INnovations in plant VarIety Testing in Europe

Deliverable D4.2
Crop growth model predictions for environmental characterization

This deliverable is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 817970.

Technical References

Project Acronym	INVITE
Project Title	INnovations in plant VarIety Testing in Europe
Project Coordinator	François Laurens
Project Duration	60 months

Deliverable No.	D4.2
Dissemination level [1]	
Work Package	WP 4 - Modelling for prediction of variety performance and application in a DSS for variety choice
Task	T 4.2 - Genotype specific crop growth modelling across management and environmental scenarios
Lead beneficiary	1-INRAE AGIR
Contributing beneficiary(ies)	1-INRAE LEPSE, 22- ACTA Terres Inovia
Due date of deliverable	31 December 2022
Actual submission date	04 January 2023

[1] PU = Public
PP = Restricted to other programme participants (including the Commission Services)
RE = Restricted to a group specified by the consortium (including the Commission Services)
CO = Confidential, only for members of the consortium (including the Commission Services)

Document history

V	Date	Beneficiary	Authors
1	15/12/2022		Franck Boizard, Pierre Casadebaig, Philippe Debaeke (INRAE, UMR AGIR, Toulouse, France, P1) ; Emmanuelle Mestries (Terres Inovia, France, P22)
2			
3			
4			

Summary

This deliverable is about the possibility to use crop growth models for predicting grain yield at variety and trial level when applied to multi-environmental trials used for varietal evaluation. This was first applied to sunflower and Terres Inovia post-registration trials (1431) from 2003 to 2020. This document pointed out the importance of an accurate soil characterization for final prediction with the SUNFLO crop growth model. In spite of uncertainties in predicting the response of a wide range of sunflower cultivars, the model could be used for the environmental characterization of each trial when sufficient crop management, climate and soil data are available.

Table of content

1 Introduction

Crop growth models (CGM) – or process-based models - simulate the dynamic responses of crops (or genotypes, G) as a function of environmental conditions (E) and management practices (M) and hence are appropriate tools to predict and explain G×E×M interactions (Chapman, 2008).
Therefore CGMs could have practical applications for improving the design and evaluation of multi-environment VCU trials (Jeuffroy et al., 2014 ; Casadebaig et al., 2016 ; Mangin et al., 2017); they could be also embedded in decision support systems used for variety choice. But for that purpose, crop models should be made more genotype specific (Wang et al., 2019).
A main objective of INVITE was to calibrate crop models for high-resolution prediction at genotype-specific level instead of at the current crop level (T4.2). This was done on wheat, maize and sunflower. However, a first deliverable (D4.2) will be focused on the application of a crop growth model developed for sunflower (SUNFLO) on a representative MET used for variety evaluation in order to evaluate the model performance at trial and variety level.
For sunflower, we have focused our study on the analysis of the Terres Inovia experimental network used for post-registration in France. It is the most extensive network in terms of years and number of trials, with the best quality data, in particular with access to the names of the varieties. This was an absolute pre-requisite for testing the performance of the model in varietal discrimination.
This document describes the results of yield simulation of the entire Terres Inovia network using the SUNFLO crop model (Casadebaig et al., 2011). After a short description of the data, we described the data curation and processing needed for simulation. Then we compared the simulation results according to different sources of soil characterization. Finally, we decomposed the results in terms of environmental and genotypic effects in order to evaluate the ability of the model to represent each term of yield variation.

2 Materials & Methods

1.1 A crop model: description of SUNFLO

SUNFLO crop model, specific to sunflower, was built by INRAE in collaboration with Terres Inovia (Casadebaig et al., 2011). Using different input variables characterizing climate, soil, cultivation practices or genotype characteristics, SUNFLO is a software program that simulates on a daily basis final crop productivity (grain yield, oil concentration), in-season variables (biomass, LAI, N uptake, N and water soil content…) and abiotic stress patterns (water, heat, cold or nitrogen) under different growing conditions (Figure 1).

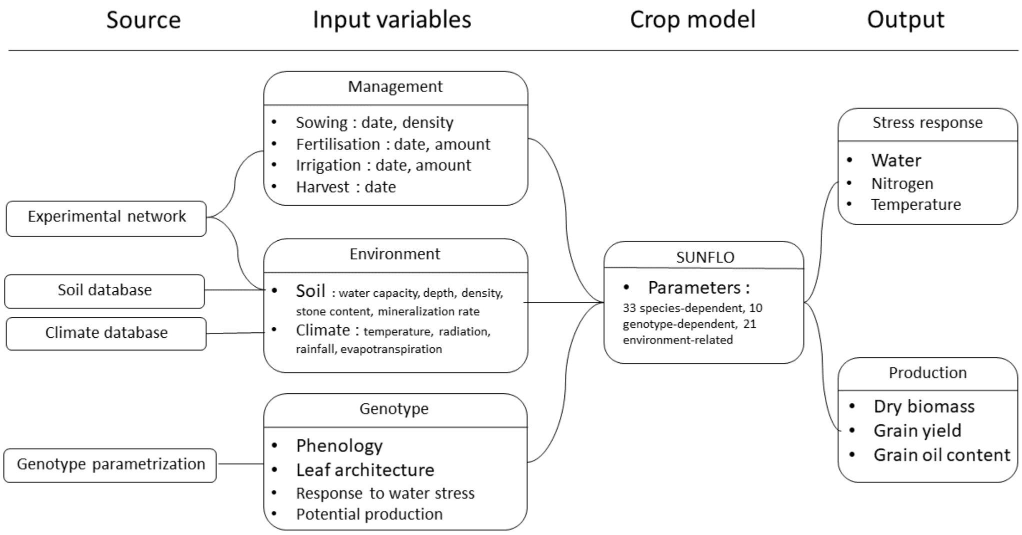


Figure 1. - ***Conceptual diagram of SUNFLO cultural model to identify different types of data.***

1.2 Experimental network: Terres Inovia (France)

The data used in this study came from the field trials used for post registration of newly registered sunflower varieties under the governance of Terres Inovia (The French technical institute for oilseeds, partner of INVITE). The dataset was composed of 1431 different trials from 2003 to 2020 spread all over France (Figure 2). The experimental network included between 36 and 105 trials each year.

Different information were collected on these plots: site identification (id, commune, year, ZIP code), plot management (date, depth and density of sowing, harvest date, N fertilization and irrigation dates and amounts, experimental design), soil characteristics (stoniness, soil depth, available water content), previous crop. Some years were particularly well documented, as 2009 and 2014-2007, in the frame of research projects in partnership with INRAE. But data from other years were often less complete for soil and management because the purpose of these trials was mainly the relative comparison of crop yield between varieties. Then these yield data are averaged by Terres Inovia to recommend the best yielding varieties at a regional level. This purely experimental approach is not based on a diagnostic of environmental conditions.

On average, a sunflower trial compares 11 different varieties. From 2003 to 2020, 397 unique varieties were compared with a total of 13916 observations (year x location x variety). Different characters can be observed on each variety: flowering date, yield (dry or standardized), thousand seed weight, oil concentration, oil quality, grain moisture percentage, impurity percentage, protein percentage, plant height. Only yield was systematically reported.

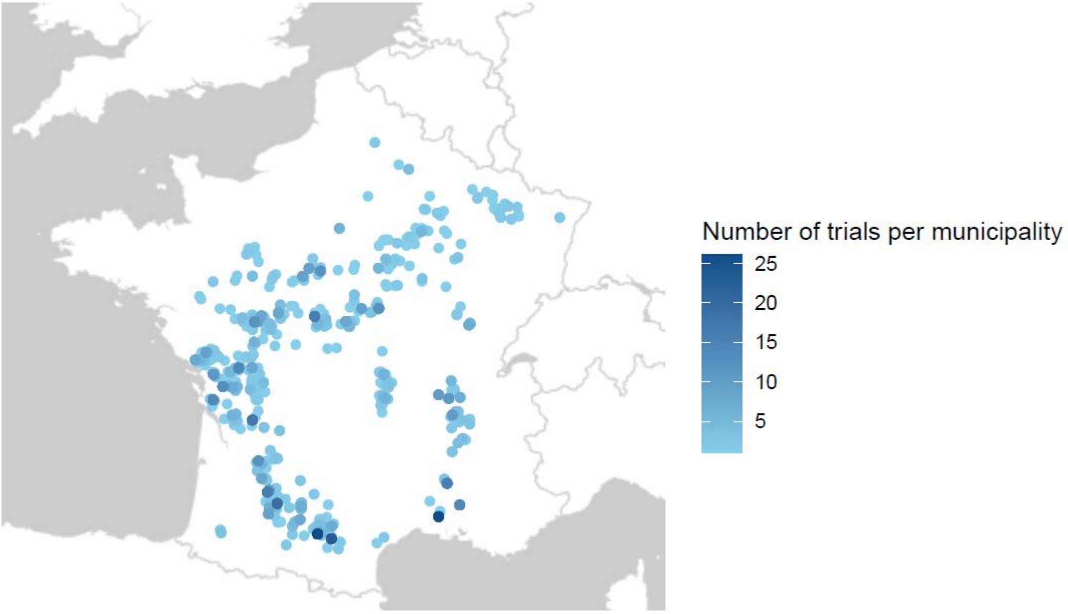


Figure 2. *Geographical distribution of field trials conducted between 2003 and 2020 by Terres Inovia.* *The trials are grouped by "commune" (township) which is the most confident geographical data we have. Each point represents a French commune and the color and size represent the number of trials carried out on the commune's territory over the years*.

1.3 Climate database

Climatic data used for each trial description were derived from the SAFRAN database. SAFRAN (Système d'Analyse Fournissant des Renseignements Atmosphériques à la Neige) uses surface observations combined with analysis data from meteorological models to produce climate variables such as precipitation, solid or liquid, hourly temperatures, wind and radiation. These profiles are then spatially projected on a regular 8 km grid over France.

This database provided the French daily climate data for all the years considered in our study. Each trial was associated with the grid cell having the closest center to the centroid of the trial's commune.

1.4 Soil database

Attached to the SAFRAN climate database, a description of the grid cells has been published (Bertuzzi, 2022). These files include several attributes attached to the SAFRAN mesh: geographic attributes (coordinates and altitude), land use, administrative division, and soil attributes. The soil data, coming from the Geographic Database of Soils of France (INRA, 2018), contains in particular the depth of the soil and the available soil water capacity, maximum and minimum. This data source will be referred to as the "French soil database" in the following.

The European Soil Data Centre (ESDAC) is the thematic center for soil related data in Europe. We used the European Soil Database derived data (Hiederer, 2013, http://esdac.jrc.ec.europa.eu/content/esdb-derived-data, 1x1 km grid cells) to obtain the quantitative soil properties corresponding to each location, i.e. total available water content, soil depth available to roots, bulk density, coarse fragments for two soil layers ([0, 30 cm],]30 cm, rooting depth]). This data source will be referred to as the "European soil database" in the following.

1.5 Genotype data

Twelve genotypic parameters are used to characterize the varieties and differentiate their functioning: four phenology parameters; four leaf architecture parameters; two parameters of response to water stress; two parameters of allocation of photosynthesis products to achenes (Casadebaig et al., 2011). These data are time-consuming to collect and were therefore not routinely estimated for all varieties grown in France. Most of these parameters are measured directly in the field or in the greenhouse (or outdoor pot platform) under semi-controlled conditions. We have phenotyped 133 varieties to date. Some additional varieties were phenotyped during the course of the INVITE project.

Table 1 – *Sunflo parameters: min, max and mean for 133 varieties*

			min 133	max 133	mean 133
TDE1	Temperature sum to floral initiation	°C.d	429	522	478
TDF1	Temperature sum from emergence to the beginning of flowering	°C.d	744	907	830
TDM0	Temperature sum from emergence to the beginning of grain filling	°C.d	991	1153	1077
TDM3	Temperature sum from emergence to seed physiological maturity	°C.d	1461	2055	1698
TLN	Potential number of leaves at flowering		22.2	36.7	29.1
LLH	Potential rank of the plant largest leaf at flowering		12.3	23.2	16.7
LLS	Potential area of the plant largest leaf at flowering	cm²	139	670	395
K	Light extinction coefficient during vegetative growth		0.78	1	0.90
LE	Threshold for leaf expansion response to water stress		-15.6	-2.1	-4.1
TR	Threshold for stomatal conductance response to water stress		-14.2	-5.8	-10
HI	Potential harvest index		0.25	0.51	0.40
OC	Potential seed oil content	% DM	47.8	62.3	56.4

3 Results

3.1 Curation of the dataset

Between 2003 and 2020, Terres Inovia conducted 1433 trials in France. A part of the trial network was not used in our study because it was incomplete. 12% of the trials were not harvested (Table 2, step 2) because they were experimentally invalidated before the harvest date. 7% of the trials were statistically invalidated after harvest because their yield was too low (diseases, heterogeneity…), some microplots were judged as outliers and some varieties were excluded from the study (Table 2, step 3). This curation led us to keep 81% of the trials that we considered as reliable. These data can be presented as the yield results of the Terres Inovia network. But a minimum of information was needed to simulate a trial with the SUNFLO model.

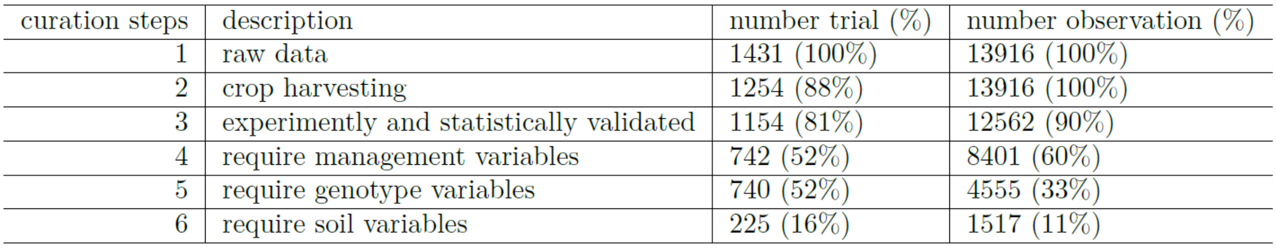
Several data were critical for the simulation with SUNFLO because it was not possible to define a default value in case of missing data. Half of the trials did not have sowing dates filled in (Table 2, step 4). Sowing date cannot be imputable with default value because it is highly variable from year to year and from site to site.

The second major limitation is the number of previously parameterized varieties. Only a quarter of the varieties, 133 out of 478 unique varieties in the raw dataset, have been parameterized (Table 2, step 5). After having selected the observations with corresponding varietal parameters, we kept 52% of the trials and 33% of the observations that can be simulated with the SUNFLO model (Table 2, step 5).

But these data were not complete from the perspective of SUNFLO. A large part of the trials were not described in terms of soil texture and depth. Only 16% of the soils were described by the experimenters. We had to input these data from other indirect data sources.

Three types of data are therefore very limiting in terms of number of observations: crop management, varietal phenotyping and soil description. We will not be able to retrieve the old crop management data and we do not have control over the phenotyping campaigns. But we can have access to soil data by different sources.

Table 2. *Curation steps and size of corresponding dataset*



curation steps	description	number trial (%)	number observation (%)
1	raw data	1431 (100%)	13916 (100%)
2	crop harvesting	1254 (88%)	13916 (100%)
3	experimently and statistically validated	1154 (81%)	12562 (90%)
4	require management variables	742 (52%)	8401 (60%)
5	require genotype variables	740 (52%)	4555 (33%)
6	require soil variables	225 (16%)	1517 (11%)

This loss of information was not the same over the years. Firstly, we observed that the number of trials conducted by Terres Inovia decreased since the year 2010 (Figure 3A). But the quality of the data, evaluated here by the proportion of trials that cannot be simulated (Figure 3B), has been increasing steadily for 20 years. The year 2014 was an exception: 66% of the sowing dates were missing.

The second major source of data loss for the simulation is the number of varieties parameterized. Here again, there is a difference depending on the year. A large amount of data was lost because we did not know the characteristics of the old varieties (Figure 4). Moreover, since 2016 we also observed this loss of data on the most recent years because the database was not updated anymore.

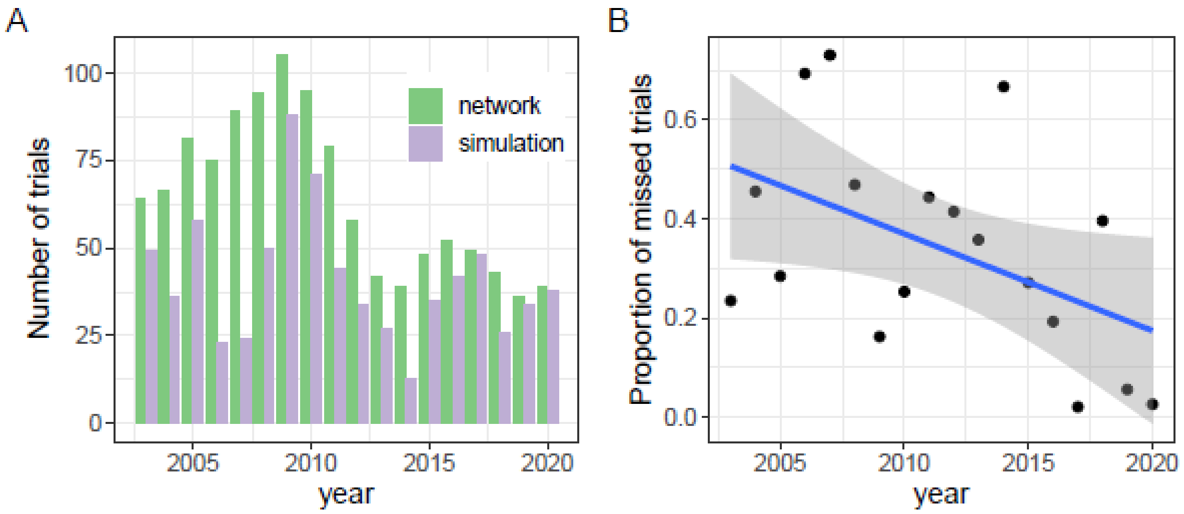


Figure 3. Comparison of the number of trials in the experimental network (A) and the proportion of possible simulations as a function of years (B).

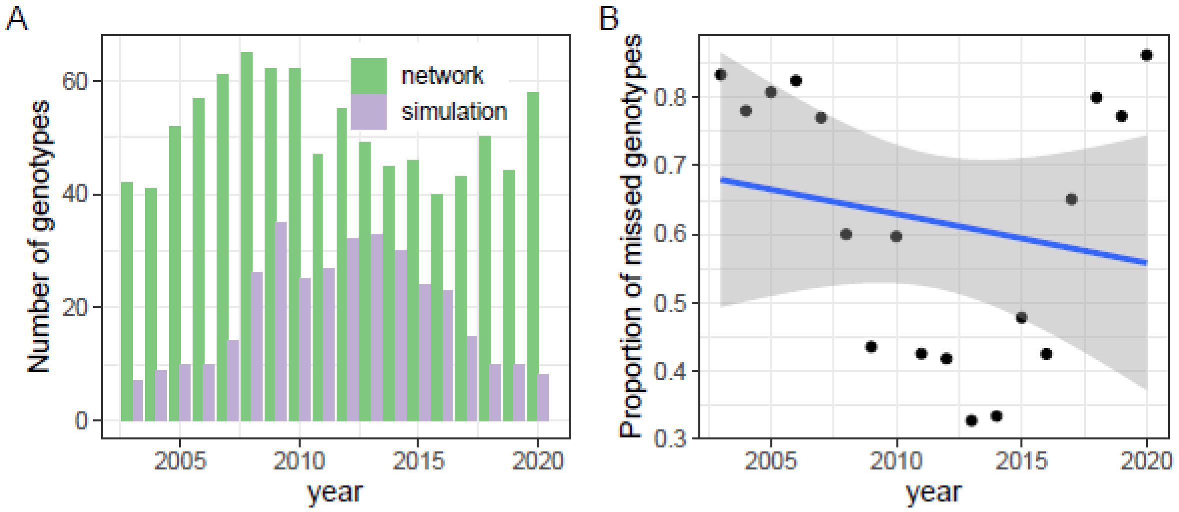


Figure 4. ***Comparison of the number of genotypes in the experimental network and the proportion of possible simulations as a function of years***

3.2 Simulation results with different soil data sources

We have seen that soil characterization is one of the main issues on our dataset to simulate a maximum number of data. Soil knowledge mainly details the accessibility of water for the plant: field capacity, wilting point, soil depth. Soil depth and water availability are sometimes measured or roughly evaluated by the experimenters (or experts). These variables have a strong influence on the simulation (Casadebaig, 2008). Due to their cost, other soil data are never measured. As they are necessary for the SUNFLO simulation, we evaluated these parameters through the available water and depth. To be able to simulate the trials with missing soil data from Terres Inovia experimenters, we used external data: imputation by standard values or databases. The soil characterization data being the main source of uncertainty in the simulation results (Casadebaig, 2008), each soil source gave different results.

3.2.1 Simulation with standard soils

The first way to impute missing data is to give a standard value for each trial. We compared the simulation results with the description of the soil by the experimenter, with 3 different standard soils: shallow soil (soil depth = 300mm, available water = 50mm), medium soil (soil depth = 1000mm, available water = 50mm) and deep soil (soil depth = 1800mm, available water = 250mm).

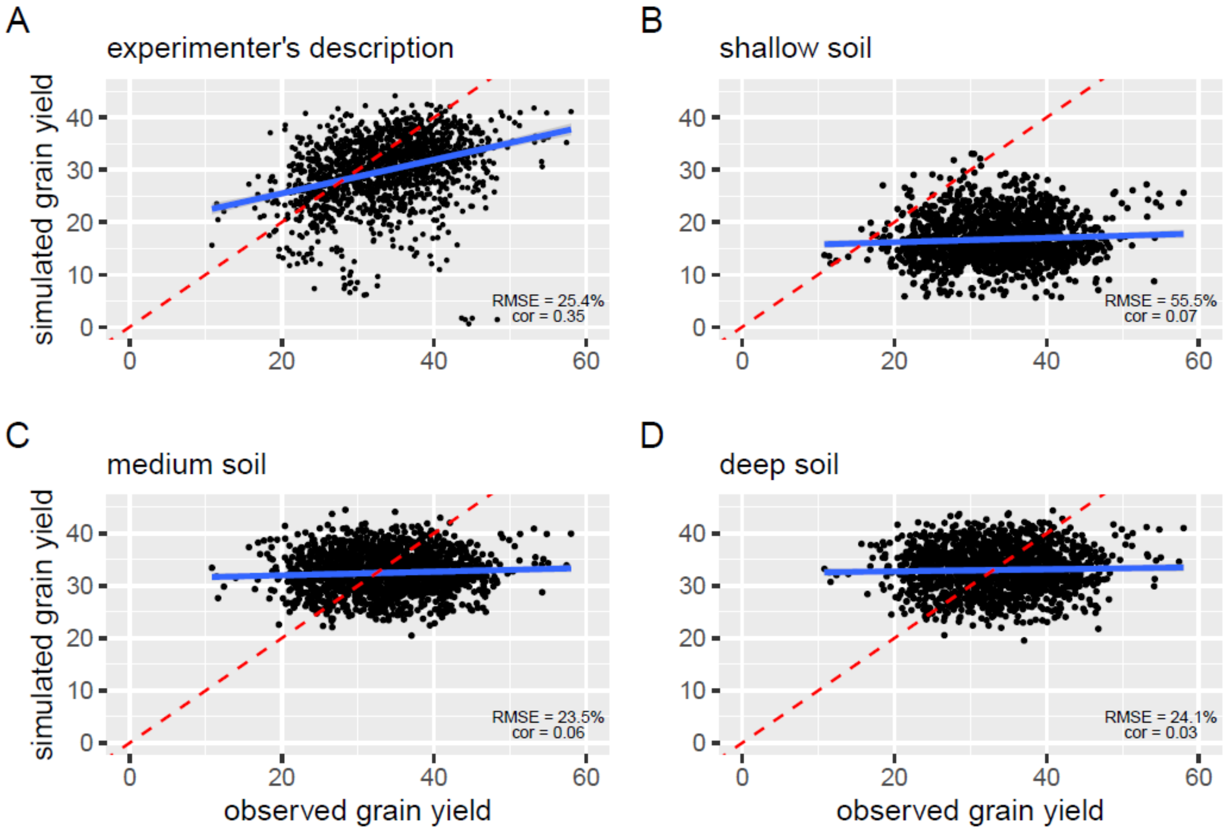


Figure 5 - ***Comparison of observed and simulated grain yield with the description of the soil by the experimenter (A) data and 3 different standard soils*** *: shallow soil (soil depth = 300mm, available water = 50mm)(B), medium soil (soil depth = 1000mm, available water = 50mm)(C) and deep soil (soil depth = 1800mm, available water = 250mm)(D). The SUNFLO model estimates a simulated grain yield to approach the experimentally observed yield. The scatterplots show the simulated yield, on the y-axis, and the observed yield, on the x-axis, according to the soil data source (A-D). To facilitate the interpretation of the results, the graph also shows the x=y axis, in red, and the linear regression*

curve in blue. In the case of perfect results, the points would be on the red line. The prediction error would be zero (RMSE = 0) and the regression line would follow the x=y axis with a slope of 1.

From the dots clouds we can observe that the soil has a great influence on the simulation results (Figure 5). The simulated yields from uniform soils (Figure 5B-D) are much less variable than results with soil determined by experimenters (Figure 5A). The simulations with standard soils did not succeed in separating much differences between the observations. Moreover, we see that the deeper the soil, the higher the simulated yield. The prediction error (RMSE), which is lower in the case of very deep soil, shows that the Terres Inovia trials were rather set up in very favourable soil conditions. Moreover, we observe that the correlation between observed and simulated yields is much better with the experimenter's soil data (A) than with a uniform soil (B-D). This approach emphasizes the added value of a sound knowledge of the soil type to improve the simulation results.

3.2.2 Comparison of soil characteristics from databases

Our first source of data, hereafter called 'experimenter', was provided by Terres Inovia. Soil depth and available water content were assessed by the experimenters who are probably familiar with soil quality and hydraulic properties. The disadvantage of this data source is the small amount of data available (Table 2). Another approach, "experimenter per municipality", assumes that the soils described by the experimenter on each site in the same territory are similar. Since the municipality is the most accurate and reliable geographic feature available, we collected the soil variables, depth and available water content, by municipality. We chose the deepest soil present in each municipality assuming that the trials are systematically set up on soils most suitable for sunflower cultivation and field experimentation. This method allowed us to simulate more sites (Table 2).

Databases can also provide soil data for each site. The SAFRAN database, hereafter called "French database", the same as the one we used for meteorological data, evaluates the available water content and the soil depth on a grid of 8602 meshes on the French territory. The European Soil Data Centre (ESDAC), a "European database", also produces a soil estimate at the European level. These two sources completed our knowledge of soils making possible to simulate all sites with sufficient data.

Here we compared the available water content (awc) for the 4 different data sources, taking as reference the most confident source, the experimenter's description of soils (Figure 6).

There is logically more water available in soil when considering the soil per municipality compared to the experimenter's data, as we kept the deepest soils per municipality (Figure 6A). Beyond that, the soils were very similar in these two designs. The French database was farther from the estimation of the experimenters (Figure 6B). The French database overestimated the amount of soil available water. The European database, on the other hand, did not result in bias but was still very far from the experimenter's estimates (Fig. 6C).

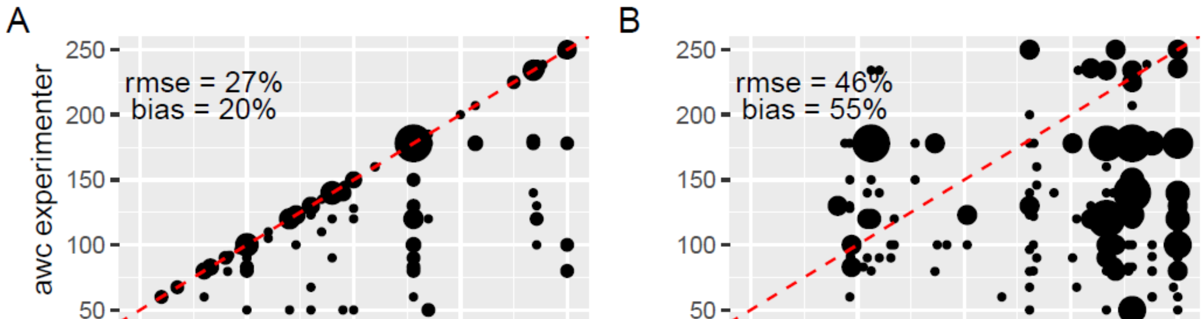


Figure 6. ***Pairwise comparison of experimenter available soil water content (awc) compared to other indirect data sources:*** *maximum awc per municipality (A), french database (B), european database (C).*

3.2.3 Results with different sources of soil characterization

We represent here the simulation results with the 225 trials (1517 observations) (Table 2) on which Terres Inovia evaluated the soil data in order to properly compare the results of the different soil data sources (Figure 7). From a global point of view, we observed an underestimation of the simulations compared to the observed yields, especially for high yields. Indeed, the simulated yield rarely exceeded 40 q.ha^{-1}, whereas some observed yields reached 60 q.ha^{-1}. This could be due to yield estimation on microplots where some border effects are sometimes difficult to control resulting in very high yields. The SUNFLO model was calibrated with data from larger experimental fields where no such border effects were encountered. Conversely, the low yields simulated by the SUNFLO model fall to 10 or less while the observed low yields are down to around 15 q.ha^{-1}. This is probably due to the fact that Terres Inovia selected favorable testing environments and discarded unsuccessful trials.

From a statistical point of view, the RMSE estimates a similar error between the different datasets with a fairly large error between 25 and 30%. But in terms of correlation coefficients and therefore
ir n
m r
tc

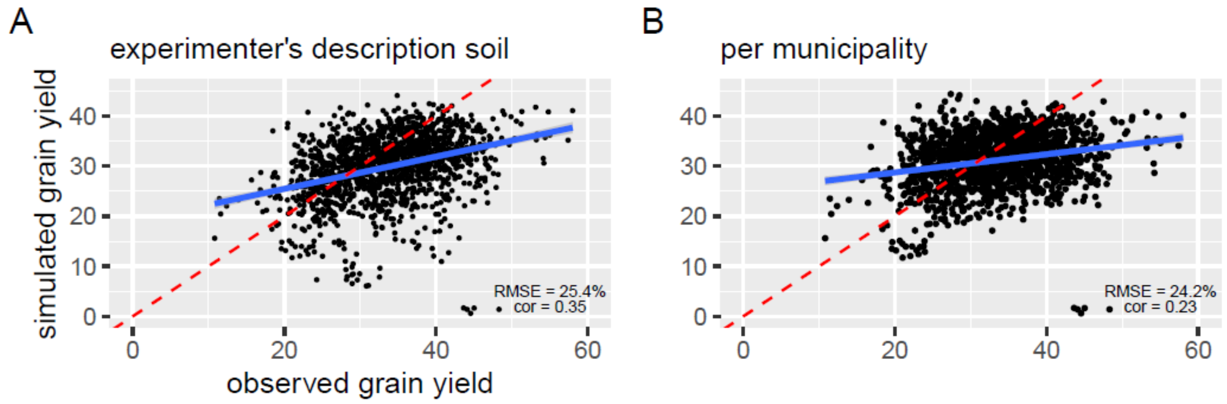


Figure 7. ***Comparison of observed and simulated grain yield with the experimenter's description of soil (A) data and 3 different strategies of imputation: imputed soil data per municipality (B), using a french soil database (C) and a european soil database (D).*** *The scatterplots show the simulated yield, on the y-axis, and the observed yield, on the x-axis, according to the soil data source (A-D). To facilitate the interpretation of the results, the graph also shows the x=y axis, in red, and the linear regression curve in blue.*

3.2.4 Results of simulation depending on the effort of data collection

We have seen that the data from experimenters gave the best simulation results (Figure 6-7, Table 2). But the quality of these data is not homogeneous depending on the years. In fact, the years 2016-2017 and in particular 2009 were characterized by a stronger collection effort, because Terres Inovia trials were integrated into research projects with Terres Inovia (CTPS and CASDAR funding). We can therefore consider that in these conditions the characterization of soils (and crop management) was of better quality.

The scatterplots and the related statistical indicators (Figure 8) show two types of years. Years with routine soil data collection (2002-2008, 2010-2015, 2018-2020), where the simulation results are less variable, with a low correlation coefficient. For years with soil data collected as part of a research project, simulation results were obviously better. This experience highlights once again the importance of a precise characterization of the soils for getting satisfactory simulation results. This is especially true for sunflower grown in spring and summer under rainfed conditions.

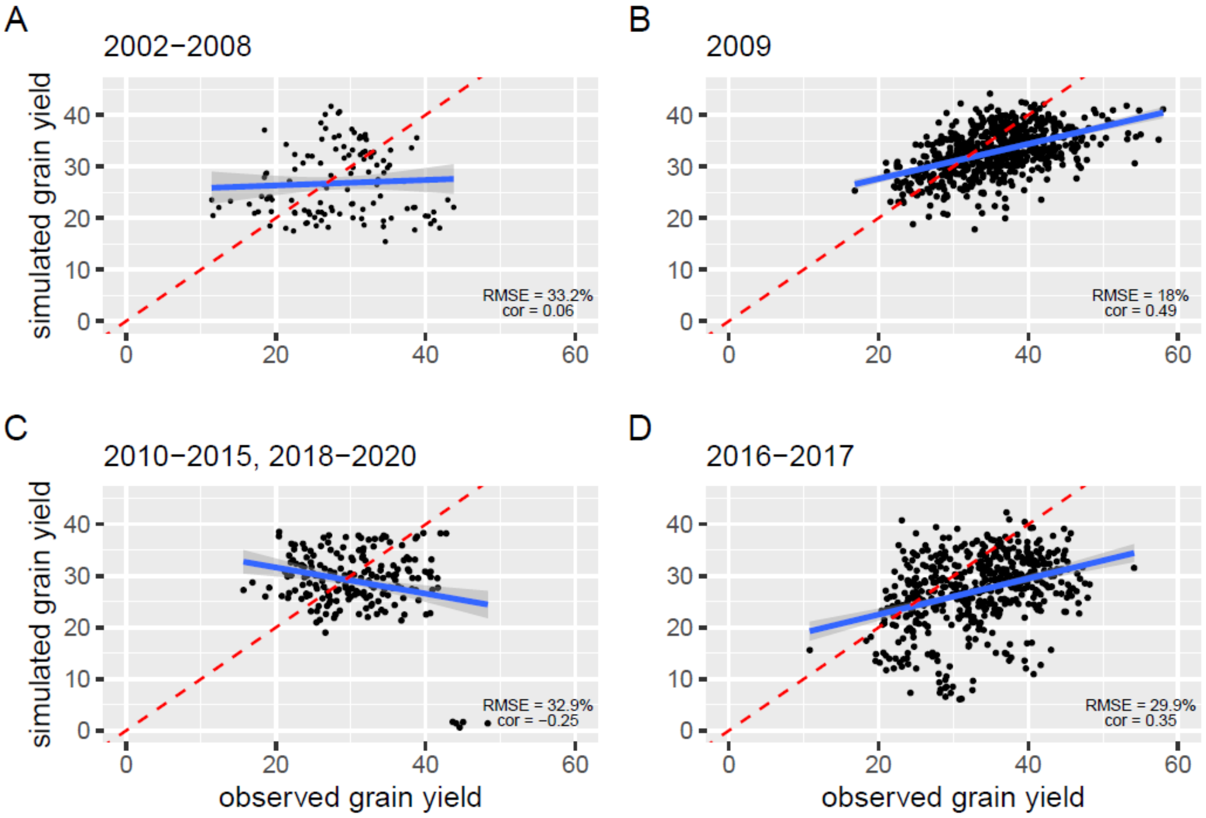
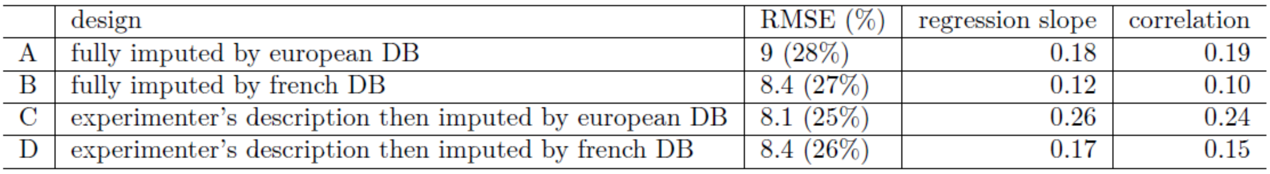


Figure 8. ***Comparison of observed and simulated grain yields depending on the year of the trial:*** *2002-2008 (A), 2010-2015 + 2018-2020 (B), 2009 (C), 2016-2017 (D)*

3.2.5 Selection of the best pipeline

In order to simulate as many trials as possible, we tested different ways to implement the soil data for all trials with enough data to be simulated by the SUNFLO model (740 trials, 4555 observations) (Table 2). The first solution was to implement all the soils using databases: the French one (Table 3A) or the European one (Table 3B) databases. Another solution was to keep the experimenter's description soil data when possible at the trial scale, and at the municipality scale, and then complete the data from the French (Table 3C) or European (Table 3D) databases.

Table 3 – Simulation results for different complete designs



	design	RMSE (%)	regression slope	correlation
A	fully imputed by european DB	9 (28%)	0.18	0.19
B	fully imputed by french DB	8.4 (27%)	0.12	0.10
C	experimenter's description then imputed by european DB	8.1 (25%)	0.26	0.24
D	experimenter's description then imputed by french DB	8.4 (26%)	0.17	0.15

The statistical indicators (Table 3) suggested two conclusions. Simulation using experimenter's description of soil achieved better performances than simulation only imputed by soil databases. The European database gave better results than the French database from the SAFRAN mesh.

3.3 Characterization of the environmental and genotypic effects

The simulation results were split here in two components: an environmental effect (Figure 9B), due to soil, management and climatic conditions and a genotypic effect (Figure 9C).

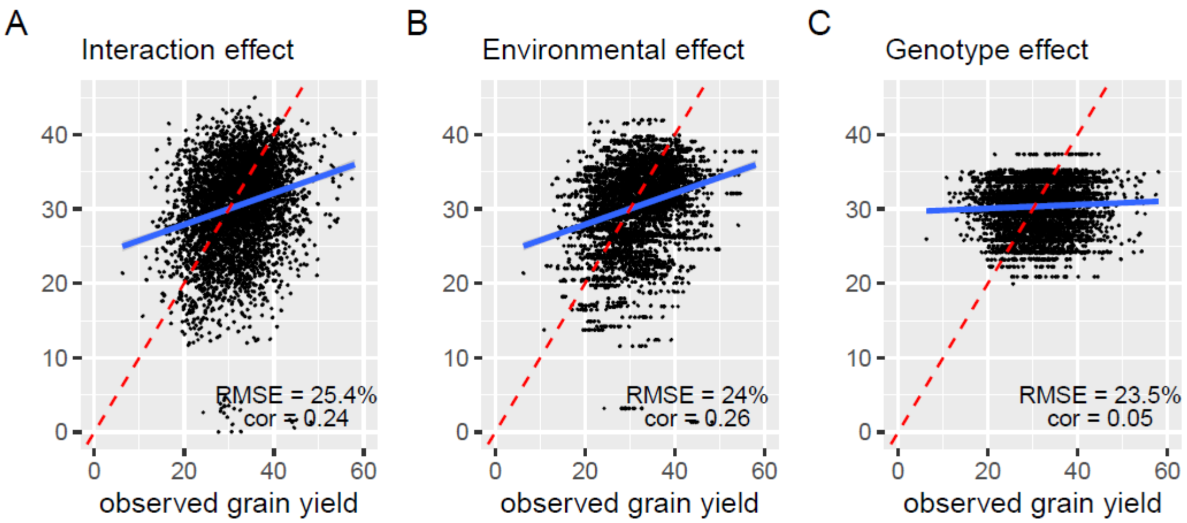


Figure 9. Comparison of observed and simulated grain yield (q/ha) decomposed in effects of the environment (B) and the genotype (C)

The environment effect was measured by assigning to each location the average simulated yield of all varieties at that location. To approximate the genotype effect, we have similarly assigned to each variety its average on all the locations of the network. Firstly, what we observe here is that the environmental effect is much better represented than the genotypic effect. Secondly, the environment effect alone is very close to or even better than the results of the interaction effect. The G by E interaction does not really add any information and even adds some noise compared to the environmental effect alone. This means that we are simulating mainly the effect of the location and it is difficult to access to the genotype effect, and even less to access to the response of the genotype to its environment.

4 Conclusion

Terres Inovia's experimental network provided a huge amount of data and detailed description of each trial. However, these data do not correspond to the requirements of a crop growth model: many crop management and soil input variables are missing, and the quality of the data is variable according to the data collection effort. We have seen that the quality of the data is essential to obtain accurate simulation results.

With this data quality and considering the field of application, we concluded that the results were insufficient to separate the best varieties. Nevertheless we demonstrated that the model could simulate the environment effect alone. We therefore want to use these modeling results to explore the issue of envirotyping. With the objective to create a numerical experiment that accurately describes farming conditions and define groups of environments to provide year-independent context. To this end, we will explore methods to cluster time series of simulated stressors by functional data analysis.

The objective of our study is to prototype a decision support system for recommending variety choice at sowing time, taking into account varietal characteristics as well as the cropping context. This tool, already in development using the R language and the Shiny package, is based on three main areas of advice:

(i) classify varieties in terms of performance: use observed performance results, in which we strongly trust and use simulation to access to the stability of these results over time.

(ii) express the agronomic merit of a variety according to its own disease resistance characteristics and the environmental biotic risk

(iii) envirotyping to describe growing conditions and characterize performance in this context.

5 References

Bertuzzi P., Clastre P., 2022, “Information sur les mailles SAFRAN”, https://doi.org/10.57745/1PDFNL, Recherche Data Gouv, V2, UNF:6:g0bDYiBhAL9atsAInFUW3Q== [fileUNF]

Casadebaig P., 2008. Analyse et modélisation des interactions génotype - environnement – conduite de culture : application au tournesol (Helianthus annuus L.). Thèse INP Toulouse, 195 p.

Casadebaig P., Guilioni L., Lecoeur J., Christophe A, Champolivier L., Debaeke P., 2011. SUNFLO, a model to simulate genotype-specific performance of sunflower crop in contrasting environments. *Agricultural Forest Meteorology* 151, 163-178.

Casadebaig P., Mestries E., Debaeke P., 2016. A model-based approach to assist variety evaluation in sunflower crop. *European Journal of Agronomy* 81, 92-105.

Casadebaig P., Gauffreteau A., Landré A., Langlade N.B., Mestries E., Sarron J., Trépos R., Vincourt P., Debaeke P., 2022. Optimized cultivar deployment improves the efficiency and stability of sunflower crop production at national scale. *Theoretical and Applied Genetics (in press),* https://doi.org/10.1007/s00122-022-04072-5

Chapman S.C., 2008. Use of crop models to understand genotype by environment interactions for drought in real-world and simulated plant breeding trials. *Euphytica* 161: 195-208.

INRA, 2018, “Base de Données Géographique des Sols de France à 1/1 000 000 version 3.2.8.0, 10/09/1998”, https://doi.org/10.15454/BPN57S, Recherche Data Gouv, V1, UNF:6:CZ5MUg5ncyO8Bu+SHAWk9w== [fileUNF]

Jeuffroy M.H, Casadebaig P., Debaeke P., Loyce C., Meynard J.M., 2014. Use of agronomic models to predict cultivar performances in various environments and cropping systems. A review. *Agronomy for Sustainable Development* 34, 121-137

Mangin B., Casadebaig P, Cadic E, Blanchet N., Boniface M.-C, et al. 2017. Genetic control of plasticity of oil yield for combined abiotic stresses using a joint approach of crop modeling and genome-wide association. *Plant, Cell and Environment* 40, 2276-2291.

Wang E., Brown H.E., Rebetzke G.J., Zhao Z., Zheng B., Chapman S.C., 2019. Improving process-based crop models to better capture genotype x environment x management interactions. *Journal of Experimental Botany* 70, 2389-2401