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

D4.2 Data-based farming services, V1

Revised version, 27 January 2025

Author		
Name	**Organization**	**Release date**
Matti Pastell	Luke	27/06/2024
Annimari Hartikainen	Luke	27/06/2024
Sarah Verbeke	UGent	27/06/2024
Dainis Jakovels	IES	27/06/2024
Panagiota Louka	NP	27/06/2024
Hanna Huitu	Luke	27/06/2024

Review on behalf of the Executive Board		
Name	**Organization**	**Date of review**
Nick Berkvens	ILVO	17/06/2024
Isabelle Piccard	VITO	18/06/2024
Gunnar Große Hovest	ATB	20/06/2024

Revision records			
Version	**Date**	**Changes**	**Authors**
1.0	28/06/2024	Original document	Luke
1.1	27/01/2025	Revised version that takes the comments of the EC and external reviewers into account	Luke

Acronyms and Abbreviations

Acronyms and Abbreviations	
AgMIP	Agricultural Model Intercomparison and Improvement Project
AIM	Agriculture Information Model
APSIM	The Agricultural Production Systems sIMulator (*Simulation model*)
C	Carbon
CAP	Common Agricultural Policy
CERES	Crop Environment Resource Synthesis (*Simulation model*)
CNHi	CNH Industrial Belgium
DEMETER	Horizon 2020 project on Internet of Things platform for smart farming
D	Deliverable
DGG	Discrete Global Grid
DHI	DHI A/S (Denmark)
DSS	Decision Support System
DSSAT	Decision Support System for Agrotechnology Transfer (*Simulation model*)
EC	Electric Conductivity
EO	Earth Observation
ET	Evapotranspiration
ETSI	European Telecommunications Standards Institute
EU	European Union
FMIS	Farm Management Information System
H3	Hexagonal hierarchical geospatial indexing system
IACS	Integrated Administration and Control System
ICASA	The International Consortium for Agricultural Systems Applications
ICCS	Institute of Communication and Computer Systems
ISOBUS	ISO 11783, standardised communication protocol in agriculture
JSON	JavaScript Object Notation, data interchange format
KMI	Royal Meteorological Institute of Belgium
LAI	Leaf Area Index
LPIS	Land Parcel Identification System
Luke	Natural Resources Institute Finland
M	Month
MIT	Massachusetts Institute of Technology
N	Nitrogen
NDVI	Normalized Difference Vegetation Index
NGSI-LD	Next Generation Service Interface – Linked Data (*Context Information Management API*)

NP	Neuropublic SA
OGC	Open Geospatial Consortium
PBM	Process Based Model
PTF	PedoTransferFunction
R&D	Research and Development
RIL	Research and Innovation Lab
SAFY	Simple Algorithm for Yield Estimates (S*imulation model*)
SDK	Software Development Kit
SPAD	Soil Plant Analysis Development
STICS	Simulateur mulTIdisciplinaire pour les Cultures Standard (S*imulation model*)
TERRA-REF	The Transportation Energy Resources from Renewable Agriculture Phenotyping Reference Platform
UGent	Universiteit Gent
UI	User Interface
USDA	United States Department of Agriculture
VRI IES	Foundation "Institute for Environmental Solutions"
WOFOST	WOrld FOod STudies (*simulation model*)
WP	Work Package

Table of Contents

List of tables

List of figures

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 into 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 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 (RILs) have been identified within the project, across various bio-geographical 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

This document provides a methodological framework that employs digital twin modeling. The framework combines and integrates multiple data sources into actionable information for farming. The provided guidelines, digital twin data model, and Python codes enable RI Labs to validate these technologies before being implemented.
Work described in this deliverable lays the basis for developing new data assimilation methods, creating automatic calibration of models, and finally creating a framework that enables training of intelligent agents for making management decisions based on multiple objectives.
This deliverable contributes to Innovation Area 5 “From data assimilation to service development”, and summarizes work done in WP 4 (Product and service development) Task 4.1 (Data-based farming services) within the respective RILs.

1.3. Document structure

This document is structured as follows:

- Section 1 provides a project overview and describes the scope, responsibilities, and structure of this deliverable.
- Section 2 provides a brief overview on the digital twin approach, both in context of smart farming and in the context of potential application cases within the ScaleAgData project. Multiple data sources, use of data models/vocabularies, and suitable modeling approaches are presented. Three potential application cases from the ScaleAgData RILs are described.
- Section 3 contains a description of and links to reference software for 1) entering and retrieving field data into/from a digital twin data model and 2) setting up the first worked example on wheat crop model, based on a co-operation with one of the RILs (Yield Monitoring Lab).
- Section 4 contains a list of references.

1.4. Evolution of the document

Version 1.0 of this document, submitted on 30 June 2024, described the developments in the initial phase (M1-M18) of the ScaleAgData project.

The present version of the document, version 1.1, submitted on 27 January 2025, includes minor additions that take the comments of the EC and external reviewers on this deliverable into account. References to clarify co-operation between work packages are added to introduction of Chapter 2. Methodological framework and its implementation, and to 2.4.2.1. Interoperability. Description on what developments were already available before the implementation and what developments were performed in the project is added to 3.1 Software description.

An updated version of this deliverable, version 2.0, is foreseen for December 2025 and will contain more general guidelines for a larger audience.

2. Methodological framework and its implementation

This chapter introduces the concepts of a digital twin and introduces potential modeling approaches, data sources and the use of a dedicated data model as an interface between different data sources and model outputs. Initial design considerations are presented for several potential application cases. The design work of these application cases was initiated in WP2 co-design workshops and planning meetings and contributed to the project architectural design. Analysis-ready EO data for modeling was brought available by WP3, and planning and carrying out the data collection for model input and calibration was and will be carried out in WP5.

The following steps were identified for developing a digital twin for supporting data-based farming services in the ScaleAgData project:

1. Definition of the target system and decision problem including exact decisions that are expected from the model and the required spatial and temporal resolution. Estimating the expected pay-off from developing the system.
2. Defining the physical counterpart to be modeled and available digital and analog data sources. Evaluation of the relevance of available data sources for the decision problem and identification of additional data collection requirements. Iterative mapping between the obtainable data sources and selection of suitable models considering also the available data collection resources and scaling up potential of the solution.
3. Identification of suitable models. The selection of a model depends on the decision problem, system definition and data availability.
4. Defining a data model for unified representation of model input data sources and digital twin outputs.
5. Evaluation of the scaling up potential of the solution. This evaluation will also affect the choice of data collection and model selection.

These steps were applied to several case studies and are explained under section 2.5 “Potential application cases.”

2.1. Digital twins

A digital twin is a combination of measured data and models, replicating the state of an object in real time. Digital twins are used to observe processes that are difficult to measure directly, or in planning and optimization of products or processes. The concept of digital twin is used in various specific applications in agriculture.
Several Decision Support Systems (DSS) tools which combine sensor data or EO data with process-based models (PBMs) of e.g., crop development or disease risk prediction, have already been developed and some are also already in use by farmers (Verdouw, C. et al. 2021). These systems are mainly focused on monitoring the system state with limited predictive capabilities (Pylianidis et al. 2021).
ScaleAgData aims to develop a methodological framework for combining multiple models and sensor data sources into prescriptive digital twins, enabling multi-objective (e.g. maximizing yield while minimizing nutrient emissions) decision making on management actions. Methods for automatically initializing and calibrating well known simulation models such as APSIM (Holzworth et al. 2014) and/or WOFOST (de Wit et al. 2019) will be developed based on a combination of existing data sources (soil scanning, machine operations, e.g. ISOBUS tasks, and EO data products).
The potential of digital twins for supporting smart farming operations has been recognized recently and the potential of the concept has been demonstrated in early applications. Precision farming technology enables site specific management (e.g. fertilization, working depth) based on e.g. yield

maps, soil scans, soil sensors and remote sensing. However, many decisions are currently made individually by each farmer, and the effects of different decisions are difficult to assess as each location in a field has specific characteristics (e.g. soil type), and each year is different.
Use of digital twins in agricultural research opens new possibilities to observe farming systems and to develop and apply modeling for projecting the responses of the systems to altered management or changing environment. For the agricultural sector, new economic and environmental gains can be acquired through technology transfer and the use of new decision support tools that help optimizing farming operations.

2.2. Modeling approaches

Digital twins need accurate real-time models of the target system. We provide a brief overview of used and potential approaches in the context of precision farming.

2.2.1. Biophysical models

Several process-based biophysical cropping system models or agroecosystem models have been developed. These models aim to describe how crops interact with the environment and agronomic management. Crop models have several use cases including simulating crop yields under climate change, understanding and optimizing crop rotations, designing plant breeding targets and optimizing management. Well known models include APSIM (Holzworth et al. 2014), STICS (Brisson et al. 2003), WOFOST and DSSAT.
Crop models typically simulate the daily total biomass potential of the crop based on global solar radiation, simulated canopy light interception and ambient temperature. The actual growth is then obtained by limiting the total potential with water or nutrient stress. The simulated growth is then partitioned to different organs based on the phenological stage. Complex crop models have quite high requirements for calibrating new cultivars and for setting the soil properties, e.g. water holding capacity and soil nutrient contents. In very simple crop models, such as SAFY (Duchemin et al. 2008), the phenology is greatly simplified, only the total biomass is simulated and fixed constants are used to estimate e.g. grain yield. The use of simple crop models can be a good option in data-constrained scenarios and in combination with machine learning models.
Simulations are typically point-based; however, the models can be spatialized for precision farming use cases by using a dense grid. Spatial correlation of observation is typically ignored or accounted for in postprocessing of model outputs. Biophysical models have been shown to be useful for management decisions when calibrated correctly. Gobbo et al. (2022) present an approach to use crop model and N uptake maps for planning site-specific fertilization, and an approach for using APSIM in digital twins was developed at Luke (Bloch et al. 2023) and it will be extended in ScaleAgData.

2.2.2. Machine learning and hybrid models

Machine learning models learn the relationships between input and output variables from training data. The advantage of using machine learning is that the models can learn complex nonlinear relationships between variables, given enough training data, however the models can perform very poorly and yield unexpected results outside of the domain of the training dataset.
The key challenge in agricultural applications is to obtain training data that covers the variability in data across weather conditions, soil types and management alternatives. One possible approach is to conduct management response trials and use the data as input for machine learning models. For instance, Tanaka et al. (2024) compared the performance of several different models setting economically optimal site-specific fertilization rates and concluded that model predictions were very sensitive to the choice of algorithm and the selection of covariates.
One promising way to tackle the need for extensive training data is to combine process-based and machine learning approaches. This can be done in several ways: simulation models can be used to

generate training data for a machine learning model to augment observational data (Pylianidis et al. 2022), a machine learning model can be integrated into a process-based model to represent part of the processes, e.g. phenology (Droutsas et al. 2022), or simulation model outputs can be used as a feature for a machine learning model. In ScaleAgData, these approaches will be compared to pure process-based models.

2.3. Data sources

Data collection on agricultural fields of operational farms has rapidly increased during the last decades (Kayad et al. 2022). Crop monitoring throughout the season can be carried out in real time using remote sensing (drones, aerial and satellite imagery) and finally by crop monitors in harvesters. Crop growing conditions such as temperature or water availability can be monitored directly in real time using sensors, and parameters such as nutrient availability or pest pressure can be monitored using a combination of data and models. Soil scanning campaigns map variation in many soil parameters. All these data sources and streams from different sensors enable building a digital twin for smart farming. Data to be integrated into a digital twin varies by model implementation and availability of useful data. Table 1 below lists the types of data from cultivated fields that we now foresee to be utilized in digital twins during this project, and the role of each in modeling. ScaleAgData RILs have extensive set-ups for sensor installation beyond what is listed here or available on an operational farm (see deliverable D3.2 and the catalogue within).

Table 1. Data sources for Digital Twins

Field borders	Field borders are used and produced in multiple phases of a cultivation process. Many FMIS contain field border data. Also, ISOBUS task files for each cultivation task contain spatial information about where each task was carried out and whether the intensity varied in the plot. In EU countries, accurate field borders for CAP eligible parcels are part of the Land Parcel Identification System (LPIS) and used in the Integrated Administration and Control System (IACS). Borders define the area of interest for modeling and boundaries for retrieving EO data.
Soil properties	Spatial variation in soil's physical and chemical properties (such as texture, water holding capacity or amount of organic matter) affect water and nutrient intake of the crop. Soil maps can be interpolated from georeferenced soil samples, and for some soil parameters they are acquired via soil scanning. In the context of digital twins, soil properties are static information during the growing season and used for model initialization. Observations at differing depths enable predicting, for instance, water movement and availability.
ISOBUS task files	Task files are recorded by tractors and their implements and store georeferenced information on cultivation events, including sowing dates, cultivars, seeding densities, and fertilizers. ISOBUS task files from sowing and fertilization tasks are used for model initialization and updating during the growing season to obtain fertilization levels and seed density.
Field management data	Field management data contains information such as crop type, cultivars, cropping events, fertilization, pesticide applications etc. On the farm, field management data is often stored and managed digitally in Farm Management Information Systems (FMIS). This data set complements and partly overlaps with information contained in task files. Field management information is used for model calibration, initialization, and simulation.
Weather data	Weather data on a field originates from a local weather station or from some other weather data source. Temperature, solar radiation, and precipitation are the most used parameters, but also others such as relative humidity, wind speed and direction, as well as air pressure may be of interest. Historical weather data is used in running the model and weather forecasts for predicting future state.
Soil sensor data	Data streams from soil sensors, measuring parameters such as soil temperature, moisture, and salinity, can be used for data assimilation when running the model for yield forecasts, and directly as input for a controller in crop irrigation.
Satellite remote sensing products	Indices such as NDVI and different EO-based estimates are based on current satellite information. For cropping purposes, interesting products include crops' biophysical parameters as well as soil water and crop water stress related parameters. Depending on the desired set-up, they can be used in running the model, model calibration, and data assimilation.
Drone products	Drone based imagery can provide accurate crop status information covering each field. Data can be used to augment or replace EO-based data products. The main advantage of drone data can be higher spatial and temporal resolution.
In situ crop observations	Crop information can be extracted from samples taken during growing season or at harvest. Physical samples may be analyzed in the laboratory, or observations can be made by scanning the crop or plant/leaves optically with hand-held devices. The acquired information is useful for quantifying crop status, for instance the nutrient needs. In digital twins, these observations can be used in model calibration and also in data assimilation, if they are available in near real time
Yield maps	Yield maps are data products created by combine-harvesters equipped with sensors and possibly enriched with other sources of data such as satellite-based information. Maps contain spatial information on yield quantity and sometimes quality. Yield maps and data streams from harvester yield sensors act as feedback to modeling and are used in calibration. Historical yield maps can potentially be useful for model initialization by estimating legacy effects from previous crops, or for evaluating the model's capability to identify areas of (recurrent) poor crop performance.

2.4. Data model

2.4.1. Harmonizing digital twin input and output data with a data model

As described in chapter 2.2, digital twins can utilize different process-based, machine learning and hybrid models, i.e. digital system models. The input data for these models will come from varying suppliers and can have multiple different formats (as listed in chapter 2.3). The data format must be changed to match the specific format required by each digital system model. Applying a digital twin to a new use case should be easy and changes in digital system model and data sources should be flexible. While data sources and digital system models are changing, the required input and the provided output datasets for the digital twin should be clear. Existing standard data models and ontologies for agricultural applications are used as a basis and they are extended both for the use cases presented earlier in ScaleAgData D3.1, and for the use in the digital twin as described here. The digital twin has specific data requirements not considered in D3.1., so a specific digital twin data model needed to be declared. The interoperability of the digital twin data model with D3.1 is discussed in more detail in chapter 2.4.2.1.
The digital twin data model describes the general data requirements and formats that data sources need to adhere to. In every use case, the input data is first converted to a harmonized digital twin data model format. The digital twin data model should follow existing ontologies and vocabularies as much as possible, e.g. ICASA for crop vocabulary. Once the data is in this harmonized digital twin data model format, the data can be altered further to formats required by each digital system model. Conversion of data from digital twin data model -format to the model-specific formats required by each digital system model is to be carried out automatically with code.
End users, in this case the ScaleAgData RILs, adopting a digital twin, should convert their source data to a declared digital twin data model format. After that, the digital twin with its multiple digital system model possibilities is available without further data formatting. Also, a digital twin provides outputs in a published digital twin data model format, not in varying output formats specific for each digital system model, enabling straightforward output data usage. Figure 1 visualizes the purpose of the digital twin data model.

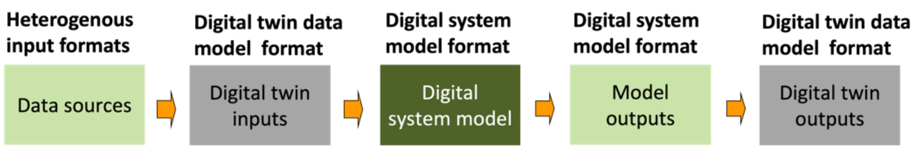

Figure 1. Digital twin data model usage between heterogenous data source inputs, digital system model and digital twin output.

2.4.2. Technical implementation and interoperability

2.4.2.1. Interoperability

The technical implementation of digital twin data model is in line with D3.1 interoperability requirements in chapter 2.5. D3.1 interoperability refers mainly to semantic interoperability, which can be improved by specifying data and data models. In D3.1 the extendable DEMETER AIM (Palma et al. 2022) ontology is chosen. The Open Geospatial Consortium (OGC) is also setting up an AIM Standard Working Group (https://www.ogc.org/requests/public-comment-requested-agriculture-information-model-standards-working-group-charter/, https://github.com/opengeospatial/aim-swg). As the DEMETER AIM ontology does not consider data requirements of specific digital system models, a separate digital twin data model is required for the digital twin and its digital system models.

The need for digital twin data model was agreed and its interoperability with DEMETER AIM was planned with WP3.

Entity is an informational representative of something existing in the real world (ETSI, 2021). Digital twin data model entities can be mostly linked to DEMETER AIM entities, and additional entities required by digital system models are added as extensions, see example in Figure 2. Figure 2 presents the link between digital twin data model FieldParcel entity and DEMETER AIM AgriParcel entity. Some attributes required by digital system models are missing from the DEMETER AIM entities. The missing attributes are added to the digital twin data model entities, like attributes category, h3resolution, h3parcel and official parcel ID are added to FieldParcel entity in Figure 2. DEMETER AIM and additional entities are also used in D3.1 chapter 2.5.4 Reuse of AIM in ScaleAgData.

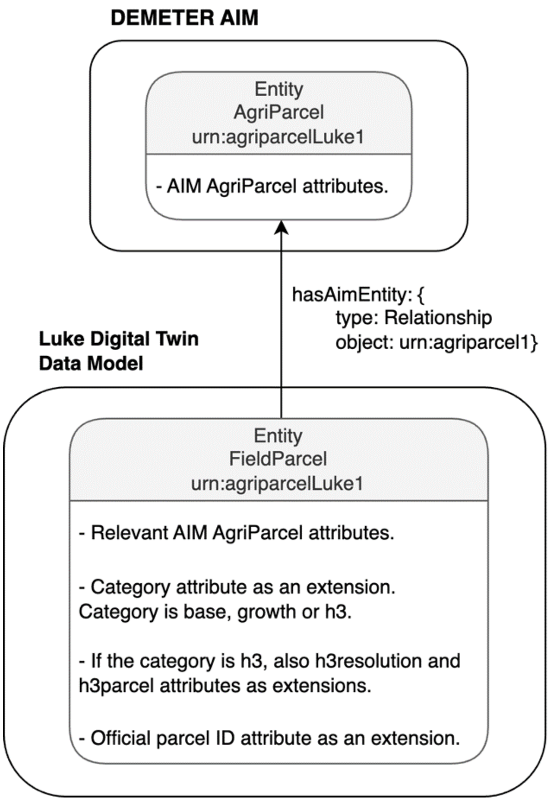

Figure 2. Digital twin data model compatibility with DEMETER AIM ontology

2.4.2.2. Entity specifications

All digital twin data model entities can be transferred through the same NGSI-LD context brokers. Context broker is a component implementing NGSI-LD interfaces, like saving and fetching the entities (ETSI, 2021). The entities defined so far are briefly described in Table 2. These entities are created with soil properties, crop management and field boundary data mentioned in Table 1. Data sources excluded from this round of entity specifications may come from standardized files or will be added to the digital twin data model in the next iteration of ScaleAgData and presented in V2 of this document. If a vocabulary and ontology were used as a basis for an entity, it is mentioned in the table. If the existing ontologies did not sufficiently meet the requirements, some entities were defined manually from scratch. The digital twin data model, including entity specifications and example Jupyter notebook, is part of the reference software. Vocabularies and ontologies mentioned in Table 2 are:

- Smart Data Models AgriFood vocabulary (https://smart-data-models.github.io/dataModel.Agrifood/),
- TERRA-REF ICASA ontology (https://terraref.github.io/icasa/1.0-alpha/core/),
- AgMIP ICASA vocabulary (https://docs.google.com/spreadsheets/u/0/d/1MYx1ukUsCAM1pcixbVQSu49NU-LfXg-Dtt-ncLBzGAM/pub?output=html).

Table 2. Entities of the digital twin data model

Entity	Vocabulary	Ontology	Extensions	AIM counterpart
AgriFarm	Smart Data Models Agrifood	-		AgriFarm
FieldParcel	-	-		AgriParcel
SoilProfileLayer	AgMIP ICASA	TERRA-REF ICASA		Soil
Genotype	AgMIP ICASA	TERRA-REF ICASA		CropSpecies
TillageEvent	AgMIP ICASA	TERRA-REF ICASA	X	AgriParcelOperation
Planting	AgMIP ICASA	TERRA-REF ICASA	X	AgriParcelOperation
FertilizerApplication	AgMIP ICASA	TERRA-REF ICASA	X	AgriParcelOperation
OrganicMaterial Application	AgMIP ICASA	TERRA-REF ICASA	X	AgriParcelOperation
HarvestEvent	AgMIP ICASA	TERRA-REF ICASA	X	AgriParcelOperation

2.4.2.3. Spatial representation

Since our digital twin describes variation within a field, a spatial representation of a field is required. The technical implementation should be computationally efficient, with a fixed grid location and suitable grid resolution. The digital twin data model supports running the model either at parcel level or using a discrete global grid (DGG) system H3 (https://h3geo.org/), see Figure 3 for example. H3 is a hexagonal hierarchical geospatial indexing system that stores cell indexes instead of coordinates, which makes spatial queries more efficient. The system is also supported by several geospatial databases and mapping libraries. H3 grid cells have a fixed location and shape, however, the cell area varies relative to its position in the grid's icosahedron vertices. In the case of H3 resolutions, 11 (~0.2 ha) and 12 (~0.03 ha) are possible resolution choices for precision farming applications, but the digital twin data model can easily be extended to support other local or global grid systems.

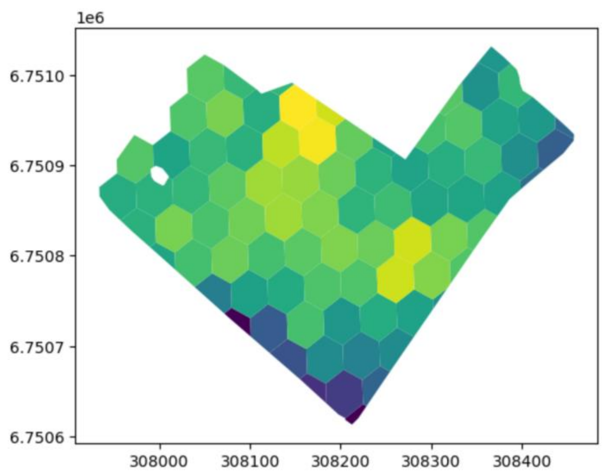

Figure 3. Example of H3 spatial representation of a yield map with H3 grid resolution 11

2.5. Potential application cases

The methodological framework will be tested together with the RILs. The following application cases have been defined during the 1st iteration phase of the project. First one of these application cases (co-operation with Yield Monitoring Lab) has now proceeded to a stage where the example code for setting up the simulation can be presented (Chapter 3). For the Water Productivity Lab and Crop Management Lab, the work will now continue according to each RIL:s schedules. The experiences gathered here will be translated into new or updated guidelines in the second version (V2) of this document.

2.5.1. Yield monitoring lab

2.5.1.1. Target system and decision problem

This use case in cooperation with CNHi and UGent 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. Combining data from multiple harvesters/sensors operating on a single field can result in erroneous yield maps. The aim of using a digital twin is to simulate yield variation and wheat protein content variation within a field 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.

2.5.1.2. Physical counterpart / characteristics

The field data that are used to set up this digital twin model are collected on a set of wheat fields provided by CNHi and located in and around Nijvel (Belgium). Historical data from 2022 or 2023 (yield, soil EC, vegetation index) are available for 8 fields. Soil moisture and evapotranspiration estimates for these fields are provided by DHI. Daily weather station data (minimum, maximum and average temperature, relative humidity, wind speed, precipitation, and radiation; interpolated data from several weather stations) are provided by KMI.
In addition, a more extensive dataset is collected during the summer of 2024 on 4 fields. Sensor data (yield, soil EC, vegetation index) and management data is again provided by CNHi. Soil moisture and evapotranspiration data will be provided by DHI and weather station data by KMI. UGent is responsible for additional measurements. Six locations on each field are subjected to measurements every month (Soil moisture, soil N content, plant fresh and dry weight, plant N content, leaf characteristics (chlorofyll, SPAD, ...), LAI).

2.5.1.3. Expected pay-off for creating a digital twin for the target system

The initial motivation for this case is to estimate spatial variation in yield to fill in yield maps with missing data. However, it provides the opportunity to use the developed model for other applications as well, such as improving variable rate fertilizer maps, or providing farmers who do not have yield sensors on their harvesters with an estimation of the yield variation in their field.

2.5.1.4. Types of models

For estimating the spatial variability, any model able to simulate differences in yield based on spatial variation present in EO data is suitable for this case study, and a simple process-based model or a machine learning model could suffice. For supporting optimal nitrogen application rates, the model needs to provide reliable biomass and protein yield estimates; in this case a more complex model, such as APSIM or a hybrid model, may be more suitable.

2.5.1.5. Potential scaling up of the approach

Key-point in scaling up is to nullify the need to perform in-field measurements. This case study can be used to test if weather station data and EO data suffice to provide accurate estimates of yield variation in a field. The absolute value of the yield estimates is of less importance as yield sensor data will often not be available for a farmer. Also, the aim of this application case is to fill in gaps that are due to missing sensor data. Relative yield values within a field are more important, and more feasible, since a cultivar-specific calibration is not possible without in-field measurements.

2.5.2. Water productivity lab

2.5.2.1. Target system and decision problem

This use case, in cooperation with IES and MIGAL, is targeting to optimize irrigation of peppermint in Latvia and quinoa in Israel by predicting potential yield / biomass as well as crop water status (normal / stressed). The aim is to support irrigation decision making to maximize yield and minimize water consumption. It is expected to demonstrate at least 20% increase in productivity and at least 20% decreased water consumption using the developed support system.

2.5.2.2. Physical counterpart / characteristics

Field data are collected on individual cultivated and irrigated fields situated in Latvia (peppermint case) and Israel (quinoa case) with an average size of 1 ha. At least four different irrigation regimes are tested in these fields. Local meteorological stations are located near the fields for continuous monitoring of natural precipitation, air temperature, humidity, pressure, wind speed and direction as well as solar irradiation. Soil moisture and temperature probes are placed in each field in ~10 cm depth. Data acquisition frequency is at least twice an hour. In parallel, airborne spectral and thermal data acquisition is planned for potential spatial upscaling of the point-based results to show result variation within fields. It is also planned to explore the potential of satellite data (ET, soil moisture and vegetation indices data products) to upscale the model to a larger area.

2.5.2.3. Expected pay-off for creating a digital twin for the target system

Timely information on crop status is required to more support effective irrigation control, which would lead to minimization of water consumption while maximizing the yield.

2.5.2.4. Types of models

Any models able to predict yield and water status of the target crop, primarily, based on meteorological data, secondarily, based on EO data could be used as a baseline.

2.5.2.5. Potential scaling up of the approach

It is expected to obtain the best prediction performance when onsite meteorological and soil sensor data will be used. However, it is planned to test the upscaling potential of the model in a simplified way based on EO data products – ET, soil moisture, vegetation indices.

2.5.3. Crop management lab

2.5.3.1. Target system and decision problem

The target of this use case is the support of crop management, by integrating data collected by IoT sensors, EO imagery and farm log data, aiming at the optimization of agricultural practices. Also, it will support policy makers to promote more efficient and sustainable agricultural practices.

2.5.3.2. Physical counterpart / characteristics

Data are collected on wheat fields located in Northern Greece, in the municipality of Kilkis. In total, 4 pilot parcels are included in the research with an average field size of 2.0 ha. In two of the pilot parcels, IoT stations are installed, continuously collecting data about the atmospheric and soil conditions of the fields, during the growing period. Also, in all 4 parcels farm log data regarding cultivation practices and phenological stage of the plants are being recorded.

2.5.3.3. Expected pay-off for creating a digital twin for the target system

The main motivation for this simulation is to analyse the variation of nutrients and water within the fields and promote more efficient use of available resources. Through this, the expected yields will be increased while the environmental footprint of the farms will be reduced. At a higher level, policy makers will be provided with the data needed to support best agricultural practices and programs.

2.5.3.4. Types of models

Several models could be utilized for the simulation of crop growth, soil dynamics, weather conditions and all the factors that are related to crop management. APSIM is one of the suitable models for this application, but also other models could be employed.

2.3.3.5. Potential scaling up of the approach

Scaling up of this approach involves extending the application of the digital twin from individual farms to larger regions, such as administrative regions. This scaling up can lead to the achievement of broader benefits in agricultural productivity and sustainability and the promotion of more cost-effective and environmentally friendly agricultural practices on a larger scale.

The key points for the up-scaling include the need of extensive data collection and standardization, cloud computing resources, the use of enhanced models to analyse the collected data and the development of user-friendly platforms through which the users will interact with the system. All the above, along with the engagement of a broad range of stakeholders and the adoption of supportive policies, will help ensure the successful scaling up of the digital twin approach in crop management.

3. Reference software

3.1 Software description

Reference software for digital twins has been implemented using Python and C# programming languages. The code is provided as part of Python package farmingpy (https://github.com/TwinYields/farmingpy), which is released under MIT license. The installation process is documented, and a Docker file is given for reproducible installation.

The software library provides the following functionality:

- Interfacing APSIM simulation model for setting up high resolution spatial simulations, running model ensembles for assimilating EO and sensor data with the APSIM simulation models, reading the simulation outputs and optimizing model parameters.
- Reading ISOBUS task data from tractors and combine harvesters.
- Unified interface to USDA Rosetta and EUPTF2 pedotransfer function models to estimate soil water holding capacity based on soil data available from farms.

C# components are required to interface the APSIM simulation model and are called from Python code using the Python.NET library. The user of the library does not need to use C# language, but will need to install the dotnet SDK. The interface to EUPTF2 models is implemented by calling the official R package (https://github.com/tkdweber/euptf2) using the rpy2 library. The installation process for Ubuntu Linux is detailed in the documentation and a docker configuration is provided to use the library and run the example code in a reproducible environment. The APSIM interface was significantly extended in the ScaleAgData project and entire data assimilation extension for APSIM wheat model was implemented and tested, the pedotransfer function interfaces were also implemented in the project. Additionally, the documentation for the library was extended and example notebooks were produced and published.
The core modeling component of the library for digital twins is the interface to the APSIM Next Generation simulation model. APSIM is very often used with a graphical user interface (UI). However, the implementation separates user interface from model implementation to separate C# projects (https://github.com/APSIMInitiative/ApsimX). The APSIM interface in the farmingpy library wraps the APSIM “Models” library and does not depend on the user interface. However, it is recommended to use the APSIM UI to create a base model, which can be modified in the Python library.

Software for the digital twin data model converts field data to NGSI-LD compatible entities defined by the schema of the digital twin data model. The schema, defined with OpenAPI specification, is converted to Pydantic (https://docs.pydantic.dev/) models. The Pydantic models are used for creating NGSI-LD compatible entities, firstly as Python class objects, secondly as JSONs sent to the context broker. The documentation for the digital twin data model is currently given in the form of two example notebooks. “Data_to_datamodel.ipynb” converts the example data of one field into NGSI-LD entities and sends those to a NGSI-LD broker. The notebook “data_from_datamodel.ipynb” then retrieves the data from the broker. Example notebooks can be downloaded from Zenodo: https://doi.org/10.5281/zenodo.12566023 .

3.2. Example use case and notebook: Yield lab

In ScaleAgData, the APSIM (The Agricultural Production Systems sIMulator) Next Generation model (Holzworth et al. 2014) was selected for wheat yield estimation in the Yield Lab. The selection was based on model completeness, modularity and availability of simulated soil and crop conditions as well as number of supported crops. In the future, other crop models and modeling approaches may be considered. The APSIM model simulates the whole cropping system including soil nutrient and carbon flows, water balance and crop development. The development of the crop is divided into several phenological stages, which are controlled by model parameters that need to be calibrated for each cultivar. The development of different crop organs and C and N allocation between them are also simulated. The model simulates all the processes with a daily time step. APSIM simulated the development of grain protein based on CERES Wheat model as described in (Asseng et al. 2002).

Example code is provided for interfacing APSIM simulation model ScaleAgData Yield Lab use case as a Jupyter notebook "Introduction_to_APSIM_interface.ipynb" on Zenodo https://doi.org/10.5281/zenodo.12566023. The notebook provides example code for setting up a wheat simulation model including:

- Setting the weather data and simulation date.
- Setting management data: sowing date, fertilizer amounts and fertilization dates.
- Setting soil properties based on farmers' soil data using pedotransfer functions.
- Changing cultivar parameters which will be adjusted in calibration.
- Reading simulation model outputs.

The model, as set up in the example, provides grain yield, protein content and harvest date forecasts with daily time step. The model also provides estimates of nitrogen stress and water stress which can be used to make management decisions. The example code covers initialization of the model and the use of model ensembles which can be used in data assimilation.

Figure 4 shows a schematic on how the model will be used in ScaleAgData. The simulation will first be initialized based on FMIS and machinery data. Once the model has been initialized it will be run daily to simulate crop and field state and provide forecasts. Before initialization, the cultivar used in the model needs to be calibrated with historical data. Remotely sensed LAI and potentially other biophysical parameters will be assimilated with the model. In the real time setting, historical daily weather data, weather forecast for the end of the season and remote sensing data need to be obtained separately.

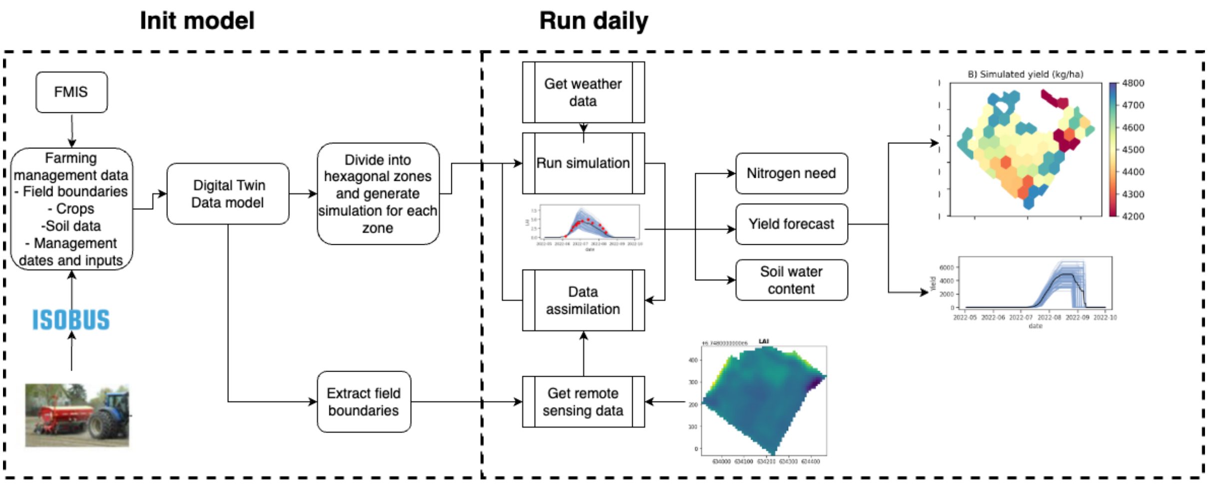

Figure 4. Schematic view of model use

4. References

Bloch, V., Palosuo, T., Huitu, H., Ronkainen, A., Backman, J., Pussi, K., Suokannas, A., Pastell, M., 2023. 46. Towards a digital twin for optimal field management, in: Precision Agriculture ‘23. Wageningen Academic Publishers, pp. 377–383. https://doi.org/10.3920/978-90-8686-947-3_46

Brisson, N., Gary, C., Justes, E., Roche, R., Mary, B., Ripoche, D., Zimmer, D., Sierra, J., Bertuzzi, P., Burger, P., Bussière, F., Cabidoche, Y.M., Cellier, P., Debaeke, P., Gaudillère, J.P., Hénault, C., Maraux, F., Seguin, B., Sinoquet, H., 2003. An overview of the crop model stics. European Journal of Agronomy, Modelling Cropping Systems: Science, Software and Applications 18, 309–332. https://doi.org/10.1016/S1161-0301(02)00110-7

Duchemin, B., Maisongrande, P., Boulet, G., Benhadj, I., 2008. A simple algorithm for yield estimates: Evaluation for semi-arid irrigated winter wheat monitored with green leaf area index. Environmental Modelling & Software 23, 876–892. https://doi.org/10.1016/j.envsoft.2007.10.003

Droutsas, I., Challinor, A.J., Deva, C.R., Wang, E., 2022. Integration of machine learning into process-based modelling to improve simulation of complex crop responses. in silico Plants 4, diac017. https://doi.org/10.1093/insilicoplants/diac017

European Telecommunications Standards Institute [ETSI] (2021) ETSI GS CIM 009 V1.5.1, ETSI [Online]. Available at https://www.etsi.org/deliver/etsi_gs/CIM/001_099/009/01.05.01_60/gs_cim009v010501p.pdf (Accessed 26 June 2024).

Gobbo, S., De Antoni Migliorati, M., Ferrise, R., Morari, F., Furlan, L., Sartori, L., 2022. Evaluation of different crop model-based approaches for variable rate nitrogen fertilization in winter wheat. Precision Agric. https://doi.org/10.1007/s11119-022-09957-5

Holzworth, D.P., Huth, N.I., deVoil, P.G., Zurcher, E.J., Herrmann, N.I., McLean, G., Chenu, K., van Oosterom, E.J., Snow, V., Murphy, C., Moore, A.D., Brown, H., Whish, J.P.M., Verrall, S., Fainges, J., Bell, L.W., Peake, A.S., Poulton, P.L., Hochman, Z., Thorburn, P.J., Gaydon, D.S., Dalgliesh, N.P., Rodriguez, D., Cox, H., Chapman, S., Doherty, A., Teixeira, E., Sharp, J., Cichota, R., Vogeler, I., Li, F.Y., Wang, E., Hammer, G.L., Robertson, M.J., Dimes, J.P., Whitbread, A.M., Hunt, J., van Rees, H., McClelland, T., Carberry, P.S., Hargreaves, J.N.G., MacLeod, N., McDonald, C., Harsdorf, J., Wedgwood, S., Keating, B.A., 2014. APSIM – Evolution towards a new generation of agricultural systems simulation. Environmental Modelling & Software 62, 327–350. https://doi.org/10.1016/j.envsoft.2014.07.009

Tanaka, T.S.T., Heuvelink, G.B.M., Mieno, T., Bullock, D.S., 2024. Can machine learning models provide accurate fertilizer recommendations? Precision Agric. https://doi.org/10.1007/s11119-024-10136-x

Palma, R., Roussaki, I., Döhmen, T., Atkinson, R., Brahma, S., Lange, C., Routis, G., Plociennik, M., Mueller, S., 2022. Agricultural Information Model, in: Bochtis, D.D., Sørensen, C.G., Fountas, S., Moysiadis, V., Pardalos, P.M. (Eds.), Information and Communication Technologies for Agriculture—Theme III: Decision. Springer International Publishing, Cham, pp. 3–36. https://doi.org/10.1007/978-3-030-84152-2_1

Pylianidis, C., Snow, V., Overweg, H., Osinga, S., Kean, J., Athanasiadis, I.N., 2022. Simulation-assisted machine learning for operational digital twins. Environmental Modelling & Software 148, 105274. https://doi.org/10.1016/j.envsoft.2021.105274

Pylianidis, C., Osinga, S., & Athanasiadis, I. N., 2021. Introducing digital twins to agriculture. Computers and Electronics in Agriculture, 184, 105942. https://doi.org/10.1016/j.compag.2020.105942

Verdouw, C., Tekinerdogan, B., Beulens, A., & Wolfert, S., 2021. Digital twins in smart farming. Agricultural Systems, 189, 103046. https://doi.org/10.1016/j.agsy.2020.103046

De Wit, A., Boogaard, H., Fumagalli, D., Janssen, S., Knapen, R., van Kraalingen, D., ... & van Diepen, K., 2019. 25 years of the WOFOST cropping systems model. Agricultural systems, 168, 154-167. https://doi.org/10.1016/j.agsy.2018.06.018