Big Data Analytics for Democratizing Crop Modeling Adoption to Farmers

Crop modeling is an essential input to make decisions over climate change assessment, irrigation management, fertilizer management, crop improvement, pest and disease management, spatial analysis, tillage simulation and crop rotation, for low and big scale farming. Nowadays, monitoring crops is an effort that has been improved through years of technology enhancement that provides more and more accurate data for agro-climatic forecasts. Nonetheless, model parametrization is currently an activity that counts with few ready to use inputs, difficult understanding software, extensive outputs for analysis and criteria from professional experts hard to find.

The proposed solution is to build **a powerful scientific workflow as a service that can aid stakeholders to better understanding of models to represent crop varieties and conditions**, by maximizing the use of the current limited field-collected data in a novel methodology to make a massive batch processing of current crop models, such those available in DSSAT© and ORYZA©, with a plethora of combinations of the genetic values of each cultivars and planting conditions. This simulation outputs structured in an analytical database will be used for a quantitative analysis that help to sort the universe of results at big scale according to their goodness-of-fit. Providing a very precise configuration of very specific variety and agroecosystem conditions of interest to the farmers and scientist, without make them deal with the whole model use process but empower them with knowledge to support the take actions.

In simulations, primary data is a crucial input that represents a major cost, the more accurate and precise the better for model results. Even though the gather of this data is a necessary task, the information generated during simulations runs and their analysis is where the computational challenge resides. Current crop model calibration process only use single average measurements, take several months to perform correctly, and large human effort. Sometimes even if the results have some accuracy they are not quite precise and requires sensitive analysis to validate the findings.

For calibration, this proposal aims to consider not only single measurements but also observations of the whole crop life-cycle, a wider range of the inputs to explore and the capacity to identify the proper values from a massive amounts of generated data, giving results with better resolution. Then**, by using response surface methodology and machine learning techniques it can be selected the single combination that fits the best for the cultivar characteristics**. Taking in count that runs configurations can be build up with different conditions of location, soil, fertilizing, irrigation, pest and disease, and other features supported by the models giving the possibility of scale to process quicker since it will be installed into a distributed and fault-tolerant processing system. This analysis will provide enough information to **ask users only meaningful variables for models use and not make them interact with whole system configuration.**

The service will require and use of internationally accepted agrobiodiversity and ecological data standards to assure interoperability and reusability of the data. This project will focus on the processing and analysis, oriented in the use of data available in literature and digital repositories as CGIAR-AgTrials and *Agricultura Específica Por Sitio* (AEPS), targeting in corn, common bean and rice crops. This data will be indeed curated to accomplish the quality standards required to be used for configuration and evaluation.

The pilot is composed of three principles: **1)** the appliance of Big Data Analytic methodologies to generate simulated data oriented to crop modeling calibration, its understanding and exploitation. **2)** The knowledge transfer to targeted users of crop simulation systems by reducing the steps and the number of inputs to run models. **3)** Develop of methodologies of field sampling to lesser the error prone and its derived cost.

The scope of this pilot is to develop the Minimum Viable Product (MVP) that support the all automatized run of the simulations and the analytical results of them using modularized and orchestrated computing containers for system automating deployment, scaling and management; statistically identification the relevant variables in models; and first version of user interfaces.

The deliverable will be a Software as a Service (SaaS), for both **precise crop variety calibration** and **cross-platform user-friendly model simulation,** as an upper layer that can be used as tool for universities, research centers, NGOs, agricultural federations, grain traders, risk analysts and local farmers to enhance current agro-climatic forecast and empower data-driven farming. Initially, the service will be tested with corn information gathered in Philippines and Ecuador, which are in a current calibration development, the review from literature and data from repositories. Besides, the transition of the methodology of corn models, from CERES that is already tested to IXIM.

Budget distribution	**Value**
Cloud computing services and infrastructure	$12.000
Big Data Information Architect	$10.000
Modules development (ETL, analytical engine, APIs and visualization)	$40.000
Containers configuration, orchestration and test	$6.000
Software Quality Assurance (QA) Analyst	$12.000
Big Data Scientist	$8.000
Crop modeling expert	$12.000
Total	**$100.000**

Once the pilot is up and running, the next step is to develop the solution complete front-end, with User Centered-Design (UCD) interfaces for Web and mobile environments. Also, replication of the methodology in rice on the involved places using in ORYZA model.

Believing in the expertise of the parts involved, CIAT with crop modeling develop and appliances, and Universidad Complutense de Madrid (UCM) competences of using computational-statistics for information solutions, this will a great opportunity to enable a strong collaboration for applied sciences in agriculture.