PROJECT DESCRIPTION – master's thesis, Department of Digital Agriculture

Project title: Assessing the usability of long-term experiment data for crop model validation

Supervisor

Prof. Dr. Senthold Asseng

Professor, chair of Digital Agriculture, TUM School of Life Sciences

senthold.asseng@tum.de

Background

Crop simulation models have become important tools in agricultural research and crop systems analysis. Process models that simulate biological, physical, and chemical processes governing crop growth and development are particularly data-hungry, requiring the input of various agronomic and environmental data for model validation and calibration. The integration of these heterogeneous source materials requires experts in a range of disciplines for locating, accessing, and transferring data, as well as workflows for data integration, transformation and quality control. Currently, the heterogeneity of data quality and formats, the scattered nature of data sources and a lack of widely adopted domain-specific standards for research data management make data preparation for crop modelling a cumbersome and hardly replicable process. One still largely untapped source of agronomic data for model validation is long-term field experiments (LTE). This is mainly due to low accessibility and interoperability, and data quality issues that are compounded by the long duration of the experiments. These issues complicate the reuse of LTE data, which currently requires tedious data curation processes. The German funding initiative BonaRes aims to facilitate publication and improve the reusability of LTE data by developing an infrastructure compliant with the FAIR principles (Findable, Accessible, Interoperable, Reusable). It notably includes a metadata catalogue of more than 570 LTEs and a data repository leveraging a semi-structured data model to publish experimental results in a standardized format. This infrastructure significantly enhances the potential for reuse by addressing key aspects of findability, accessibility, and interoperability. In contrast, data quality challenges need to be addressed with regard to the target application for reuse.

The FAIRagro consortium develops a data infrastructure harmonizing existing databases, repositories, and data standards to make agrosystem research data compliant with the FAIR principles. FAIRagro is structured around research projects that span various disciplines, data types, and scales. These use cases provide requirements, datasets, code, methods, and iterative feedback to support the development of FAIRagro services. One such use case, led by scientists at the Hans Eisenmann-Forum for Agricultural Sciences and the department of Digital Agriculture of the TUM, focuses on creating a workflow to acquire, standardize, and integrate data sources for crop model validation. It is primarily focused on the development of ETL (Extract, Transform, Load) processes to automate the integration of experimental data (manual measurements, IoT sensors, airborne remote sensing…) with secondary data (e.g., weather

time series and forecast, soil profiles, land-use maps…) and routines to calibrate model parameters and run simulations in the DSSAT framework.

Master's thesis project

The purpose of this master's thesis project is twofold: (1) to extend the ETL pipeline to incorporate published LTE data as an additional source of agronomic information and (2) to identify LTE data quality issues and implement solutions to address them. The data science component of the project will build on an existing prototype (R scripts) that transforms a specific LTE dataset into the ICASA format, the standard data model for multiple crop modeling frameworks. The student will generalize the pipeline to process all LTE datasets published in the BonaRes repository. The second objective will be addressed by attempting to reproduce experimental results of multiple years of selected LTEs, using the NWheat model in the DSSAT framework. The student will identify data quality issues and apply various methods to resolve them, such as imputation techniques, enrichment with secondary data, and retrieval of missing information from data producers. Throughout this process, they will document their approach while iterating through simulation/validation runs. The student is expected to conduct a detailed analysis of the simulation results, examining both the influence of data quality and the impact of experimental and environmental conditions on simulated crop yields.