR, includes a pixel-level tare estimate, based on RGB images (resolution to be determined) or on EO data alone (20m resolution).

Currently, the tare weight of the potato harvest, i.e. the soil / dirt attached to the tubers and the harvester’s conveyor belt, is estimated by sight, without any data input. Tare weight is thought to be dependent on soil type and moisture.

In autumn 2024, an experiment is planned on three fields with different soil types to improve the tare estimate based on either RGB images from a camera, installed in harvesters, or EO data (soil moisture estimates). In-situ measurements of soil moisture will be used to train machine-learning models to calculate soil moisture based on the RGB images and to validate EO estimates of soil moisture. Soil moisture sensors installed in one of the fields can complement the dataset.

In-situ measurements will be performed at harvest by pausing the harvester at several points in the field during harvest. The soil attached to the potatoes and the conveyor belt will be collected to measure the soil weight and moisture. Soil samples on the field can also be taken to increase the amount of soil moisture data.

Sensor input data:

RGB camera installed in AVR’s Puma 4.0 harvesters. Soil moisture sensors still to be determined.

Sensor data	Data provider	AOI/test sites	Nr. Fields/season	Meas. frequency	Season(s)	Nr. training data
RGB camera	AVR	Belgium	3	Once, at harvest	2024	3 fields* 20 subfields
Soil moisture sensors	ILVO	Belgium	1	Continuous (daily?)	2024	1-4 sensors
Soil scan	ILVO	Belgium	1	Once, at harvest	2024	Complete field?

EO input data:

EO products	EO data provider	AOI/test sites	Time period	Temporal frequency	Spatial resolution
Soil moisture	DHI	Sentinel Tile 31UFS	2024	Harvest day + days after	15m x 25m (subfield aggregated)

Other data

Data		Data provider	AOI/test sites	Time period	Temporal frequency	Spatial resolution
In-situ measurement	soil moisture	UGent	Belgium	2024	Harvest day	20 per field?
	Tare weight	UGent	Belgium	2024	Harvest day	20 per field?

Methodology and validation results:

The tare estimation is a two-step process. First, soil moisture needs to be estimated. This will be done based on the soil samples that will be scraped from the potatoes and the conveyor belt. The “calculated soil moisture” that will be obtained this way will be used to calibrate a machine learning model to estimate soil moisture from the RGB images that are taken by the camera installed in the harvester.

Soil moisture estimates will also be calculated from EO data (see methodological framework described in section 2.3.2). The in-situ soil moisture measurements, both from the scraped potatoes and from the field itself, can be used to validate these EO estimates.

The second step is estimating the tare weight. This can be done with the RGB images as an input (whether soil moisture is needed as an additional input is to be tested) and the measured tare weight as calibration data. Again, machine learning will be employed for this step. For farmers without a camera mounted on the harvester, it will also be tested whether the soil moisture from EO data is enough to estimate tare weight. The in-situ measured tare weight will be again used as calibration data.

Which machine-learning models will be used will be determined after the completion of the dataset.

Sensor-integrated data products:

Sensor-integrated data product	AOI/test sites	Time period	Temporal frequency	Spatial resolution	Accuracy
Tare estimate based on RGB camera	Belgium	Summer 2024	Once, at harvest	To be determined, dependent on frequency of RGB imaging on field	N/A
Tare estimate based on EO data	Belgium	Summer 2024	Once, at harvest	15m x 25m	N/A

The resulting data product will be derived using two different models. One that estimates the tare weight (pixel-level) based on RGB images of the potato conveyor belt inside the harvester. However, this requires the presence of such a camera inside a harvester. For farmers that do not have such a harvester, a second model will be available that estimates the tare weight based on EO derived soil moisture. The downside of this model will be the lower resolution and lower accuracy.

During the first iteration round, as we are still in the development phase, no specific use cases have been planned yet, nor any validation experiments with AVR customers.

One of the limitations will probably be the soil type for which the models were calibrated. They will therefore only be accurate for those soil types. Model validation is necessary to determine the accuracy for different soil types.

Use case(s):

If an accurate model can be developed, this can be used to correct the harvester yield data and provide more accurate yield estimates to the farmers.

3.3.3 Winter wheat yield estimates (LUKE)

This data product, developed by LUKE, in collaboration with UGent and CNHi, focuses on simulating spatial variation in yield estimates of winter wheat. Yield sensors that are installed on harvesters are often prone to inaccuracies resulting in incomplete yield maps. The calibration of the sensors is not always very accurate, making it difficult to compare data collected from different sensors. When combining data from multiple harvesters/sensors operating on a single field, this results in erroneous yield maps.

LUKE is developing a digital twin in collaboration with UGent, CNH and VITO to model the behavior of winter wheat fields. The aim of using a Digital Twin model is to simulate wheat yield and protein content variation within a field (in a hexagonal grid) based on the spatial variation in EO data. Missing data in yield maps provided by the harvester sensors can then be estimated based on the yield variation estimates from the digital twin model. Modeled yield forecasts, as well as crop and nutrient status during the growing season can support variable rate fertilization application. More detailed information can be found in deliverable D4.2.

During the first iteration round of the ScaleAgData project, the development was targeted to setting up the digital twin’s modelling framework (see Chapter 2). Aim is to automatize digital twin set-up as far as possible for the selected use cases. During the time when data from the lab was not available, development and testing utilized LUKE’s test field data collected in Finland.

Sensor input data:

Data of two seasons will be used to set up the digital twin. A limited dataset of eight fields in Flanders (region around Nijvel/Nivelle, see map below) monitored in 2023 will be used to test the model first. An extensive dataset of four fields (same region, see map below) monitored in spring/summer 2024 is being gathered to calibrate the model for different winter wheat varieties. More repetitions of sensor data are planned for this dataset.

General data sources including sensor input data are also listed in deliverable D4.2 section 2.3.

Sensor data	Source	Data provider	AOI/test sites	Nr. fields/season	Meas. frequency	Season(s)
Yield		CNHi	Belgium	8 fields	once	2023
Yield		CNHi	Belgium	4 fields	once	2024
VI	Augmenta	CNHi	Belgium	8 fields	once	2023
VI	Augmenta	CNHi	Belgium	4 fields	3 times during season	2024
VI	(drone)	UGent	Belgium	4 fields	Twice during season?	2024
Soil EC		CNHi	Belgium	8 fields	once	2023
Soil EC		CNHi	Belgium	4 fields	3 times during season	2024

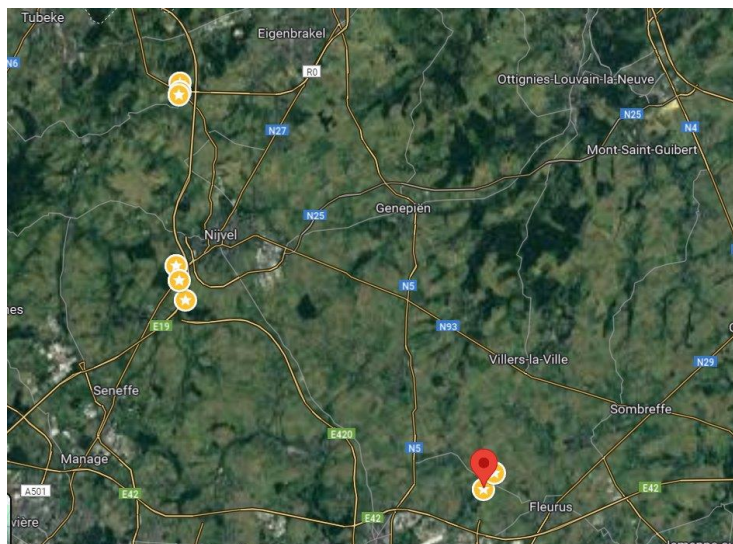


Figure 16: Fields monitored in 2023

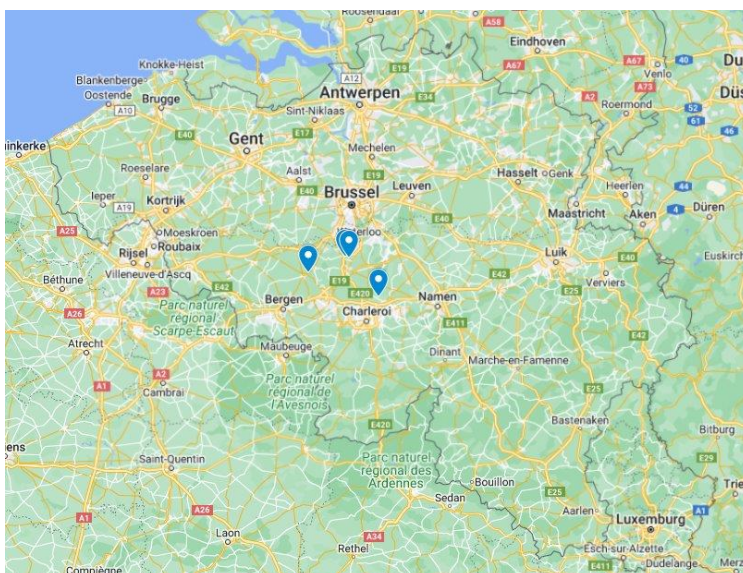


Figure 17: Fields monitored in 2024

EO input data:

Below are listed the EO data that was gathered for the 2023 season. The same dataset will be gathered for the 2024 season after the harvest.

General data sources including EO input data are also listed in the deliverable D4.2 section 2.3.

EO products	EO data provider	AOI/test sites	Time period	Temporal frequency	Spatial resolution
Cropsar index	VITO	Belgium (8 fields)	Jan 1 – Aug 14, 2023	Every 5 days	10m x 10m
Evapotranspiration	DHI	Sentinel Tile 31UFS	Sep 2022 - Sep 2023	Every day	15m x 25m
Soil moisture	DHI	Sentinel Tile 31UFS	Sep 2022 - Sep 2023	Every day	15m x 25m

Other input data:

Additional data is being gathered to improve the model. Weather station for the 2023 season is already available. The same data will be gathered for the 2024 season after harvest. Other data for 2023 and 2024 (that is already available) is also listed below.

Data		Data provider	AOI/test sites	Nr. fields/season	Meas. frequency	Season(s)
Weather station	Min T, max T, average T, RH, Wind speed, Precipitation, Radiation	KMI	Belgium	8 fields	daily	2023
Management data	Cultivar, sowing date, harvest date	CNHi	Belgium	8 fields		2023
Management data	Cultivar, sowing date, fertilisation timing and amount	CNHi	Belgium	4 fields		2024
In-situ measurements	Soil moisture, soil nitrogen, plant length, plant fresh weight, plant dry weight, plant nitrogen, leaf chlorophyll, anthocyanin, flavonoid content and nitrogen balance index, LAI	UGent	Belgium	4 fields, 6 subfields	4 times during season	2024

Methodology and validation results:

To simulate daily status of a field, yield estimates and within-field variation, the sensor and EO input data are entered into process-based and machine learning models. Data assimilation described in section 2.2 updates observed EO input data into models. Pure biophysical process-based cropping system models are compared with machine learning and hybrid models, where process-based and machine learning approaches are combined. Flexible use of various modelling approaches and easy change of the simulated field is enabled with digital twin data model. See deliverable D4.2 section 2 for further description of the methodology.

Sensor-integrated data products:

Sensor-integrated data product	AOI/test sites	Time period	Temporal frequency	Spatial resolution	Accuracy
Winter wheat yield estimates (field level)	Belgium (8 fields) and Finland (22)	2017 - 2023	Forecast 1 / day	Field level	N/A yet
Winter wheat yield variability maps (pixel-based)	Belgium (8 fields) and Finland (22)	2017 - 2023	Forecast 1 / day	0.01ha to 0.2 ha	N/A yet

Use case(s):

The spatial variations in wheat yield estimated with the digital twin model will be used to correct the harvester yield data and provide more accurate yield estimates to the farmers.

3.3.4 Winter wheat yield estimates (VITO)

Once the yield model framework that is described in section 3.3.1 has been validated for potatoes, it will also be used / finetuned for winter wheat yield estimation. Wheat harvester data will be provided by CNHi.

3.4. RIL Soil Health

3.4.1 EO based regional soil organic carbon map

This product, developed by AUTH and ILVO in collaboration with DEIMOS, includes separate pixel-based maps of Flanders and Central Macedonia with a resolution of 10m. The pixels contain estimations of the SOC in the topsoil at field level and regional level.

Sensor input data:

No sensors will be used for this data product

EO input data:

EO products	EO data provider	AOI/test sites	Time period	Temporal frequency	Spatial resolution
Sentinel 2	Google Earth Engine	Flanders	2017-2023	5-daily	10 m to 60 m
Sentinel 2	Copernicus Data Space Ecosystem	Region of Central Macedonia	2017-2023	Every 5 days	10 m to 60 m

Other input data:

Besides the EO data, other data will be used. The ML model will be trained on OC soil samples. For Flanders, a soil campaign performed during 2021 and 2022 will be used, which has provided 210 sample points for topsoil SOC throughout Flanders (Figure 18).

We expect to include new sample points from 2023 and 2024 via the Horizon EU Steropes project (<https://ejpsoil.eu/soil-research/steropes>) and the CMON project (<https://www.vlaanderen.be/inbo/en-GB/projects/monitoring-van-koolstofstocks-in-de-bodem-in->

[vlaanderen-cmon-evinbo](#)) in Flanders. For Central Macedonia 2208 sample points are available for model development (Figure 19)

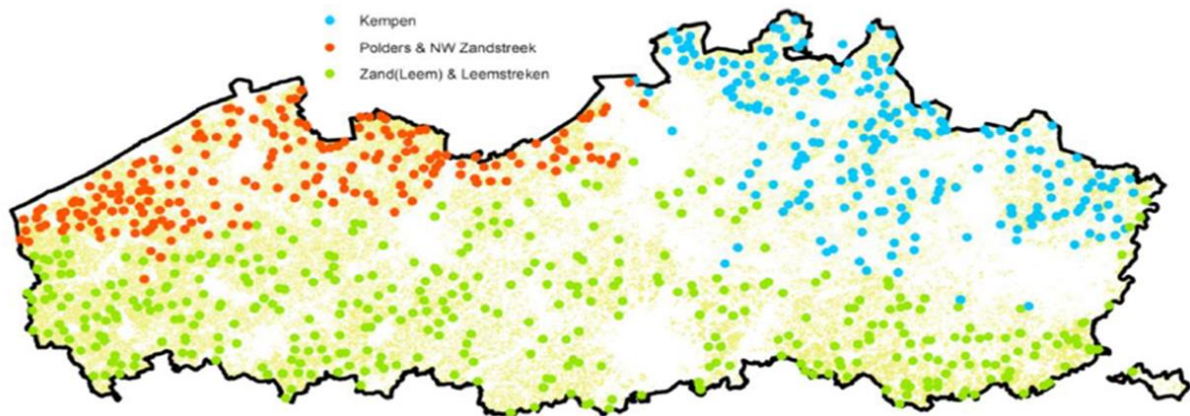


Figure 18: 210 sample points for topsoil SOC throughout Flanders, taken in 2021 and 2022.

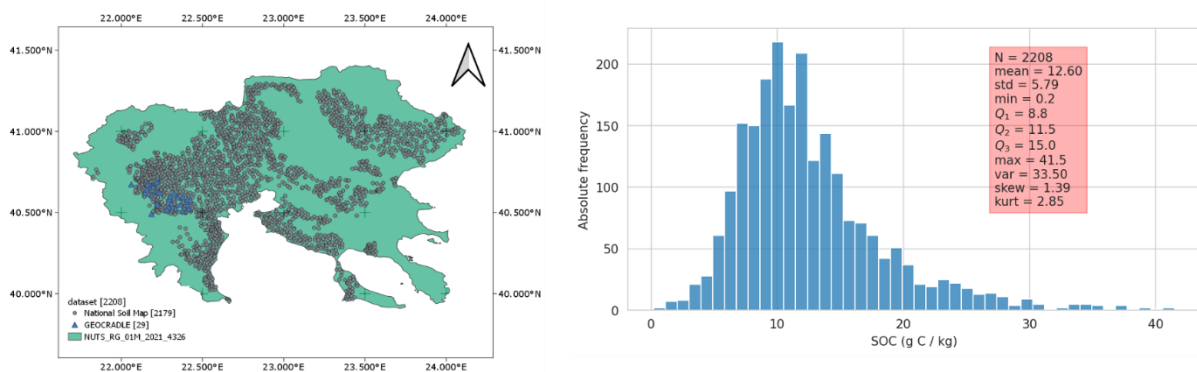


Figure 19: 2208 sample points for topsoil SOC throughout the region of Central Macedonia, and their corresponding topsoil SOC content in the form of a histogram; samples collected between 2014 and 2024.

In addition, soil property data derived from multiple European datasets provided by ESDAC-JRC (e.g. bulk density, clay, coarse fragments, sand, silt) will be included as extra contextual data to train the models

Methodology and validation results:

The following methodology is used to develop a Soil Organic Carbon (SOC) model predicting the topsoil OC in Flanders and Central Macedonia: this is performed with multiple scripts developed and run in the RI development environment. Sentinel-2 EO images are extracted and further processed (e.g. using indices and thresholds to determine if the pixels contain bare soil). In addition, soil property data derived from multiple European datasets provided by ESDAC-JRC (soil physical properties for Europe) are added to the training data. After preprocessing the EO and soil property data an Artificial Neural Network (ANN) is trained and validated for the prediction of SOC using the SOC results of the soil samples.

Due to a lack of sufficient training data and data privacy restrictions for sharing soil sample data we apply federated AI (see section 2.1) to obtain sufficient data for further training of the SOC models, This federated AI framework will be set up guaranteeing data privacy. Two approaches are applied, both using custom scripts; the first approach employing a GCP infrastructure has been set up from M10-M18, the second approach will use an existing Federated AI framework Flower.ai , which we will develop in cooperation with DEIMOS from M18-M24 (see section 2.1).

The first approach uses the following Federated AI dataflow. Multiple communications are performed between a central server-actor and multiple federated client-actors. The client-actors download the general SOC ANN, each client-actor finetunes the model on their own data for one epoch and upload the model weights and extra metadata related to the number of datapoints and evaluation metrics back to the central server-actor. The server-actor receives the files with finetuned model weights and extra metadata from all clients from the Cloud Storage bucket and performs further data processes to generate a newly federated-AI-trained model which is then again provided for a next communication round with all client actors. These communications will go on as long as the model is improved. The second approach using flower.AI will use a similar methodology but based on the infrastructure and data processes of the flower.ai framework.

The SOC map is then generated by collecting and processing EO satellite data (pixel resolution 10m) for the whole of Flanders and Central Macedonia; this is done for a certain period (e.g. 6 months to 1 year). Only the pixels of fields containing bare soil are withheld. Mean multiband spectral values are calculated for each pixel, after which the trained SOC model is used to estimate topsoil SOC values for the whole region.

Data products:

Sensor-integrated data product	AOI/test sites	Time period	Temporal frequency	Spatial resolution	Accuracy
SOC values at pixel level	Flanders, Central Macedonia	Dec 2024	Every 6 months or annually	10 m	To be determined

Use case(s):

The map will be assessed by external stakeholders, including farmers, policymakers, research institutes and government institutions and will be used for monitoring and advisory purposes.

3.4.2 Soil health indicator estimates (field level)

This product, developed by AUTH and ILVO in collaboration with VTT, includes separate pixel-based maps of selected fields at the field level in Flanders and Central Macedonia with a resolution of < 1m, developed from UAV data. The pixels contain estimations of the SOC in the topsoil.

Sensor input data:

Sensor data	Source	Data provide	AOI/test sites	Nr. fields/season	Meas. frequency	Season(s)	Nr. training data
		r		n	y		

Hyperspectral images of reflectance	VTT camera mounted on UAV	VTT	Selected fields in Flanders and region of Central Macedonia	Minimum of 2 fields / area / season	1 image / field / year	Growing season of 2024 and 2025	Minimum of 10 samples / field
Hyperspectral signatures of reflectance	Portable spectrometer in the VNIR-SWIR (PSR+3500 for AUTH and ASD FieldSpec4 Pro for ILVO)	ILVO and AUTH	As above	As above	1 spectrum per 100m ²	Growing season of 2024 and 2025	N/A

The number of training data here refers to new soil samples collected from each field where the UAV flights will take place. These will undergo the same treatment as the field samples noted in the previous section (i.e., chemical analysis in the lab).

EO input data:

EO products will be employed to augment the spectral resolution of the provided VTT hyperspectral camera. In particular, spaceborne sensors may have coarser spatial resolution but bolster a wider spectral range. This fusion will help us develop more robust soil maps.

EO products	EO data provider	AOI/test sites	Time period	Temporal frequency	Spatial resolution
Sentinel-2	Copernicus Data Space Ecosystem	Same fields as above	2024	1+ cloud-free image(s) / field / year as close as possible to the UAV data collection	10 m to 60 m
Hyperfield	KUVA space	Same fields as above	2025-2026	1+ cloud-free image(s) / field in 2025 as close as possible to the UAV data collection	As provided by KUVA

During the second iteration we will examine the integration of hyperspectral EO data provided by Hyperfield from KUVA-SPACE. The Satellite will be in orbit in Q3-2024 and data will be available by Q4-2024.

Methodology and validation results:

The plan in the selected fields in Flanders and the Region of Central Macedonia is to perform UAV flights and collect hyperspectral images using the VTT camera, and to augment these data with hyperspectral spaceborne data (first experiments with the VTT camera will be performed from M18-M24, augmentation with KUVA hyperspectral spaceborne data will be available from the end of the first iteration or from the start of the second iteration).

The steps are:

1. In the identified fields, using one of the following approaches i.e. the algorithm of ICCS, Kenard Stone or based on a SOC map, to identify optimal sampling locations from the spaceborne data (Sentinel-2 data in 2024 and Hyperfield data in 2025) collected close to the assumed day of flight for each field.
2. On the day of the UAV flight (with optimal flight conditions) on each field:
 - a. Do the UAV flight and collect hyperspectral images from the VTT camera
 - b. Collect soil spectral signatures using the portable spectrometer in the whole field
 - c. Collect the soil samples at the ICCS' algorithm identified locations
3. [Optional] If the spaceborne image's acquisition date is more than 10 days before the actual UAV flight, obtain, if possible, a new spaceborne image closer (before or after) the actual UAV flight date.
4. Fuse UAV hyperspectral data (which tentatively covers 600 to 1000 nm) with spaceborne data (e.g., RGB from Sentinel-2, complete VIS-NIR-SWIR range of Hyperfield) to obtain a fused image.
5. Train an AI model to predict SOC content on the fused image; use cross-validation to validate the result
6. Validate the sampling locations proposed by ICCS using the in-field point spectroscopy data

Sensor-integrated data products:

Sensor-integrated data product	AOI/test sites	Time period	Temporal frequency	Spatial resolution	Accuracy
Fused hyperspectral image	Same fields as above	Dec 2024 – March 2025	1 image / field / year	<1 m	N/A
Final Field SOC maps	Same fields as above	Dec 2024 – March 2025	1 image / field / year	<1 m	Target: $R^2 > 0.5$

Use case(s):

Final products can be tested by the agronomists and/or farmers who have a knowledge of the fields.

3.5. RIL Grassland

3.5.1 Gap-filled grasslands LAI maps at parcel level

This product, developed by EURAC, includes pixel-based (10m resolution), spatially gap-filled LAI time-series data to estimate grassland yield over the province of Bolzano, in the Italian Alps. The gap-filling is based on S1-S2 data fusion, developing and validating different machine learning algorithms (Gaussian Process, Random Forest, CatBoost) for the improvement of the biopars time-series over the Italian Alps. It is planned to adapt the framework also to Spain in collaboration with IFAPA.

Sensor input data:

The gap-filled LAI time-series are validated using ground measurements collected by EURAC over eight grassland parcels throughout the Provinces of Trento and Bolzano. Their location is shown in the map below together with all permanent meadows in the regions that built the AOI. This shows the potential limitation of geographical distribution of the sensor data which might be clustered and is not covering

the area equally. The sampling takes place bi-weekly from late April to the end of October of the years 2023 and 2024, spanning the growing period of the grassland before the first until after the latest mowing event of the season. At each field measurement event three replicates are collected per meadow, located in the center of a S2 10 m pixel. They consist of various single measures, spanning from the LAI over soil moisture to the vegetation height. The most important measures for the validation of the data fusion, namely the LAI and yield, can be found in the table below.

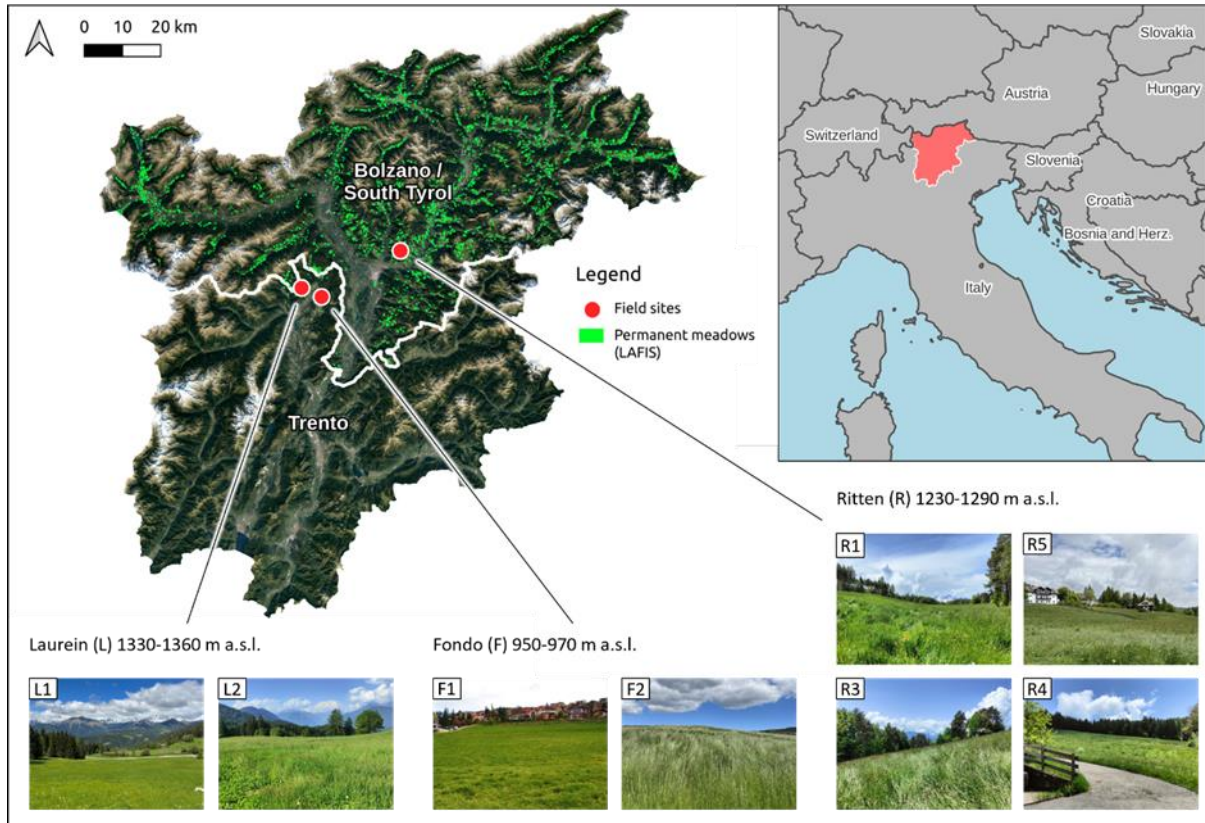


Figure 20: Location of the study area and the field sites. Imagery: Google, © TerraMetrics.

Sensor data	Source	Data provider	AOI/test sites	Nr. fields/season	Meas. frequency	Season (s)	Nr. training data
Grassland LAI	LI-COR 2200C Plant Canopy Analyzer	EURAC	Provinces of Trento and Bolzano, Italy	8	Bi-weekly	2023, 2024	+/- 250 per year
Grassland yield [t/ha]	Destructive samples						

EO input data:

EO products	EO data provider	AOI/test sites	Time period	Temporal frequency	Spatial resolution
-------------	------------------	----------------	-------------	--------------------	--------------------

Sentinel-2 products (LAI using the SNAP Biophysical Processor)	Copernicus data space ecosystem	Provinces of Trento and Bolzano, Italy	2019-2024	All available scenes (2-3 days)	10m
Sentinel-1 RTC	Microsoft Planetary Computer	Provinces of Trento and Bolzano, Italy	2019-2024	All available scenes (<3 days, since 2023 <6 days)	10m
Daily total precipitation and temperature	EURAC	Provinces of Trento and Bolzano, Italy	2019-2024	daily	250m
Digital elevation model (DEM, classification of the meadows into altitudinal belts)	European Environment Agency	Provinces of Trento and Bolzano, Italy	2019-2024	-	25m

Methodology and validation results:

S1 RTC data together with auxiliary data, namely the day of the year, daily total precipitation, daily mean temperature, and the altitudinal class derived from the DEM, are used in a data fusion approach for the spatial gap-filling of S2 LAI over Alpine grassland. For this purpose, different machine learning algorithms (Gaussian Process, Random Forest, CatBoost) are trained over the AOI for 2020-2022. A model selection with cross-validation during the model training ensures that the best-suited predictors amongst the S1 RTC and the auxiliary data are selected and that the hyperparameters are optimized. The validation is done using the in-situ LAI measurements over the eight grassland parcels and using S2 LAI of meadows that were not included in the model training. As the validation targets meadows both temporally and spatially outside of the training data, this can be seen as an independent test of the methodology.

The cross-validation during model training (2020-2022) states RF as best performing model ($R^2=0.70$, $RMSE=0.97$). However, all models perform poorly in the validation (on 2023) with CatBoost being the best model ($R^2=0.16$ and $RMSE=1.34$ against S2 LAI), and the models underestimate the in-situ LAI strongly. Additionally, the predicted LAI appears temporally and spatially smoothed, and mowing events cannot be identified accurately.

The selected features of all different models include auxiliary data which allows to disclose that S1 is difficult to interpret for the data fusion approach for S2 LAI over Alpine grassland. Due to only one S1 sensor remaining in orbit, the availability of data is a limitation for this methodology. The strong reliance on auxiliary predictors causes smoothed predicted LAI. Thus, future development of the methodology will include a greater use of S1 data without or less focus on auxiliary predictors. Also, polarimetric decomposition could ease the estimation of LAI from S1. Due to the complex terrain and size of the study area, a more conservative selection of meadows and less generalized algorithms could improve the results significantly.

Sensor-integrated data products:

Sensor-integrated data product	AOI/test sites	Time period	Temporal frequency	Spatial resolution	Accuracy
Spatially gap-filled LAI from S1-S2 data fusion	Permanent grasslands in the Province of Bolzano, Italy	2019-2024	Highest possible (1-3 days)	10 m	Yet to be improved

Use case(s):

The gap-filled LAI maps will be used to generate an improved version of the GPI (Grassland Production Index) that EURAC developed for an index-based insurance for Alpine grasslands in the Province of Bolzano.

3.5.2 Estimated grassland yield at parcel level

This product, developed by IFAPA, includes a daily estimation of grasslands net primary production (NPP) at 10m. Sentinel-2 images and meteorological information are integrated in a light use efficiency (LUE) model following an adaptation proposed for this ecosystem by Gomez-Giraldez et al. (2019).

Sensor input data:

The daily estimation of net primary production (NPP) is validated using ground measurements collected by IFAPA at 12 validation sites with 9 measurement points (S2-pixel size) distributed across five pilot farms from December 2023 to June 2024, located in the region of "Los Pedroches" in the northern part of Córdoba, Spain. These sites were selected to account for the variability of climate, topography, and soil properties within the study region. Their locations are shown on the map below.

The first validation campaign was conducted monthly from the end of December-2023 to the beginning of June-2024, covering the grassland growing period from autumn's start of season to the onset of the dry period in June. During each field measurement, 2-6 replicates (depending on the variable) per plot are collected within each 10-m S2 pixel.

Several variables were measured using different, and sometimes redundant, instruments, including grassland biomass (Grasmaster Pro, Jenquip EC20 Platometer, cutting 30x30 cm quadrants), fraction of photosynthetically active radiation absorbed by the plant (fPAR with the Accupar Ceptometer LP-80) and field radiometry (ASD FieldSpec). The most important measures for data fusion validation are listed in the table below.

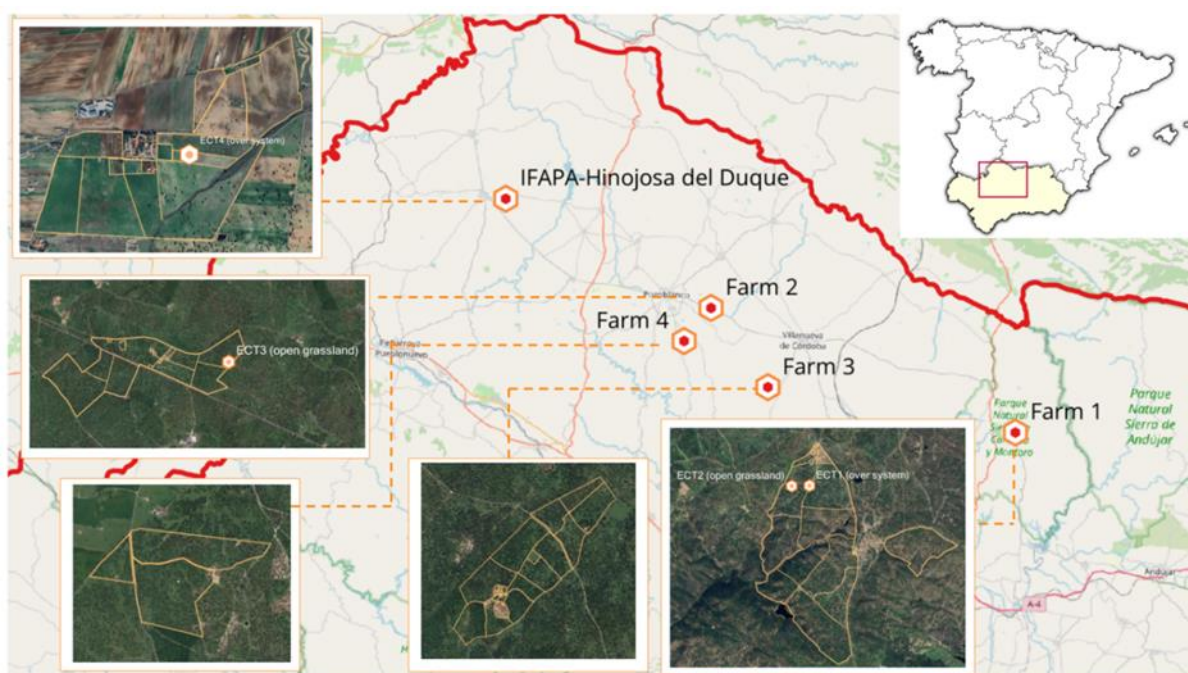


Figure 21: Location of the five pilot farms selected for section 3.5.2 and the flux towers (initially ECT2 and 3) used for section 3.5.3 in the Pedroches region (Spain).

Sensor data	Source	Data provider	AOI/test sites	Nr. fields/season	Meas. frequency	Season(s)	Nr. training data
Grassland biomass (Kg MS/ha)	Grasmaster Pro	IFAPA	Pedroches (Spain)	12 parcels x 9 points/season	Monthly during the growing season	2023/2024	+/- 600 per year
	Jenquip EC20 Platemeter						
fPAR	Destructive samples						
	Accupar Ceptometer (LP-80)						
Meteo. variables (Rad, Ta, DPV)	Weather stations (different labels)						
Field radiometry (450-2500nm)	ASD FieldSpec				only not cloudy days		+/-200 per year

EO input data:

EO products	EO data provider	AOI/test sites	Time period	Temporal frequency	Spatial resolution
Sentinel-2 reflectance	Google Earth Engine	Pedroches (Spain)	2021-2025	All available scenes (3-5 days)	10m-20m

Methodology and validation results:

To estimate grassland net primary production (NPP), an adaptation of Monteith crop production model (Monteith, 1977; Gómez-Giraldez et al., 2019) to this ecosystem was used. It focuses on (i) the presence of a variable-density tree layer influencing spectral data and (ii) the estimation of light use efficiency parameter for these seminatural grasslands using biomass field measurements. The model calculates net primary production (NPP, Kg/ha) using a remote estimation of the fraction of photosynthetically active radiation absorbed by the plant ($fPAR$), photosynthetically active radiation (PAR) measured by weather stations, and a calibrated value of light use efficiency (ϵ) following the procedure described in this scheme:

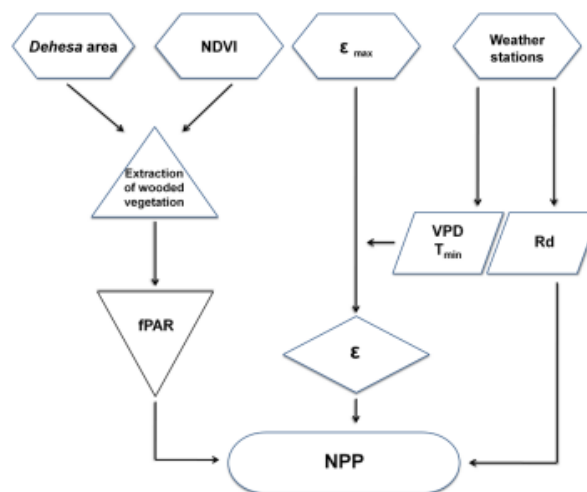


Figure 22: Schematic representation of the adapted Monteith model that is used to calculate NPP

Daily solar radiation (Rad, MJ), minimum and mean air temperature (TMin, TMed, °C), relative humidity (HR, %), and DPV for the study period were collected from 10 weather stations around the region. Photosynthetically Active Radiation (PAR) was estimated based on daily solar radiation. Fraction of PAR ($fPAR$) values were obtained from the Normalized Difference Vegetation Index (NDVI) calculated from Sentinel2 using an empirically derived equation tailored to natural Mediterranean grassland. To focus on grassland production, the contribution of oak trees was subtracted using summer NDVI values.

The light use efficiency parameter (ϵ) was calculated taking into account Tmin and VPD, which are factors known to reduce a plant's ability to use light efficiently. The maximum value of ϵ for the ecosystem (ϵ_{max}) was adjusted by a scalar minimum temperature and a scalar vapor pressure deficit derived from the daily Tmin and VPD values.

A previous version of the model was validated with good results by Gómez-Giraldez et al. (2019) using ground measurements. However, the model was revised, and a new validation was carried out here to better account for the local scale and evaluate the implementation on the studied farms. Unfortunately, the intensive drought of 2022/2023 led to a limited and less variable data collection. The current field campaign, 2023/2024, has been completed, and the data from this campaign is being processed. The outcomes will be shared around August 2024. Another field campaign is planned for 2024/2025 to complete the dataset and compensate for the loss of data caused by drought in 2023. The graph below depicts the validation with limited data for previous campaigns (2021-2023).

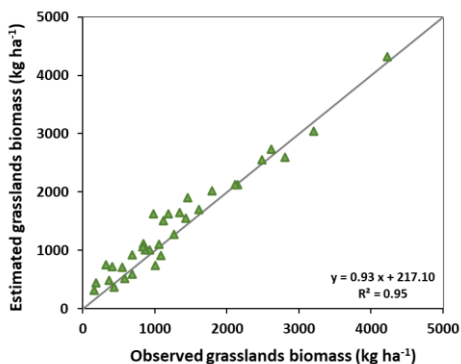


Figure 23: Validation of grassland biomass – results from previous campaigns

Sensor-integrated data products:

Sensor-integrated data product	AOI/test sites	Time period	Temporal frequency	Spatial resolution	Accuracy
Grassland accumulated NPP	Annual dehesa grasslands in the Pedroches region, Spain	2021-2025	Daily	10 m	RMSD= 224 kg ha ⁻¹ (15.5% error) using data from 2021-2023. Yet to be improved.

Use case(s):

The grassland biomass product will be shared with the pilot farms managers and the cooperative technical staff (<https://www.covap.es/meet-us/about-covap/the-cooperative>) to evaluate the best procedure for transferring the information to the farmers and how it can be used to improve the recommendations on stocking density and grazing rotations.

3.5.3 Improved grassland GPP maps based on flux tower sensors

This data product includes grassland GPP maps based on eddy covariance flux sensors, meteorological information and EO data, and is developed through a Deimos and IFAPA collaboration. It aims to produce GPP maps based on Sentinel-1 and Sentinel-2 data. The model being developed is based on a system of Feedforward neural networks trained through GPP data collected in situ by Eddy covariance and temperature sensors. The algorithm is still under development and the results obtained so far are preliminary. An algorithm has currently been tested using only the historical data possessed by IFAPA (therefore the texts with the data collected specifically for ScaleAgData are missing) and only the Sentinel-2 data (even if the algorithm has been set up to work with multi sources EO data). Over the next few months, we will test and improve the algorithm with new EO and in situ data.

Sensor input data:

Data from two eddy covariance flux towers installed over grasslands in the Pedroches area (see location in section 3.5.2) are being used. The table below provides details about the variables and sensors.



Figure 24: Sensors installed in grasslands in the Pedroches area

Variable	Sensor Source	Data Provider	Data Provider	AOI/test sites	Nr of fields/season	Meas. frequency	Seasons	Nr. Training data
GPP	CSAT anemometer LI 7500A, IRGASON	3D	IFAPA	Pedroches (Spain). Flux towers on open grassland ECT2 and ECT3 (Figure 23)	2/42 months in total	Daily (original 10 Hz averaged 30 min)	Initial historical dataset: 42 (2017-2022)	Initial historical dataset 573 ETC2 + 673 ETC3
ET, H	CSAT anemometer LI 7500A, IRGASON	3D				Daily (original 10 Hz averaged 30 min)		
Rn	4 ways radiometer NR-1					Daily (original 10 Hz averaged 30 min)		
G	HFP01 Huseflux Soil Heat Plates					30 min		
Meteo. Variables: Rad, Ta, u, P, DPV	Wheater statin (different labels)					30 min		

EO input data:

The data used are, as shown in the table, Copernicus data from Sentinel-1 and Sentinel-2.

For Sentinel-1 we used the Level 1 - GRD product.

For Sentinel-2 this is MSI level 2-A data, Top Of Canopy data provided by ESA.

At M18 of the project, we acquired data with the historical database provided by IFAPA, between 11/20/2017 and 10/20/2020. We have planned the acquisition of data between 2022 and 2024 in the

coming months, in conjunction with the measurement campaigns carried out by IFAPA in the first 24 months of the project.

EO products	EO data provider	AOI/test sites	Time period	Temporal frequency	Spatial resolution
Sentinel-2	Copernicus	Pedroches region	11/20/2017 - 10/20/2020	Officially 3 days, really, considering only the cloud free images, on average 10g (with high variance)	10m, 20m, 60m (depending on the bands)
Sentinel-1	Copernicus	Pedroches region	11/20/2017 - 10/20/2020	3/6 days	10m

Methodology and validation results:

The developed methodology aims to relate EO data (Copernicus Sentinel-1 and Sentinel-2) with Gross Primary Production (GPP) and Ecosystem Respiration (RECO) using a system of neural networks.

This methodology uses data from in-situ sensors (both for training the neural network and for validating the method) and EO data (for training the neural network and subsequently as an independent variable of the model).

The in-situ sensors were Eddy Covariance towers and environmental sensors. The first sensors measure fluctuations in wind speed and gas concentrations on an hourly basis to calculate the mass fluxes of these gases between the ecosystem and the atmosphere. The second sensors measure environmental variables such as: soil temperature, solar radiation and soil humidity.

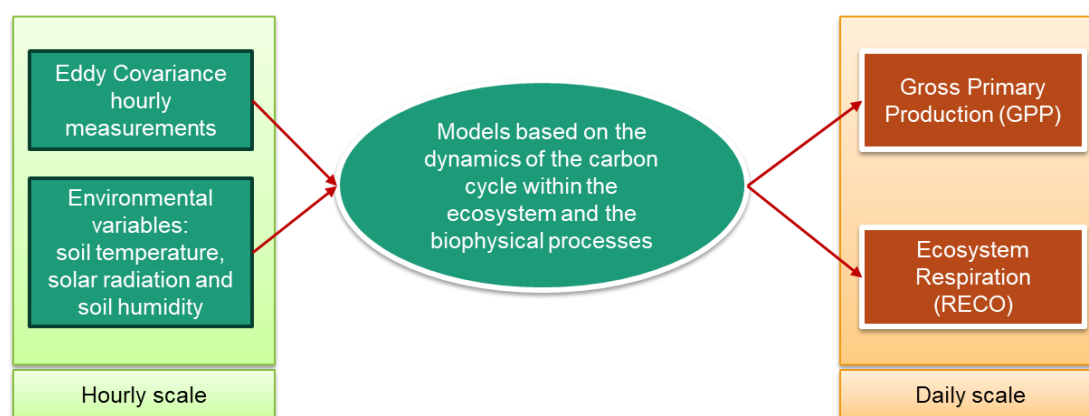


Figure 25: From field measurements to GPP and RECO variables. Data processing scheme.

For this preliminary phase, exclusively Sentinel-2 data were used to test the algorithm's performance, not only from the results' point of view but also of computational cost and processing times. We are working on introducing Sentinel-1 data into the algorithm; we plan to show the first results around the project's M20.

The Sentinel 2 data used is MSI sensor data, level 2A (Top of Canopy) for the T30SUH tile, acquired between 10/1/2017 and 8/1/2022. We have excluded from the catalog all images with cloud cover greater than 20%.

The EO data processing was divided into 6 steps:

- 1) Selection and acquisition of data from the Copernicus server
- 2) Raster clipping bat in AOI format
- 3) Pixel resampling (to 20m) for each band with pixel resolution higher than 20m
- 4) Pixel extraction inside EC footprint
- 5) Pixel Average
- 6) Data Normalization

We thus obtained a database of 294 elements where each element is characterized by: acquisition date, AOI, in situ GPP value, in situ RECO value, reflectance value for each of the 12 available S2 bands.

The algorithm used is a Feedforward neural network, composed of an input layer, 2 hidden layers (we also carried out tests for networks with 1 layer), and an output layer. The function tested for hidden layers was tangent-sigmoid, but we are preparing tests with other functions such as ReLU (and its evolutions), Swish and Maxout. A different number of nodes was tested for both layers, setting the maximum number equal to the number of elements in the input layer and trying different combinations.

The Output Layer was constantly defined by a node and the Linear Transfer Function.

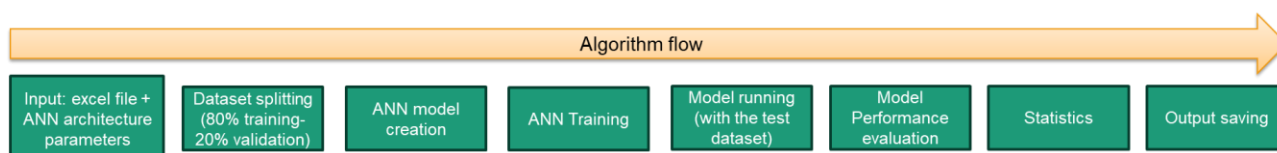


Figure 26: EO and in Situ data for GPP estimations. Algorithm flow chart.

The methodology proposed so far is under development, is being tested and adequately modified in order to optimize its performance based on the data available and the results obtained from the various tests.

The validation of the method was done by randomly dedicating 20% of the dataset to the comparison between variables simulated by the model and variables measured in situ (using EO data as independent variables and GPP/RECO as dependent reference variables). As soon as we have a larger database available, we will structure it as follows: 70% dedicated to training, 15% dedicated to method validation (intended as a tool to avoid overfitting) and 15% dedicated to testing (intended as a tool to quantify and qualify model performance).

Sensor-integrated data products:

Sensor-integrated data product	AOI/test sites	Time period	Temporal frequency	Spatial resolution	Accuracy
Grassland GPP maps	Pedroches/Spain	2017/2024	On average 10 days	20m	Preliminary results: 80%

Use case(s):

Currently the idea is to have the product tested by IFAPA to use it for research purposes. Moving on to a production phase of the service is not discarded.

3.6. RIL Dairy

The data products explained in this section are preliminary work results, ready for initial validation in the Dairy RIL at the time of writing. Depending on the results of that planned validation, decisions will be made about further development and potential upscaling of the data products. In case they would be considered useful also for users outside the Dairy RIL, further aggregation and anonymization of the data may be required as prerequisite before making them available, because these data products are partly based on confidential data.

3.6.1 Regional productivity of dairy farms

This product, developed by OHB in collaboration with DMK and ATB, includes regionally aggregated time series of milk quality and quantity, allowing analyses of the productivity of dairy farms.

Input data:

Milk quality and quantity parameters are collected for individual farms in the counties of Cuxhaven and Stade in northern Germany all year round from 2018 to 2023. The spatial distribution of the farms is visualized in Figure 27 below. Along with the amount of milk as well as the percentage of fat and protein, the dataset contains the classification of respective farms regarding their feeding (indoor housing of cows or cows grazing outside), their geolocations, further quality parameters and metadata.

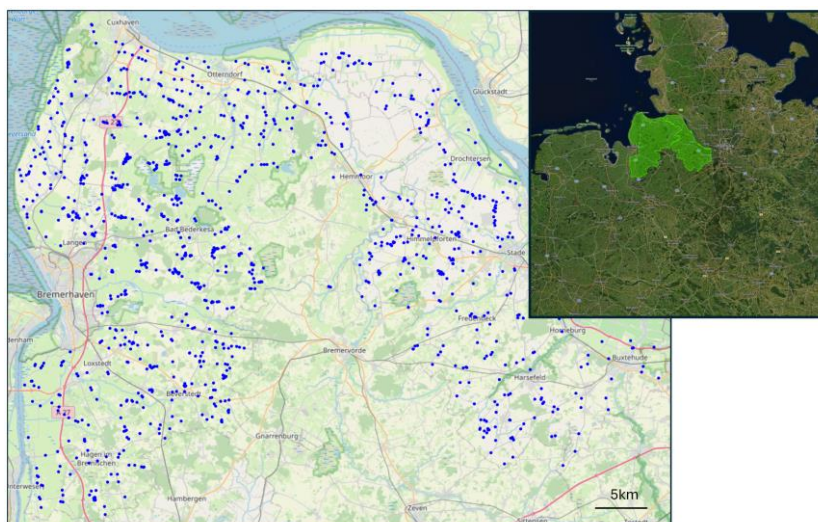


Figure 27: Spatial distribution of dairy farm locations in the sample region in Northern Germany.

Input data	Source	Data provider	AOI/test sites	Nr. farms/season	Meas. frequency	Season(s)
Milk quantity	Lab analysis	DMK	Counties of Stade/Cuxhaven, Germany	Several hundred farms	2-daily	6
Milk fat percentage	Lab analysis	DMK	Counties of Stade/Cuxhaven, Germany	Several hundred farms	2-daily	6
Milk protein percentage	Lab analysis	DMK	Counties of Stade/Cuxhaven, Germany	Several hundred farms	2-daily	6

Methodology:

At first, appropriate geographical areas for the initial validation are selected, focusing on regions with a high density/coverage of dairy farms. For the selected sample regions, raw data is exported from the ERP system of the dairy cooperative. Each data record represents information about one delivery of one farm to a processing plant.

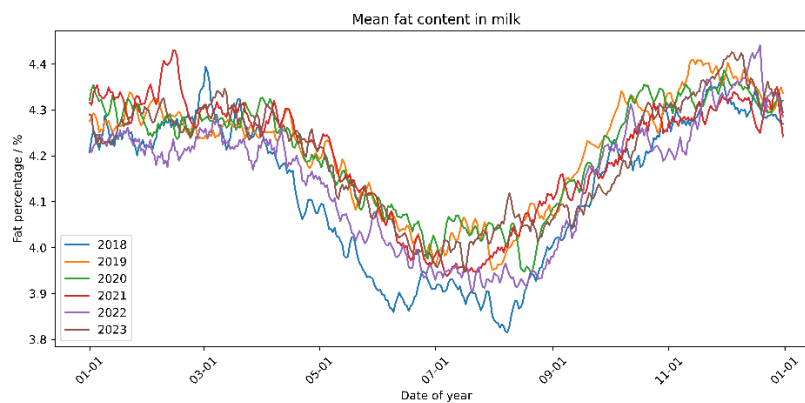
Those raw data are analyzed to identify the relevant parameters:

- the county where the farm is located
- type of farm (cows grazing outside or cows fed in the stable)
- the date of the milk delivery
- the quantity of milk per delivery
- the percentage of fat and protein (results of the lab analysis which is done for each milk delivery)

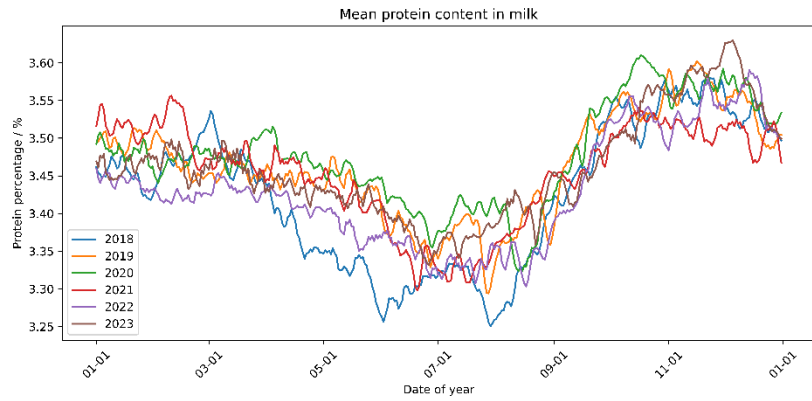
Data is then cleaned/filtered to only contain the relevant data records and fields.

The preprocessed basic data set can then be aggregated over different dimensions, summing up the milk quantity and calculating mean values for the quality (percentage of fat and protein). Based on these aggregations, the productivity can be compared between different groups of farms, e.g. between farms in different regions or between farms feeding cows in the stable vs. farms where cows are grazing outside.

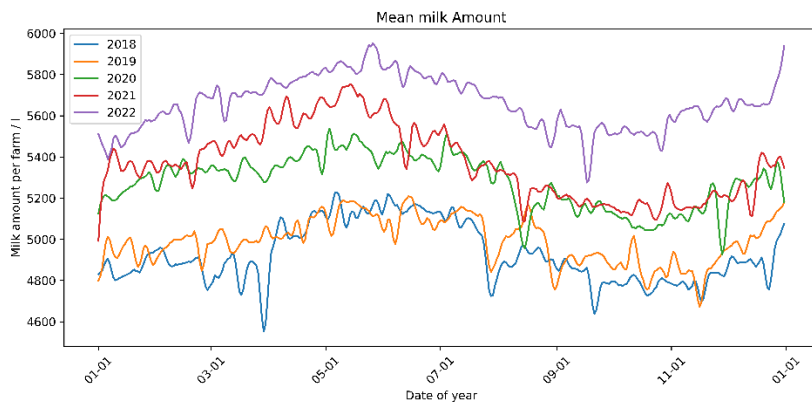
Time series of the aggregated data can be extracted from the dataset for further analysis and can also be plotted to allow visual comparison between groups of farms or to show the variation of the productivity over the course of each year. As an example, Figure 28 shows time series of daily means of fat and protein content and mean amount of milk delivered by each farm for the years 2018-2023. Recurring annual trends like minima in the quality parameters and a maximum of milk amount at mid-year can be directly identified. Figure 29 shows a visual comparison for the same parameters for the two sample counties of Cuxhaven and Stade.



a)



b)



c)

Figure 28: Annual time series for the milk quality parameters a) fat and b) protein percentage, as well as c) mean milk quantity per farm. Values are daily means from all farms.

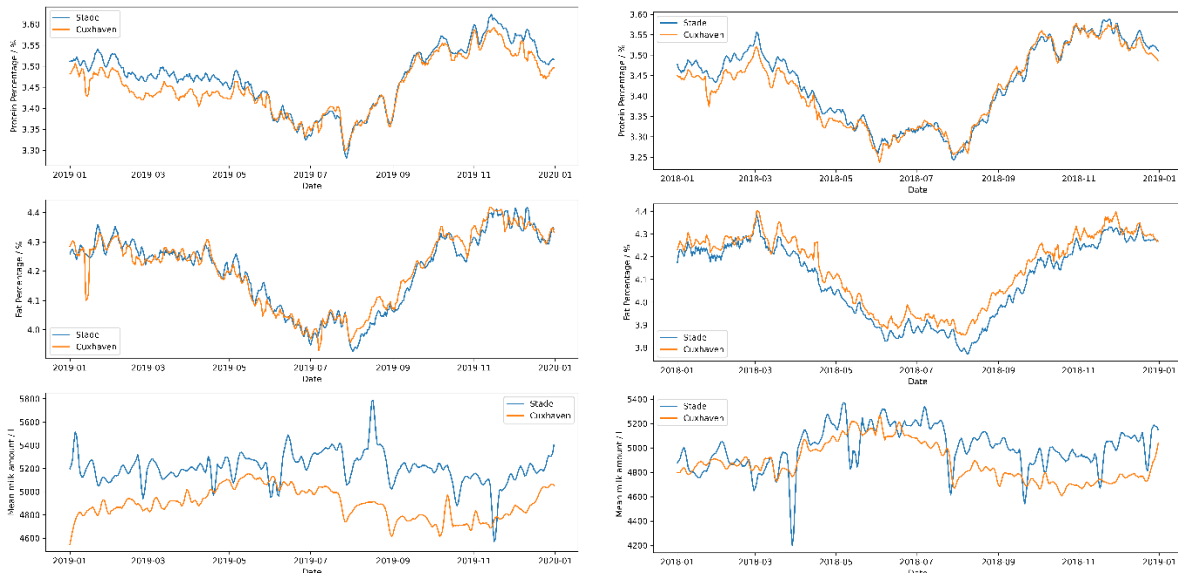


Figure 29: Comparison of milk quality and quantity parameters between counties of Cuxhaven and Skåne.

Data products:

Data product	AOI/test sites	Time period	Temporal frequency	Spatial resolution
Milk quantity (total kg) per region or per type of farm	Sample counties in Northern Germany (Stade/Cuxhaven)	2018-2023	2-daily	Region/County
Milk quality – mean % of fat per region or per type of farm				
Milk quality – mean % of protein per region or per type of farm				

Use case(s):

The dairy lab will use this data product as basis to analyze the productivity of farms in different regions and of different types of farms, the development of productivity in the course of each season and the development of productivity over several years.

3.6.2 Deviation of milk quality & quantity

This product, developed by OHB in collaboration with DMK and ATB, includes the deviation of individual (groups of) farms’ productivity from the regional productivity, as well as the deviations between the different regions.

Input data:

This builds on top of the data about regional productivity of dairy farms, i.e. the same input data is used.

Input data	Source	Data provider	AOI/test sites	Nr. farms/season	Meas. frequency	Season(s)
Milk quantity	Lab analysis	DMK	Counties of Stade/Cuxhaven, Germany	Several hundred farms	2-daily	6
Milk fat percentage	Lab analysis	DMK	Counties of Stade/Cuxhaven, Germany	Several hundred farms	2-daily	6
Milk protein percentage	Lab analysis	DMK	Counties of Stade/Cuxhaven, Germany	Several hundred farms	2-daily	6

Methodology:

Building on top of the time-series data about regional productivity of dairy farms, different deviations are being calculated:

- Deviations of each region’s aggregated productivity from the total aggregated productivity over all farms/regions
- Deviations of individual farms’ productivity time series compared to the aggregated data
 - o Aggregated for all farms in the same region
 - o Aggregated for farms of the same type (grazing cows vs. cows fed in the stable)

These deviation time series allow for the assessment of the productivity of individual farms or groups of farms. The resulting data can be plotted to allow visual detection of deviation patterns and can also be used as basis for correlation analyses between individual farms and groups of farms.

Examples regarding the quality parameter “percentage of fat” are shown in the following. Figure 30 shows the distribution of individual farms’ degree of deviation from the mean value over all farms for two sample years, showing a higher variance in 2019 compared to 2018. Figure 31 shows examples of individual farms’ seasonal development of milk quality, compared to the mean value over all farms.

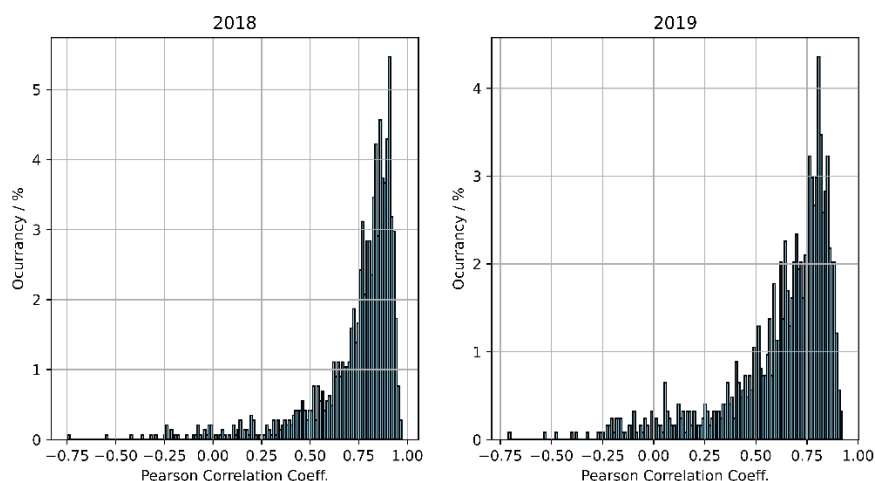


Figure 30: Distribution of Pearson coefficients for the correlation between time series of fat content in milk of individual farms to the time series of mean fat content over all farms.

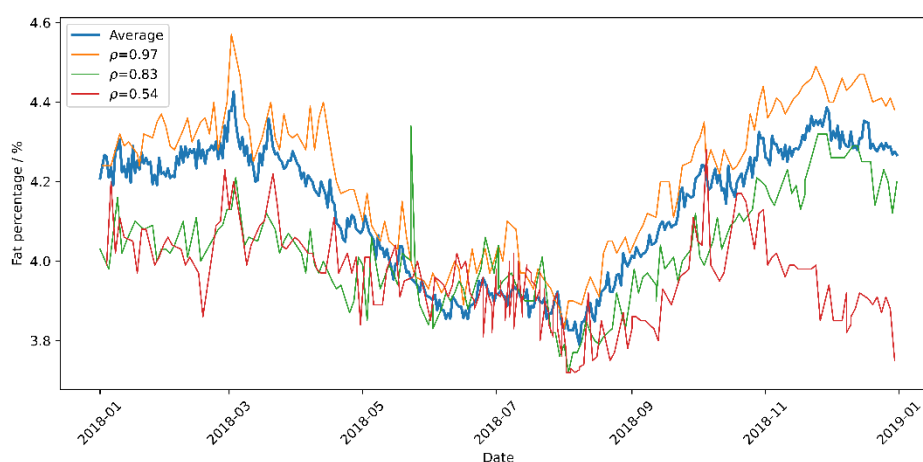


Figure 31: Example for timeseries of fat percentage in milk of individual farms with diverse values of Pearson correlation coefficients to the mean values for year 2018.

Data products:

Sensor-integrated data product	Reference values to calculate deviations	AOI/test sites	Time period	Temporal frequency	Spatial resolution
Deviation of milk quality (% of fat, % of protein) and quantity (kg) of individual	Aggregate over all farms Aggregate over farms in one region	Sample counties in Northern Germany (Stade/Cuxhaven)	2018-2023	2-daily	Per farm/region

farms or groups of farms	Aggregate over farms of one type				
--------------------------	----------------------------------	--	--	--	--

Use case(s):

The dairy lab will use this data to analyze how far the productivity of individual farms or groups of farms deviates from the mean/totals over different regions or types of farms. This data product may also support milk quality/quantity benchmarking.

3.6.3 Assessment of grass yield at regional level

This product, developed by OHB in collaboration with ATB and DMK, enables assessment of the grassland productivity on regional and field-based scale.

Sensor input data:

Sensor data	Source	Data provider	AOI/test sites	Nr. fields/season	Meas. frequency	Season(s)
Grass yield per area (ha)	Forage harvester	Machine telemetry	Counties of Stade/Cuxhaven, Germany	N/A	N/A	6
Grass' dry matter percentage	Forage harvester	Machine telemetry	Counties of Stade/Cuxhaven, Germany	N/A	N/A	6

EO input data:

EO products	EO data	AOI/test sites	Time period	Temporal frequency	Spatial resolution
Vegetation Indices based on Sentinel-2	OHB	Full Counties of Stade/Cuxhaven	2018-2023	5-daily	10m
Biomass production	VITO	Representative fields for several individual farms	2018-2023	daily	10m
ESA WorldCover	ESA	Full Counties of Stade/Cuxhaven	2021	static	10m

Spatial data	provider	AOI/test sites	Time period	Temporal frequency	Spatial resolution
Agricultural land use classification	State ministry	Full Counties of Stade/Cuxhaven	2023	static	1m

Methodology:

Sensor data collection starts with the export of raw data collected by forage harvesters, encompassing following parameters:

- Machine type
- Fruit classification (manually inserted)
- Vehicle speed
- The timestamp of yield data points
- Yield per area
- Dry matter percentage

Parameters like vehicle speed, fruit classification and the distribution of yield data are used for data points filtering to ensure that only data points associated with grassland harvest are taken into account. The aggregation of data points within a county and a given temporal extent enables historical estimates of harvest quality and quantity per area to assess the productivity of grassland on a regional scale, based on analysis of the distribution of yield and dry matter content data.

EO data enables grassland productivity assessment at local and regional scale. For regional scale analysis, the collection of data encompasses automated processing of Level 2 Sentinel-2 data. The processing encompasses reprojecting, clipping and cloud masking of the data products, as well as masking with land cover classification layers and the calculation of vegetation indices, such as NDVI, NDMI or EVI and respective statistics for the full spatio-temporal extent of the AOI and the periods 2018-2023. Further, the collection of growing rates for dry matter productivity and biomass accumulation for field-sized polygons in cooperation with the project partner Vito enables quantitative estimation of grassland productivity of individual fields at local scale.

Time series and statistics of the EO data products and in-situ harvest data can be provided for the analysis of yearly or seasonal productivity and between different regions.

Sensor-integrated data products:

Sensor-integrated data product	AOI/test sites	Time period	Temporal frequency	Spatial resolution
In-situ Grassland Productivity (dry matter yield per area)	Counties of Stade/Cuxhaven	2018-2023	Yearly	Per County
EO Grassland Productivity	Counties of Stade/Cuxhaven	2018-2023	Mission-dependent	Index-dependent

Use case(s):

The dairy lab will utilize this data product to analyze grassland productivity across various regions and scales, as well as to examine the development of grassland productivity over multiple years. This also allows correlation analyses with dairy farms’ productivity data, aiming to get new knowledge about the influence of grass yield on the productivity of dairy farms in different regions.

4. Availability of results in the RI environment

The RI environment is the main repository of the different common datasets, applications and products that are needed by the RI Labs during the Technology Validation stage or that are part of other development activities conducted in WP3, WP4 and WP5.

An overview of the RIE as well as a user manual were already provided in D4.1. To complement this, this chapter's focus is to provide best practices for data handling within the RI environment. To provide support on this and other subjects all users should use the RIE service desk available here: <https://service4eo.atlassian.net/servicedesk/customer/portal/18>

The RIE team (managed by Deimos) held meetings with the RI Labs to understand their data collection and sharing requirements. From those meetings three profiles were established:

Data that is for the exclusive use of the person/lab team and does not need to be shared with other members: in this case each user has a 1Gb personal storage in the RI environment and they should use it for this purpose. The team members can use them as any other drive (create folders and subfolders, add/delete files, etc.). If more space is required a user can request it via a support ticket in the service desk

Data that needs to be shared between different ScaleAgData partners: these datasets should be available in the RIE data catalogue. It is not feasible for the RIE team to assess all the needed datasets for each RI Lab (including, for instance, required Aol and Tols). To overcome this limitation each RI Lab team should raise a service desk ticket to request a particular dataset to be available in the catalogue. Whenever a ticket of this type is raised, the service desk team will reply by sending a data collection metadata template for the user to fill with required information in order to be able to harvest the required dataset and define data access permission rules. If no data access rules are defined, by default, the datasets will be available for all project partners. Typically, it could take up to two weeks for the dataset to be available in the RIE environment. Once they are catalogued, the user will be notified via the corresponding service desk ticket. Details on how to access those catalogued datasets are given in the RIE user manual, which is part of D4.1.

Data available in external sources (e.g., Google Drive or One Drive) that needs to be accessed in the RI environment: in this case there are already examples developed by the Soil Health RI Lab where Google Drive based files are accessed and new files are saved into that external drive using specific python libraries so no limitations are foreseen at this point to do this in the RI environment. Nevertheless, if a specific user finds a problem setting this up, they should raise a service desk ticket for the RIE operations team to analyze.

During the first 18 months of the ScaleAgData project, methodological frameworks have been defined (see Chapter 2), and collaborations have been established with several RI Labs. Methods are being co-developed for specific use cases, with numerous exchanges occurring between the technology providers and these Labs. Experiments have been set up, while test products and code have been made available to the Labs for testing.

Currently, Federated AI technologies are being tested within the Soil Health Lab. The Yield Lab is exploring few-shot learning for potato yield estimation, and Digital Twins are being set up for wheat yield estimation. Collaborations have also been initiated to set up Digital Twins with the Water management and Crop management Labs. Timeseries of ET and SM maps have been provided to Water Management, Crop Management and Yield Monitoring Labs. The provided data serves as a

benchmark against which improvements derived from integrating in-situ and EO data can be evaluated by the Labs and the partners.

Within the RI Labs, basic models are being developed to generate agri-environmental data products (see Chapter 3). While model development is still ongoing in the RI Labs and the potential use of the methodological frameworks is being explored, some Labs, such as the Yield Lab and the Dairy Lab, have already generated initial data products. After the 2024 growing season (December 2024) more consolidated sensor-integrated data products are expected.

In the coming months (M18-M24), when a sufficient level of maturity is reached, the first version of the methodological frameworks and the methods developed in the Labs will be gradually made available on the RIE, along with the data products. Currently, at this stage of the project (M18), they are not yet mature enough to be shared on the RIE for evaluation. Testing and evaluation of the methodological frameworks (task 4.4) took place through direct exchanges of test products and code between technology providers and RI Lab partners.

The next update of this document is foreseen in M36, at which stage a more complete overview of the sensor-integrated data products will be provided.

5. References

- Gómez-Giráldez, P.J., Aguilar, C., Caño, A.B., GarcíaMoreno, A., González-Dugo, M.P., 2019. Remote sensing estimation of net primary production as monitoring indicator of holm oak savanna management. *Ecological Indicators*, 106, 105526. <https://doi.org/10.1016/j.ecolind.2019.105526>
- Guo, X., Fang, X., Zhu, Q., Jiang, S., Tian, J., Tian, Q., & Jin, J. (2023). Estimation of root-zone soil moisture in semi-arid areas based on remotely sensed data. *Remote Sensing*, 15(8), 2003. <https://doi.org/10.3390/rs15082003>
- Kisekka, I., Peddinti, S. R., Kustas, W. P., McElrone, A. J., Bambach-Ortiz, N., McKee, L., & Bastiaanssen, W. (2022). Spatial–temporal modeling of root zone soil moisture dynamics in a vineyard using machine learning and Remote Sensing. *Irrigation Science*, 40(4–5), 761–777. <https://doi.org/10.1007/s00271-022-00775-1>
- Li, M., Sun, H., & Zhao, R. (2023). A review of root zone soil moisture estimation methods based on remote sensing. *Remote Sensing*, 15(22), 5361. <https://doi.org/10.3390/rs15225361>
- Monteith, J.L., 1977. Climate and the efficiency of crop production in Britain. *Philos. Trans. R. Soc. London. B, Biol. Sci.* 281, 277, LP-294.
- Souissi, R., Zribi, M., Corbari, C., Mancini, M., Muddu, S., Tomer, S. K., Upadhyaya, D. B., & Al Bitar, A. (2022). Integrating process-related information into an artificial neural network for root-zone soil moisture prediction. *Hydrology and Earth System Sciences*, 26(12), 3263–3297. <https://doi.org/10.5194/hess-26-3263-2022>
- Tseng, Gabriel, et al. "Lightweight, pre-trained transformers for remote sensing timeseries." *arXiv preprint arXiv:2304.14065* (2023). <https://arxiv.org/abs/2304.14065>
- Tseng, Gabriel, et al. "Cropharvest: a global satellite dataset for crop type classification." *Neural Information Processing Systems (NeurIPS)* (2021). [Cropharvest: A global dataset for crop-type classification \(neurips.cc\)](https://arxiv.org/abs/2106.04556)
- Yinglan, A., Guoqiang, W., Peng, H., Xiaoying, L., Baolin, X., Qingqing, F. (2022). Root-zone soil moisture estimation based on remote sensing data and Deep Learning. *Environmental Research*, 212, 113278. <https://doi.org/10.1016/j.envres.2022.113278>