Fostering sustainable legume-based farming systems and agri-feed and food chains in the EU

Deliverable D1.4

Map of achievable legume yields across studied EU-areas

Planned delivery date: M36

Actual submission date: M43

Start date of the project: June 1st, 2017 **Duration:** 48 months

Workpackage: WP1

Workpackage leader: INRA **Deliverable leader:** UNIPI

Partners contributing to the deliverable: INRA, TERIN, WU, PGRO, WR, SSSA, INIAV, AICF, SEGES, LLKC, FH-SWF, CRAN, LAMMC, FiBL

Version: V2

Dissemination Level	
Public	x
Classified, as referred to Commission Decision 2001/844/EC	
Confidential, only for members of the consortium (including the Commission Services)	

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

Summary:

In this report we summarized the major outcomes of the dataset on grain legumes yield built in Task 1.2 of WP1 of the LegValue project. According to the DoA, Task 1.2 might have delivered by month 36 a unique deliverable (D1.4) reporting on yield maps of grain legumes in different pedoclimatic conditions in the EU. Due to difficulties in gathering a significant amount of scientific data on grain legume yields and due to efforts higher than expected spent on validation of the modelling methodology adopted to generate the maps, the leader of task 1.2 proposed and the ExCom agreed upon splitting this deliverable in two parts (namely Part A and Part B).

This dataset, constituting the first part (Part A) of D1.4, was primarily intended for consolidating scientific data necessary to feed the modelling exercise adopted for the production of European maps of grain yield potential for the major pulse species (soybean, field pea, faba bean, chickpea, lentils), that will form the second part (Part B) of the D1.4.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

I. Part – A

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

Table of contents

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

1. Summary

In this report we summarized the major outcomes of the dataset on grain legumes yield built in Task 1.2 of WP1 of the LegValue project. According to the DoA, Task 1.2 might have delivered by month 36 a unique deliverable (D1.4) reporting on yield maps of grain legumes in different pedoclimatic conditions in the EU. Due to difficulties in gathering a significant amount of scientific data on grain legume yields and due to efforts higher than expected spent on validation of the modelling methodology adopted to generate the maps, the leader of task 1.2 proposed and the ExCom agreed upon splitting this deliverable in two parts (namely Part A and Part B).

This dataset, constituting the first part (Part A) of D1.4, was primarily intended for consolidating scientific data necessary to feed the modelling exercise adopted for the production of European maps of grain yield potential for the major pulse species (soybean, field pea, faba bean, chickpea, lentils), that will form the second part (Part B) of the D1.4. Additional objectives were:

- to allow more precise and realistic estimations of protein and starch yield from pulses in current and future scenarios of legumes included in cropping systems in the EU;
- to identify research gaps and technological lock-ins currently hampering a higher share of pulses in cropping systems;
- to identify sites with high productive potential for pulses that are currently unexplored in the EU;
- to set for a step up for science on legume crop management by highlighting ways to improve the existing knowledge on the topic.

The dataset that is made publicly available consists of data extracted from scientific papers published on international journals, that are complemented with data from LegValue Task 1.2 Partners own data, generated both on-station and on-farm but always under controlled conditions, and from the former EU-FP7 Project LEGATO.

Part of the unpublished data were made available by the LegValue partners only for the objectives of the project and consequently are not included in the public version of the dataset.

By outlining the 5244 entries of the dataset, we preliminarily explored the relationships between the grain yield of the five legumes and environmental (soil, climate, latitude) and agronomic (tillage, irrigation, fertilisation, weed control, organic farming practices) aspects. Grain yields varied a lot in function of pedoclimates but possibilities of cross-fertilisation among locations and contrasting conditions were identified in terms of knowledge transfer and establishment of future research.

Among management practices, special emphasis has been given in former studies on crop genotype, an evidence contrasting with poor resources allocated so far for the genetical improvement of these crops. The high presence also of combination between genotype and other management aspects reveals the high interest within the scientific community in developing new plant material adapted not only to a range of environmental conditions but also consistent with management practices.

In this “Part A” of the Deliverable 1.4 we also presented the preliminary results of the modelling activity performed on soybean yield data that were estimated by different algorithms based on former analysis on global dataset linking yield levels with climatic conditions. The preliminary tests performed on soybean yield estimations were very positive. The selected algorithm (Random Forest) allowed to predict in a reliable way soybean yields under historical climatic conditions also in the EU territory. According to preliminary analysis, although not considering the positive effect of irrigation on potential

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement N°727672

yields, the internal demand of soybean grain seems to be likely met by EU production even if the legume would be present just 1 year over 6 of crop rotations.

This modelling approach will be then fed in the next 6 months with the yield data of all the 5 major pulse species considered and included in the dataset (Annex A to this report) in order to refine the validation of the model and to be able to produce more reliable yield estimations and more accurate yield maps (Part B of this Deliverable 1.4).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

2. Introduction

Within the LegValue project, we aim at enhancing the presence of legume crops in cropping systems in the EU. Besides forage legumes, that are well known to deliver a number of different ecosystem services (e.g. higher availability of nitrogen in the soil through N_2 biological fixation, reduction of soil compaction, soil fertility restoration, weed suppression, stimulation of soil biological activity and biodiversity), also pulses (i.e. grain legumes) have potential to boost the agroecological transition towards more diversified cropping systems and viable value chains.

Despite these apparent benefits and subsidies that are dedicated in EU Rural Development Plans (RDPs) and in the Common Agricultural Policy (CAP), the area of farmland under legume production in the EU has been showed to steadily decline (except for soybean) in the last decades (Schreuder and de Visser, 2014). Pelzer et al. (2017) have recently estimated that grain legumes occupy only 1.8% of arable land in the EU. There are a number of factors causing this low percentage of land sharing of pulses, many of them being mostly related to socio-economic issues (Magrini et al., 2016), but also agronomic reasons are of paramount importance. Farmers do consider grain legumes so poorly attractive because of the uncertainty and instability of their yields, being normally affected by pedoclimates and biotic stress factors much more than other major arable crops like cereals.

So far, the research efforts paid on studying the relationships between environmental and management factors potentially affecting pulse yields have been not so extensive as for other groups of arable crops. That is why one of the objective of LegValue WP1 is to try to shed light on these aspects and to determine the actual yield potential of the most common pulse species grown in the EU (soybean, *Glycine max* (L.) Merr. ; field pea, *Pisum sativum* L. ; faba bean, *Vicia faba* L. ; chickpea, *Cicer arietinum* L. ; lentils, *Lens culinaria* Medik.).

In Task 1.2 of WP1, we intend to produce EU yield potential maps for the five major pulse species according to current soil and climate conditions with the following objectives:

- to allow more precise and realistic estimations of protein and starch yield from pulses in current and future scenarios of legume presence in cropping systems;
- to identify research gaps and technological lock-ins currently hampering a higher share of pulses in cropping systems;
- to identify sites with high productive potential for pulses that are currently unexplored in the EU;
- to set for a step up for science on legume crop management by highlighting ways to improve the existing knowledge on the topic.

To generate the maps, in Task 1.2 it was agreed on identifying a methodological approach based on simple mechanistic modelling able to link pedoclimatic conditions with grain legume performances, without the need to use highly detailed and specific information on crop physiology and crop phenology, but at the same time able to produce reliable and solid outcomes. Among the available models in the literature, it was decided to use the Ecological Niche Model approach, that revealed to be accurate enough in predicting yield levels of soybean worlwide according to climatic conditions.

In order to feed the model with science-based evidences on legume yields produced and collected in the EU for a finer model validation, we first needed to gather most of the existing knowledge and to structure it in a consolidated form, easy to use and to explore it also external to and after LegValue.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

Deviation from the DoA

This first step (i.e. collection of scientific data on legume yields produced in the EU) was very time consuming and required extra efforts to integrate scientific paper results (that were very few compared to our expectations) with other sources of scientific information (e.g. former EU project results, unpublished experimental results owned by LegValue partners and data generated in LegValue On-Farm Networks -OFN-). Furthermore, the methodological approach for yield map generation (i.e. the Random Forest Ecological Niche Model) needed to be tested and validated on a test crop (i.e. soybean) with yield data produced within the EU before being extended to the other major pulses. This activity, led by INRA, required additional efforts respect to those foreseen in the DoA.

For all this reason, during the last annual meeting of the LegValue project, held virtually from April 28th to April 30th, 2020, **the task 1.2 leader proposed to the ExCom to split the Deliverable 1.4, expected by M36 according to the DoA, in two parts**. In this first part, named **"Part A",** delivered by **M37**, we included the dataset generated in Task 1.2 with explanation of major outcomes and the preliminary results of application of the modelling approach on soybean yields, estimated based on historical climatic data for the EU, based on the results obtained on global observed yield dataset. This preliminary analysis, presented in section 3.5 and 4.3, supports the strength and reliability of the adopted modelling approach and provides an anticipation of the yield maps that will be delivered in a second part **("Part B"**), expected by **M43** (December 2020).

In Part B we will present the final version of the yield maps generated by the Ecological Niche Model for all the 5 pulse species based on calibration and validation of the algorithm with data on yields and pedoclimates included in the dataset annexed to Part A. In Part B, an in-deep analysis of relationships between pulse yields and pedoclimates, as well as a science-based estimation of the potential of grain legume yields in the EU will be also provided.

The ExCom of LegValue agreed upon this deviation from the DoA, as it was considered that:

- the dataset produced with "Part A" of the deliverable, that was not expected according to the DoA, might represent a product with added value for the other tasks of the project (in particular for Tasks 1.3 and 1.4 that could benefit from more structured information on the link between pedoclimatic and agronomic variables and legume yields) and for the scientific community;
- the delay in the delivery of the yield maps should not affect anyhow the regular development of the Work Package 1 and of the other tasks linked to Task 1.2 (in particular, Task 1.3 and Task 1.4).

3. Materials and methods

3.1. Literature review

The collection of the data on legume yields produced under controlled conditions in the EU started by extracting the data on the 5 selected legume species (soybean, field pea, faba bean -including broad bean, pigeon bean and horse bean-, chickpea and lentils) from the global dataset published by Cernay et al. (2017). The paper includes 8,386 articles published globally from 1967 to 2016 and respecting the following 6 eligibility criteria on the dataset ISI-Web of Science:

1. 1/+ legume grown as sole crop in title/abstract;

2. >1 legume grown per each site (title/abstract);

3. at least one experiment for 1/several years from seeding to harvest;

4. peer reviewed journals;

5. written in English;

6. full text available

The dataset produced by Cernay et al. (2017) followed also other eligibility criteria:

- Studies conducted on grain legumes;
- Studies reporting grain/aerial biomass yields;
- Data coming from scientific experiments;
- Studies not including forage legumes, cover crops and intercrops.

By restricting the global dataset to the EU countries, we extracted 223 papers potentially eligible for the dataset. All of them were read in full to assess for their eligibility, according to our additional criteria (i.e., grain yield data reported as tables for each legume species, precise information on field site pedoclimatic conditions), and at the end only 25 papers were first considered eligible.

To expand the dataset we then proceeded as follows:

1. we updated the dataset to 2019 by launching the same search string used by Cernay et al. (2017) on Scopus and Web of Science databases;
2. we expanded the dataset to Turkey;
3. we modified the search string in order to:
    a. include also studies involving even only one legume crop at a given site;
    b. include also studies reporting on legumes grown in intercropping (in case also legume yield in pure stands was available);
    c. specify also the name of the 5 legume species considered;

    The final search string applied was as follows:

    "crop* AND (soybean OR soyabean OR pea OR faba bean OR chickpea OR lentil) AND (yield* OR 'dry matter' OR biomass) AND (compar* OR assessment OR product* OR performance*) AND (trial* OR factorial OR experiment* OR treatment* OR condition*)";

4. we screened all the eligible papers and we also kept in papers reporting not only data in tables but also in charts, by using the WebPlotDigitizer-4.2.0 app (https://automeris.io/WebPlotDigitizer) to extract data from graphical forms;
5. we also gathered additional papers included in the literature review on ecosystem services related to legume crops performed in Task 1.1 and that were not identified according to our inclusion criteria;
6. we explored references cited in eligible papers identified as above.

3.2. Data from the EU-FP7 Legato Project

To expand the dataset, in a second moment it was agreed on including also open data coming from former EU project. The EU-FP7 LEGATO Project "LEGumes for the Agriculture of TOmorrow" (2014-2017) has delivered the results of the 2-yr experimental activities performed on field pea, faba bean, chickpea, lupins and chickling peas (*Lathyrus* spp.). The dataset was downloaded from

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

https://intranet.iamz.ciheam.org/forms/Legato/WP6/index.php and eligible data on pea, faba bean and chickpea were extracted and added to the dataset.

3.3. Data from unpublished experiments carried out by LegValue partners

In order to expand the geographical coverage of the dataset, we also asked the Task 1.1 partners to deliver their own experimental data coming from LegValue Task 1.1 trials on legume yields (see Milestone MS3), from OFN trials as well as from other field experiments performed externally to LegValue and not already published.

The summary data reported in the Results chapter of this document referred to the whole dataset, including also data owned by LegValue partners that will remain not publicly accessible but that will enrich the production of yield maps that will be included in the Part B of this Deliverable 1.4.

3.4. Structure of the dataset

The dataset was structured as a worksheet with variables as columns and entries as rows.

Each combination of year x site x crop species x experimental treatment level, averaged over spatial or temporal replications, was reported as a single entry (i.e. a single row).

To facilitate browsing the dataset, we added also filters on each column.

An "Instructions" worksheet was also added to facilitate data entry operations as well as to inform the readers about the proper interpretation of the data.

The dataset includes 63 columns (i.e. variables), as described in details in Table 1.

Table 1 – Structure of the dataset and meaning of each column (variable)

COLUMN TITLE	**CONTENT**
ID	Entry ID
Source	Experiment/Paper/LEGATO Project
Publicly available	Y/N (Y if this data could be part of an open-access publication of the dataset on a Data Journal; N if for internal use in LegValue)
Experiment_ID	Experiment acronym_Surname of responsible
Site_Country	Name of the country in full
Site_Region	NUTS 3
Site_Name	Name of the site in full
Latitude	Decimal degrees of Latitude (XX.xx).
Latitude Cardinal	N/S
Longitude	Decimal degrees of Longitude (XX.xx).
Longitude Cardinal	W/E
Site_Soil_Classification_Name	Soil classification type (USDA)
Site_Soil_Texture_Name	Soil texture class (e.g. loam, sandy, sandy loam)
Site_Precipitation_mm	Total rainfall in the period considered. "NA" if not available
Site_Precipitation_Period	annual/growing season
Site_Precipitation_Period_Month	Initial Final month of the period for which precipitations are reported (e.g. Jan Dec)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

COLUMN TITLE	CONTENT
Site_Precipitation_Period_Year	Years of registration of the precipitations (e.g. 1993).
Site_Temperature_Celsius	Average temperature in the considered period. "NA" if not available
Site_Temperature_Period	annual/growing season
Site_Temperature_Period_Month	Initial Final month of the period for which temperature is reported (e.g. Jan Dec)
Site_Temperature_Period_Year	Years of registration of the temperature (e.g. 1993).
management evaluated	e.g. tillage, irrigation, variety
Crop_Sequence_Treatment_Name	Report the name of the treatment or (in case of factorial combination) the name of the combination
Scientific name	Latin name (without author initials) of the legume crop species (e.g. Glycine max)
Previous crop	Latin name (without author initials) of the crop species grown before the legume (e.g. Triticum aestivum)
Crop	Common name of the legume crop species (e.g. Soybean)
Cv	Name of the legume crop variety
Group of precocity	Only for soybean. Report here the precocity group (000 to 10)
GM	Genetically modified variety? Y/N
Sowing date	mm/dd/yyyy. "NA" if not available
Harvest date	mm/dd/yyyy. "NA" if not available
Tillage_no tillage	"tillage" if a tillage operation is performed before legume sowing, or "no-tillage" if sod-seeding legume
Days_crop_cycle	Length of crop cycle in the experimental year (nr. of days from sowing to harvest). "NA" if not available
Plant density (plant/m2)	nr of legume plants per m2 (alternative to sowing density). "NA" if not available
Sowing Density (seeds/m2)	nr of legume seeds per m2 (alternative to plant density). "NA" if not available
row spacing (m)	inter-row space in meters. "NA" if not available
N rate (kg N ha-1)	Total amount of N (kg ha^{-1}) supplied to the crop
N fertiliser type(s)	Name(s) of each N fertiliser applied to the crop with its level of N application rate (kg N ha^{-1}), separated by comma and in chhronological order of application (e.g. Poultry manure -30-, Urea -30-)
Nr. of applications (N)	Nr. of applications of N fertilisers
% N from organic fertiliser(s)	% of total N supplied to the crop coming from organic fertilisers or amendments
P rate (kg P ha-1)	Total amount of P (kg ha^{-1}) supplied to the crop
P fertiliser type(s)	Name(s) of each P fertiliser applied to the crop with its level of P application rate (kg P ha^{-1}), separated by comma and in chhronological order of application (e.g. Diammonium phosphate -30-, Superphosphate -30-)
Nr. of applications (P)	Nr. of applications of P fertilisers
% P from organic fertiliser(s)	% of total P supplied to the crop coming from organic fertilisers or amendments
K rate (kg K ha-1)	Total amount of K (kg ha^{-1}) supplied to the crop
K fertiliser type(s)	Name(s) of each K fertiliser applied to the crop with its level of K application rate (kg K ha^{-1}), separated by comma and in chhronological order of application (e.g. Potassium sulphate -30-, Sugarbeet liquid pulp -15-)
Nr. of applications (K)	Nr. of applications of K fertilisers
% K from organic fertiliser(s)	% of total K supplied to the crop coming from organic fertilisers or amendments
irrigation	Y/N_partial/full (Y if irrigation was applied or N if not; PARTIAL if irrigation did not cover the full water need of the crop or FULL if it did)
Mean irrigation quantity (mm)	Mean amount of irrigation water (mm) applied (exact amount or mean value if a range is reported). "NA" if not available
Herbicide application	Were chemical herbicides applied or not (Y/N)
Mechanical weed control	Was mechanical weeding applied or not (Y/N)
Crop protection	Were crop protection products, including natural or biocontrol agents, applied to the crop (Y/N)
number of replicates	Number of replicates concurring to the mean yield value reported in a single site x year combination (e.g. number of blocks or spatial replicates)
number of sites	Number of different sites considered as spatial replicates for computing the mean yield value reported, if mean yield values for each site are not available

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement N°727672

COLUMN TITLE	CONTENT
number of years	Number of years concurring to the mean yield value reported, if single year mean yield values are not available
Moisture content harvest (%)	Moisture percentage of the marketable yield (e.g. "13" for 13%) as reported in the source material
Grain Yield (t d.m./ha)	Yield of the grain of the legume crop in t d.m. ha^{-1} (the humidity reported in the previous column, when available, is removed from the grain yield reported)
SE crop yield	Value of the standard error of the mean of the yield, if available. If not available, "NA"
SD crop yield	Value of the standard deviation of the mean of the yield, if available. If not available, "NA"
CV crop yield	Value of the coefficient of variation of the mean of the yield, if available. If not available, "NA"
Var. crop yield	Value of the variance of the mean of the yield, if available. If not available, "NA"

3.5. Data analysis

To summarise the major outcomes of the full version of the dataset, graphical exploration of the data was performed by using the package ggplot2 (Wickham, 2009) of the statistical software R, version 136 3.3.1 (R Core Team, 2013).

A more detailed study on the relationships between legume yields and pedoclimatic conditions will be included in the Part B of the Deliverable. However, we briefly present below the method that we plan to use to generate the maps of achievable yield for each legume crop based on the relationships between legume yields and pedoclimatic conditions. The method has been developed for soybean in Europe. Building on two recently published global datasets including historical soybean yield and retrospective meteorological forcing (Iizumi et al. 2014a,b), we developed data-driven relationships between climate and soybean yield to estimate soybean suitable areas over Europe. Several machine learning algorithms were trained and tested at the global scale (Random Forest, Artificial Neural Networks, Generalized Additive Model, and Multiple Linear Regression) to predict soybean yield as a function of monthly climate inputs (solar radiation, minimum and maximum temperature, rainfall, and vapour pressure) calculated over the growing season (April to October). A large share of the training data was taken from major soybean-producing countries (Argentina, Brazil, Canada, China, India, Italy and the United States), and zero-yield data points were randomly sampled in climate zones known to be unsuitable for soybean production (e.g. deserts and arctic areas) and added to the dataset so that they represented about 20% of the final dataset. The most accurate algorithm was selected after running a cross-validation procedure assessing model transferability in time and space. The selected algorithm (Random Forest) was then run for the entire Europe to assess potential distribution of soybean suitable area in rainfed conditions under current and future climate. Projections of soybean suitability in Europe were performed for historical climate (1981-2010). The projections assume a growing season from April to October and no irrigation, although soybean is often irrigated in Europe. The no irrigation assumption prevents from making any hypothesis about available water for irrigation, which is a complex issue especially under climate change. We therefore acknowledge that the yield projections are probably a bit conservative from that point of view.

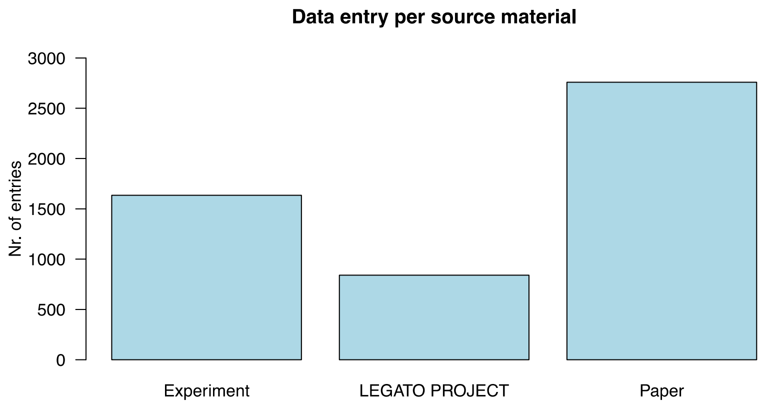
4. Results

4.1. Overview of the dataset

The data collected resulted in 5234 single entries related to the crop yield of the 5 major pulse species grown in the EU.

The distribution of the entries per type of source material (published papers, partners experimental data, LEGATO project) is depicted in Figure 1.

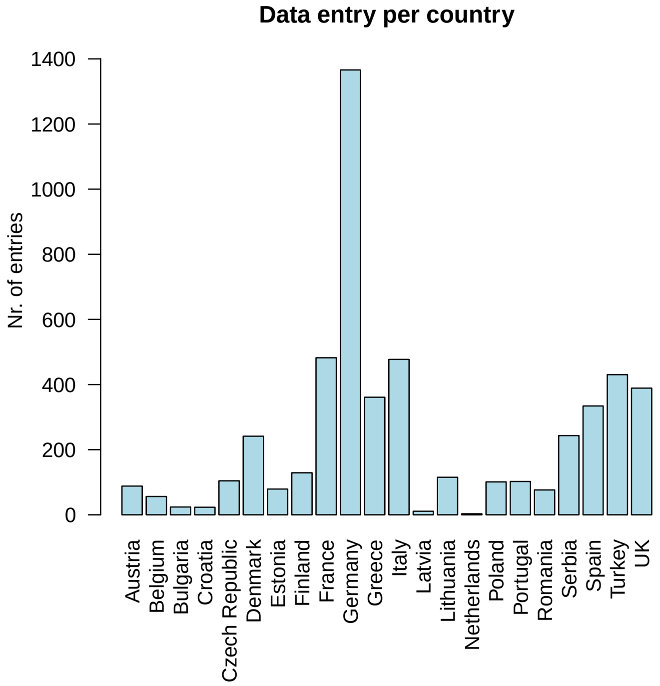
Figure 1 - Number of entries grouped per source material



Most of the entries (2759) comes from published scientific papers, whilst unpublished experimental data include 1635 entries and the results from the LEGATO project cover 840 entries.

As shown in Figure 2, Germany covers 26% of the entries. Also France (482), Italy (477) and Turkey (430) were well represented in the material. A total of 22 countries are represented in the dataset, covering a full range of different pedoclimates from Southern countries (e.g. Spain and Portugal) to Northern sites (e.g. Finland, Estonia, Lithuania) and from Eastern to Western countries.

Figure 2 - Number of entries grouped per country



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

The five pulse crops are not equally represented in the dataset (Figure 3). Pea (1553) and soybean (1511) are the most represented crops in the dataset.

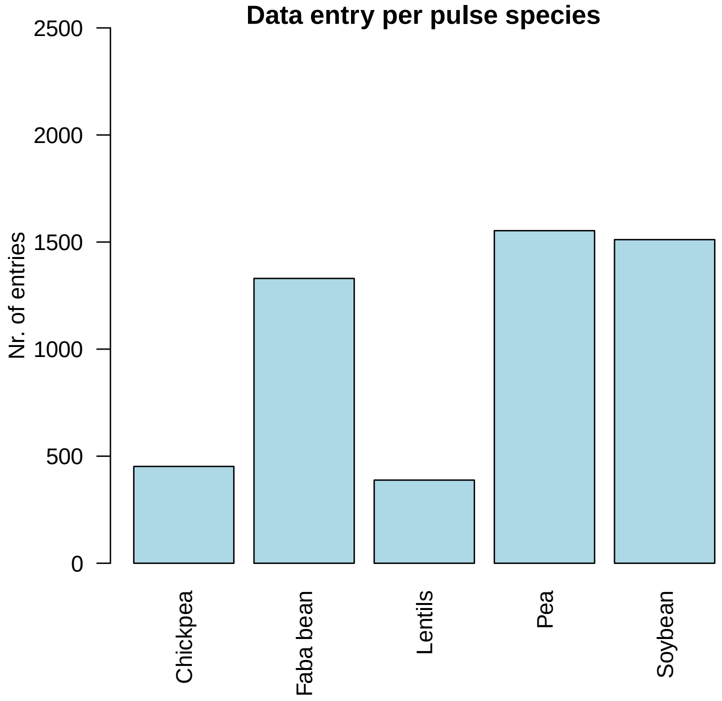
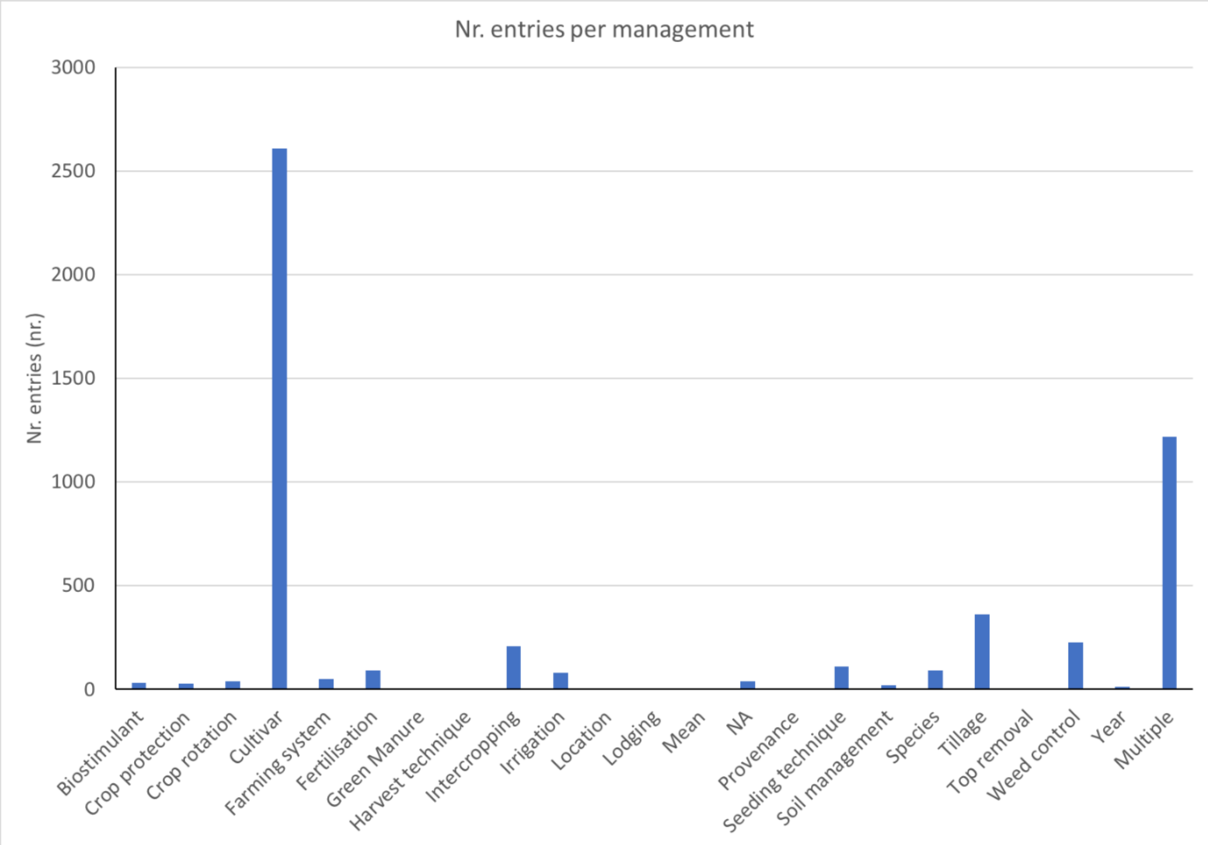


Figure 3 - Number of entries grouped per pulse species

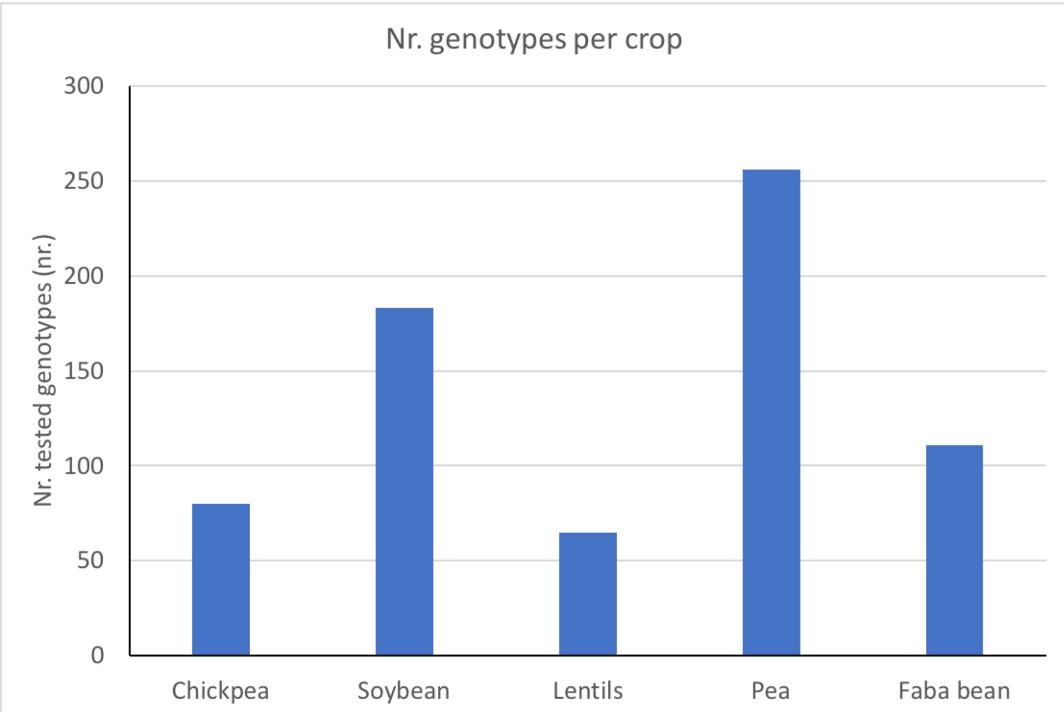
Genotype is the management aspect most represented, covering around 50% of all the entries of the dataset. Among the other agronomic management issues, interactions among several aspects, tillage, weed control and intercropping are the other more common elements (Figure 4).

Figure 4 - Number of entries grouped per management



The number of tested varieties grouped per crop is depicted in Figure 5. The highest number of different genotypes tested in the source material is observed for field pea (256), followed by soybean (183) and faba bean (111). For chickpea and lentils, only 80 and 65 genotypes are represented.

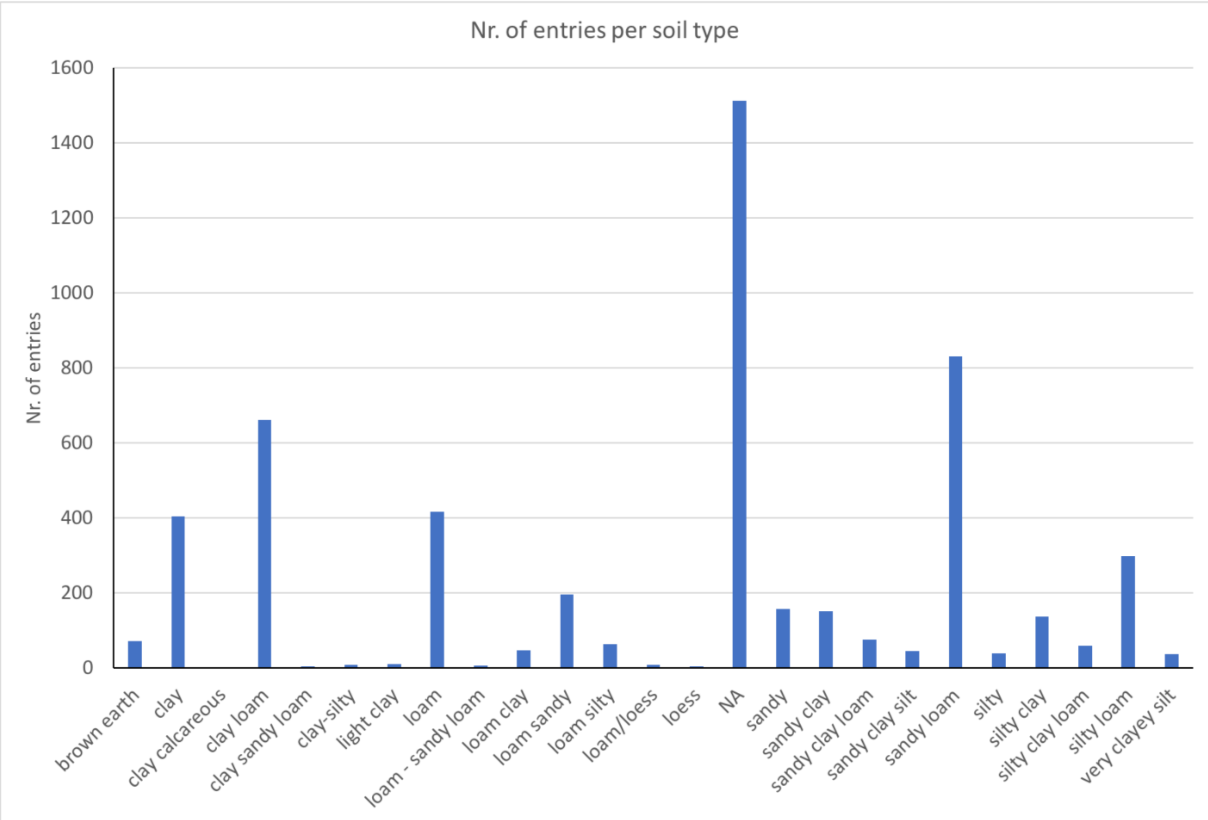
Figure 5 - Number of varieties grouped per legume species



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

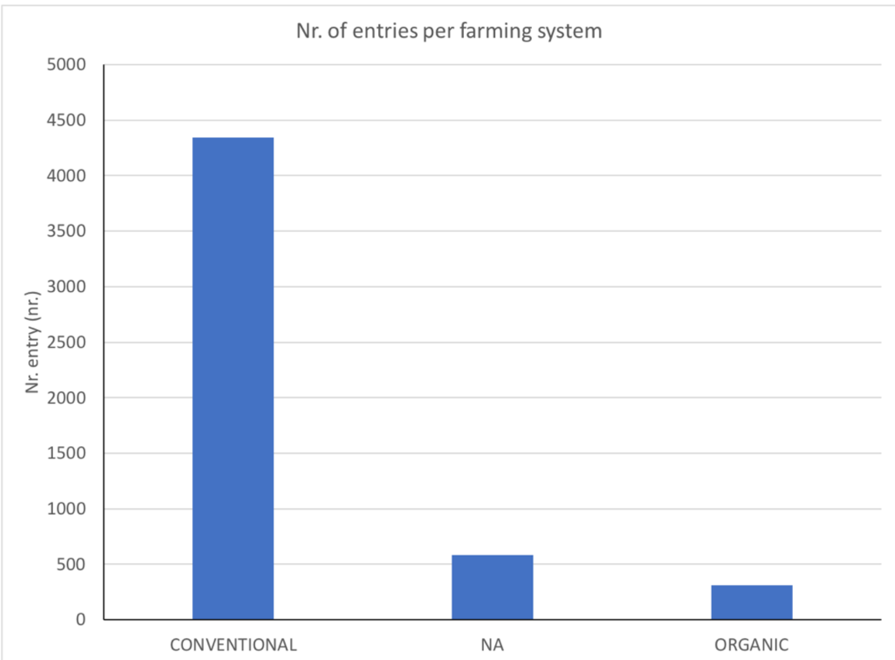
The dataset covers a wide range of different soil types and texture, from very sandy to very clay and calcareous soils, with highest presence of intermediate texture (Figure 6).

Figure 6 - Number of entries grouped per soil texture



For what concerns the farming system, most of the entries refers to studies or treatments managed according to non-organic management, whilst organic farming is represented only by 6% of the entries (Figure 7).

Figure 7 -Number of entries grouped per farming system



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

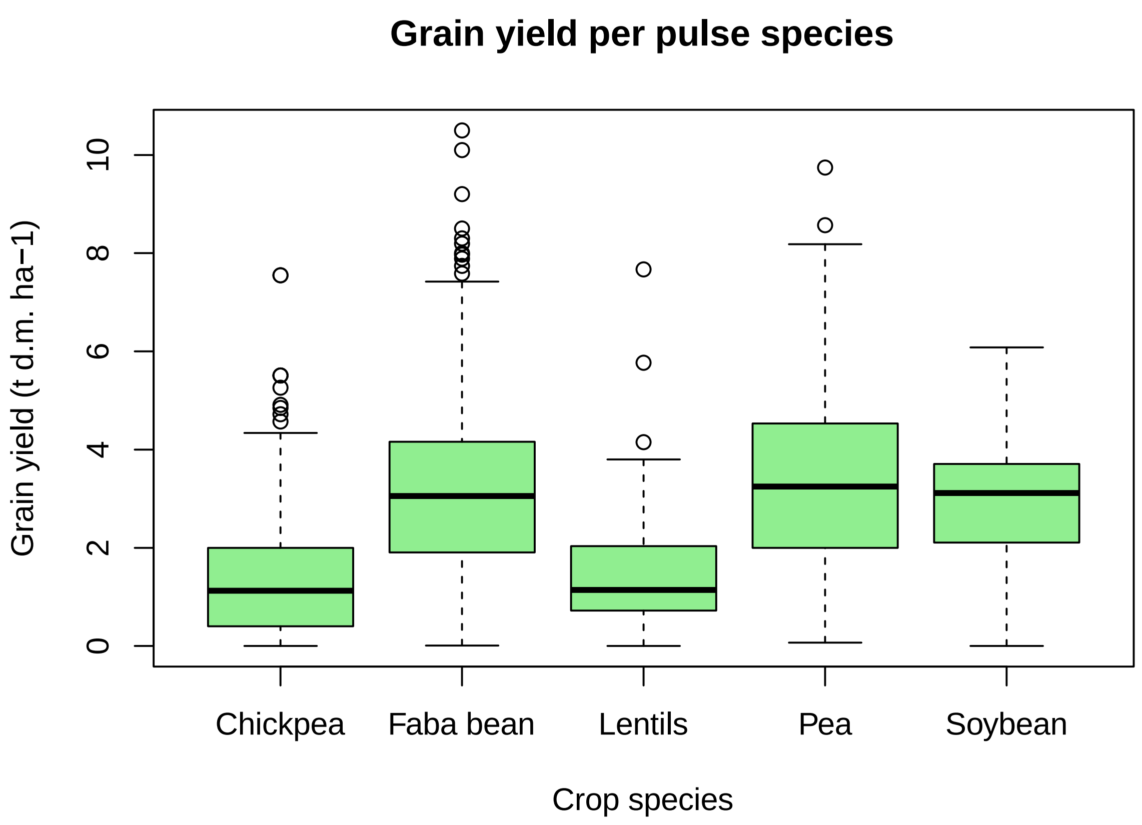
4.2. Effects of factors on crop grain yield

The grain yield in the dataset is normally expressed as t ha^{-1} of dry matter. Anyway, for many papers and experimental data the moisture level of the grain was not reported or not measured, affecting somehow the comparability of the results. In many cases, the mean values of grain yield were not accompanied by measures of variability (standard error, standard deviation, coefficient of variability, variance), thus hampering the possibility to fully exploit the dataset for meta-analyses.

Papers and experimental data not clearly stating the plant produce considered (e.g. grain or total aboveground biomass) were also removed from the dataset.

Concerning the trend of grain yield as affected by crop species, the dataset confirms that the three major pulses (i.e. soybean, field pea and faba bean) grown in Europe are the more productive ones, whilst a clear difference was observed for chickpea and lentils (Figure 8).

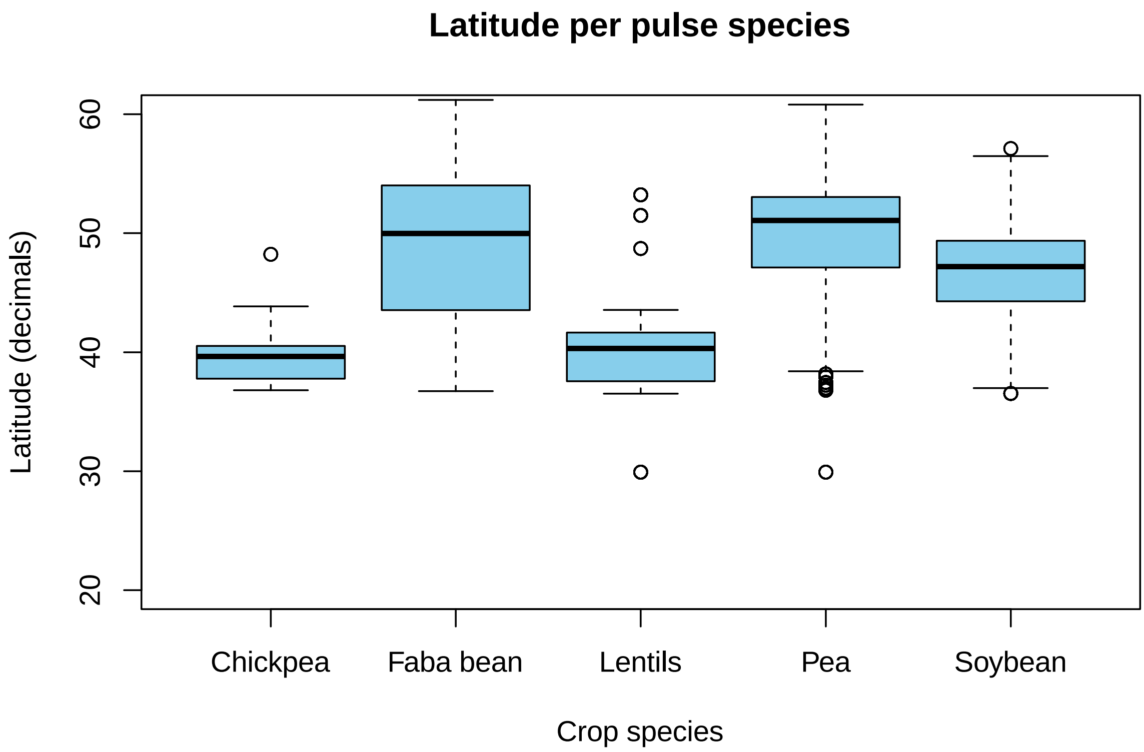
Figure 8 - Boxplot of grain yield (t d.m. ha^{-1}) as affected by legume species



As shown in Figure 9, the trend of grain yield reflects also the geographical distribution of the five crops, with chickpea and lentils mostly studied at lower latitudes.

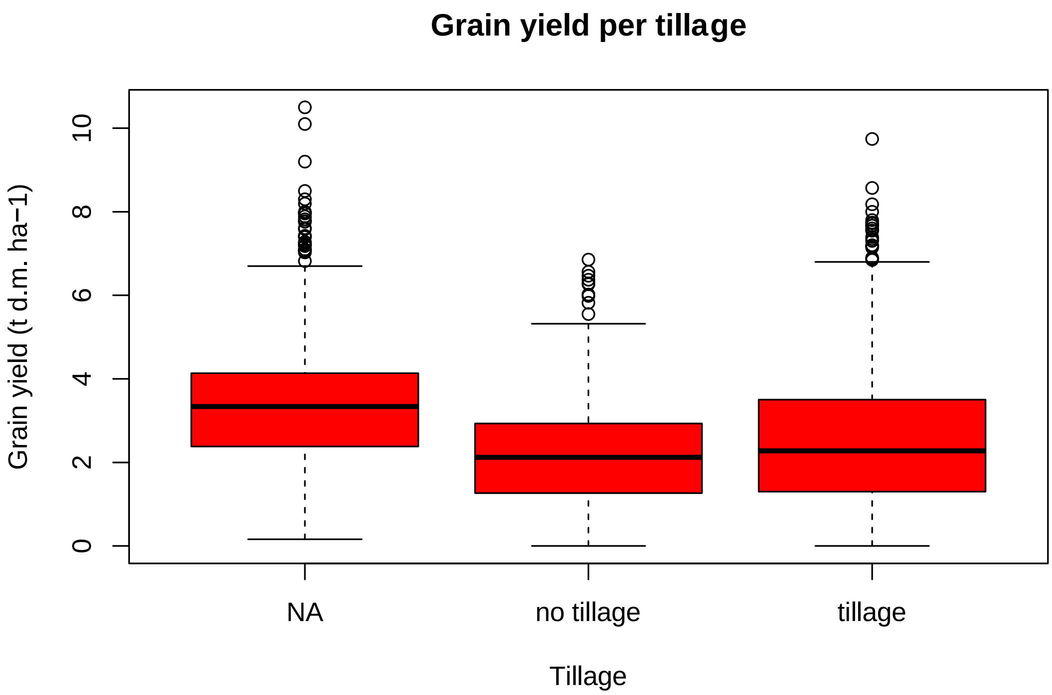
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

Figure 9 - Boxplot of latitude decimals as affected by legume species



Among management practices, overall tillage systems (Figure 10), herbicide application (Figure 11), irrigation (Figure 12) and organic farming (Figure 13) did not clearly affected the levels of variability in legume yields.

Figure 10 - Boxplot of grain yield (t d.m. ha-1) as affected by tillage systems



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

Figure 11 - Boxplot of grain yield (t d.m. ha-1) as affected by herbicide application

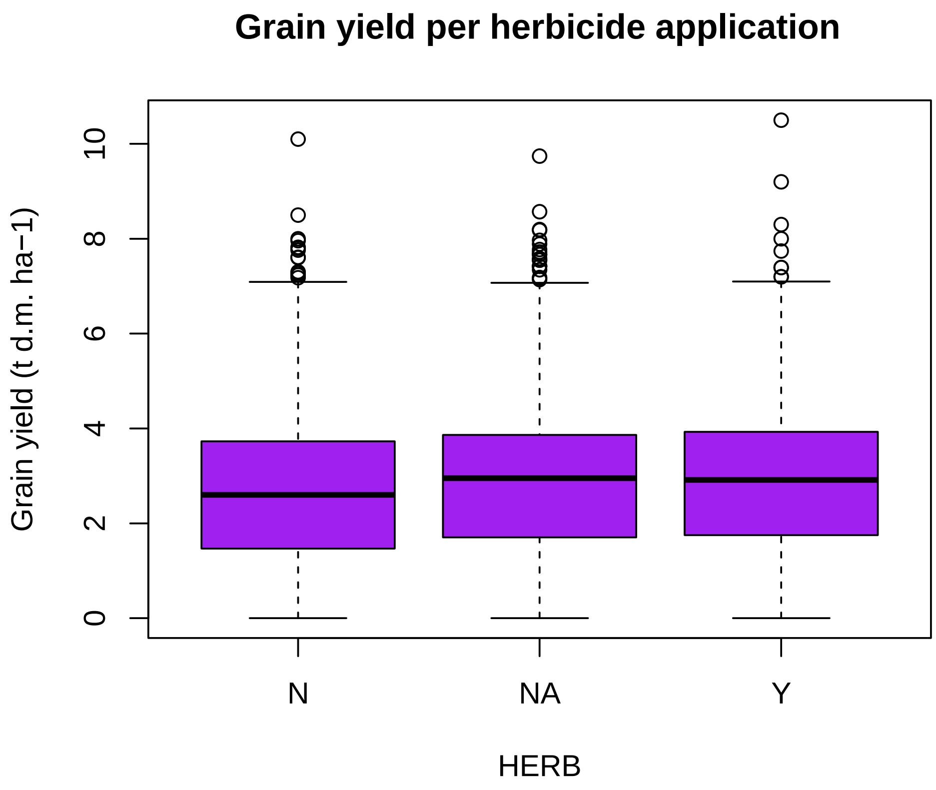
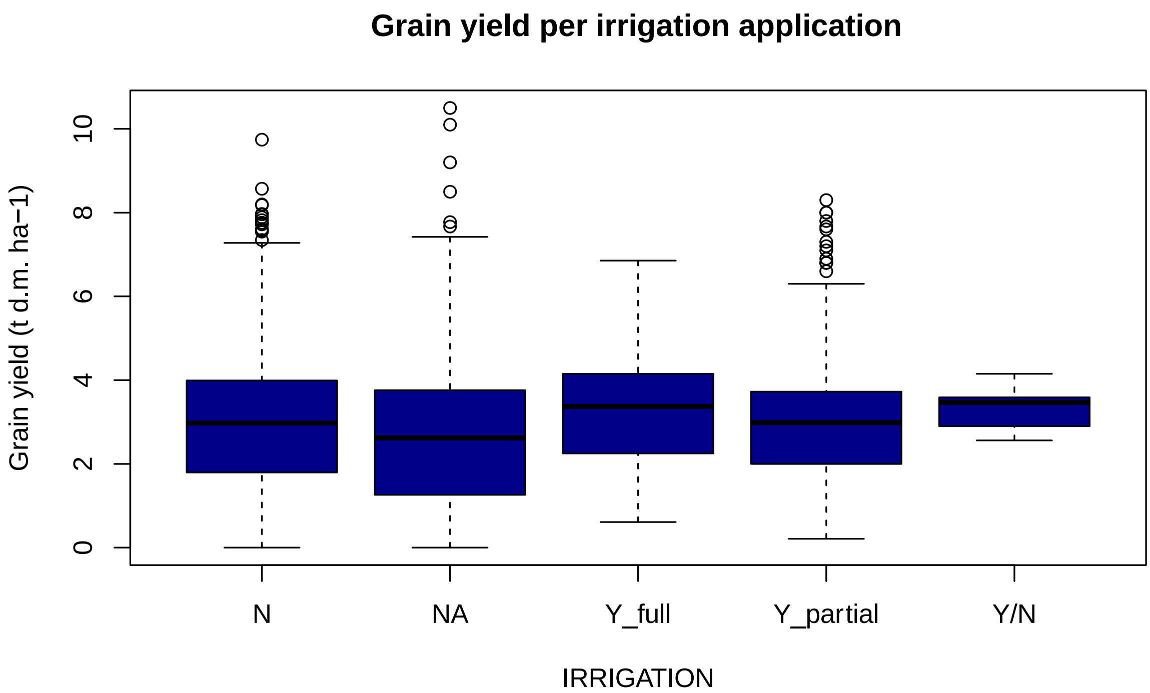
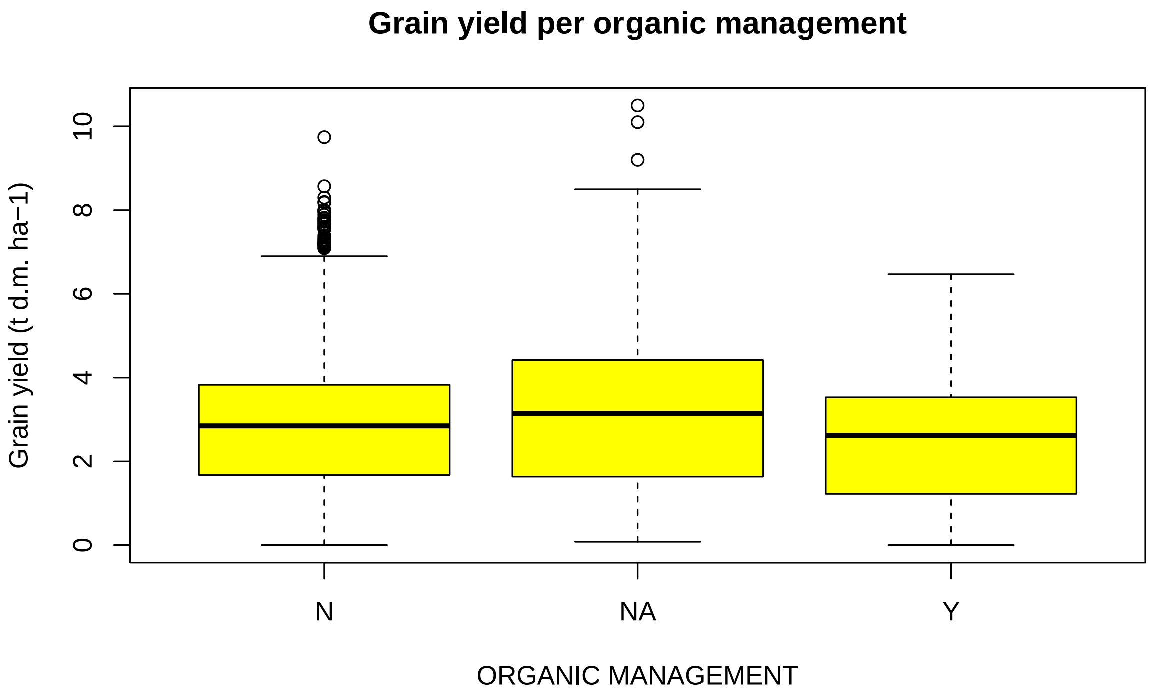


Figure 12 - Boxplot of grain yield (t d.m. ha-1) as affected by irrigation (covering part or full water needs of the crops)



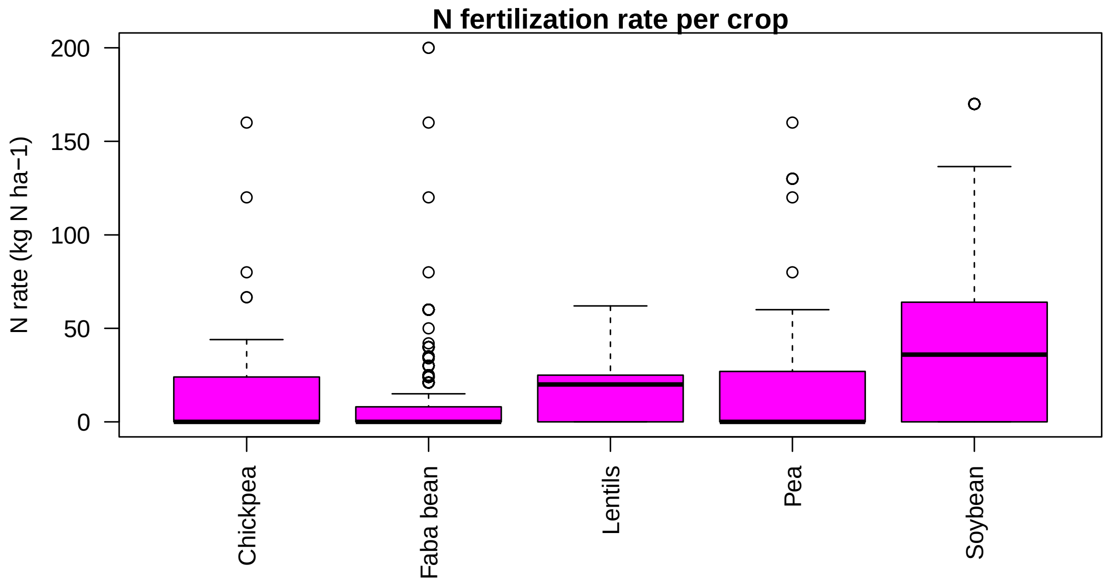
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

Figure 13 - Boxplot of grain yield (t d.m. ha-1) as affected by organic farming management



Concerning the fertilisation, nitrogen is normally applied to soybean at higher levels respect to the other legumes (Figure 14).

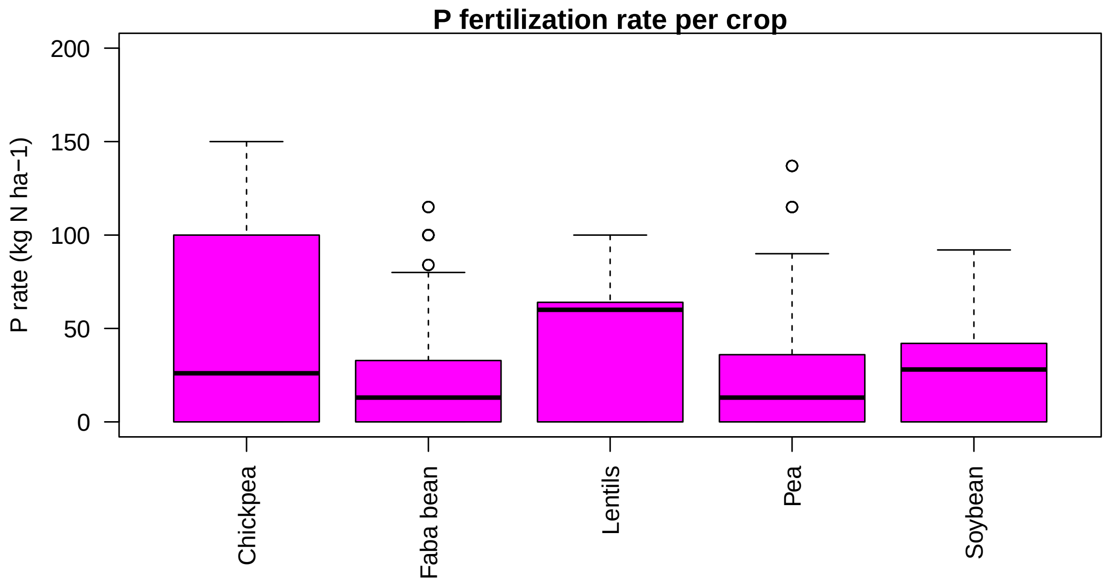
Figure 14 - Boxplot of N fertilisation rate (kg N ha^{-1}) as affected by legume species



For phosphorus fertilisation, the data included in the dataset reveals an opposite trend, with higher levels often tested for less productive pulse species (chickpea and lentils).

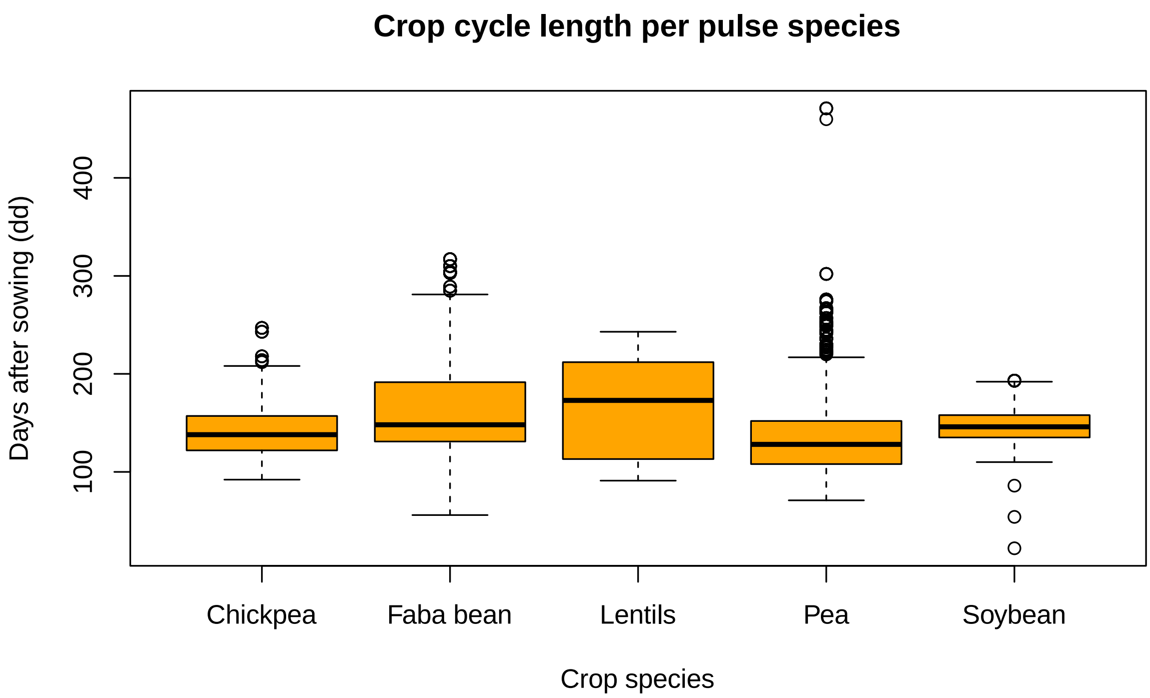
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

Figure 15 - Boxplot of P fertilisation rate (kg P ha^{-1}) as affected by legume species



Concerning the length of the crop growing cycle, the data included in the dataset reveal narrower ranges for soybean and chickpea than for faba bean, lentils and pea (Figure 16).

Figure 16 - Boxplot of the length of growing cycle (expressed in terms of days from sowing to harvest)



As shown in Figure 17, the data included in the dataset do not cover evenly all the crops in terms of geographical distribution for chickpea and lentils, whereas for soybean, pea and faba bean a wider distribution can be observed. For all the crops we observed a high variability of the grain yield, likely related to interannual or spatial variation.

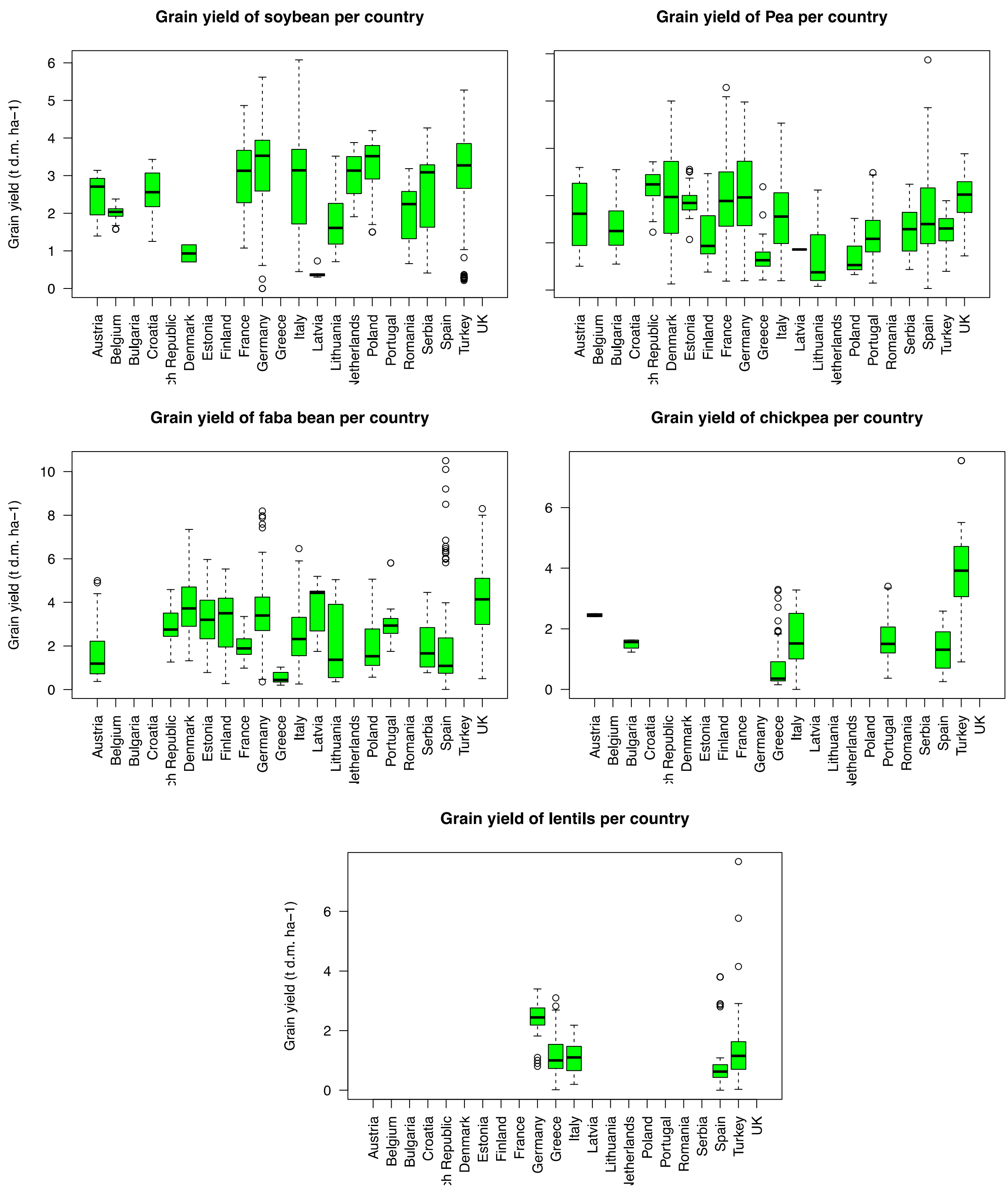
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

For what concerns the soil type, the data in the dataset show a trend towards higher yield value for soybean in loam soils, and poor adaptation of all the legumes to silty soils (Figure 18).

Finally, for soil tillage we observed positive (soybean and lentils), neutral (faba bean and chickpea) and negative (field pea) effects of no tillage on the grain yield of the 5 legumes in terms of median and range of variability (Figure 19).

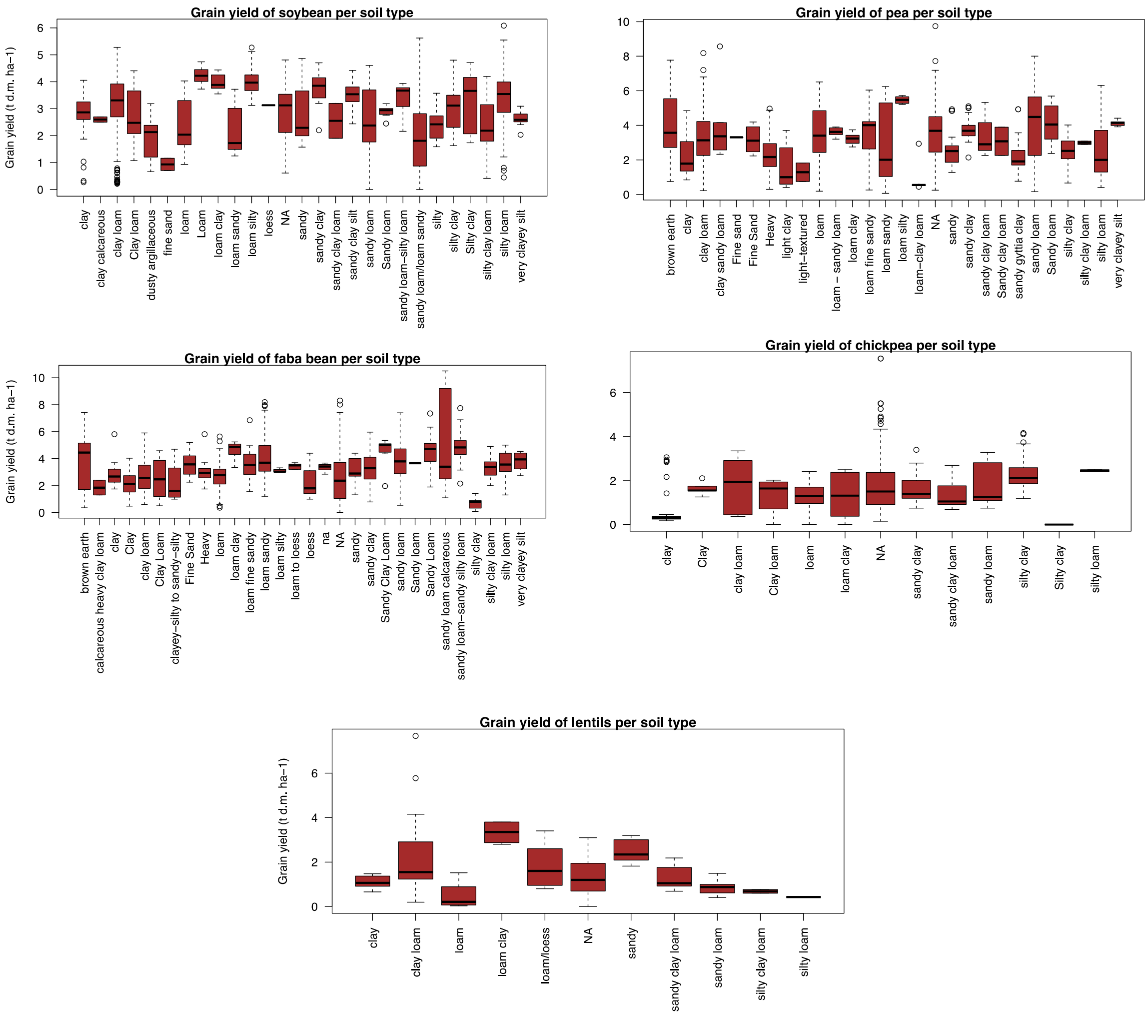
For the other management aspects tested for this report (i.e. irrigation, herbicide application, fertilisation), clear outcomes were not identified.

Figure 17 - Boxplot of grain yield (t d.m. ha-1) of the 5 legume species as affected by organic farming management



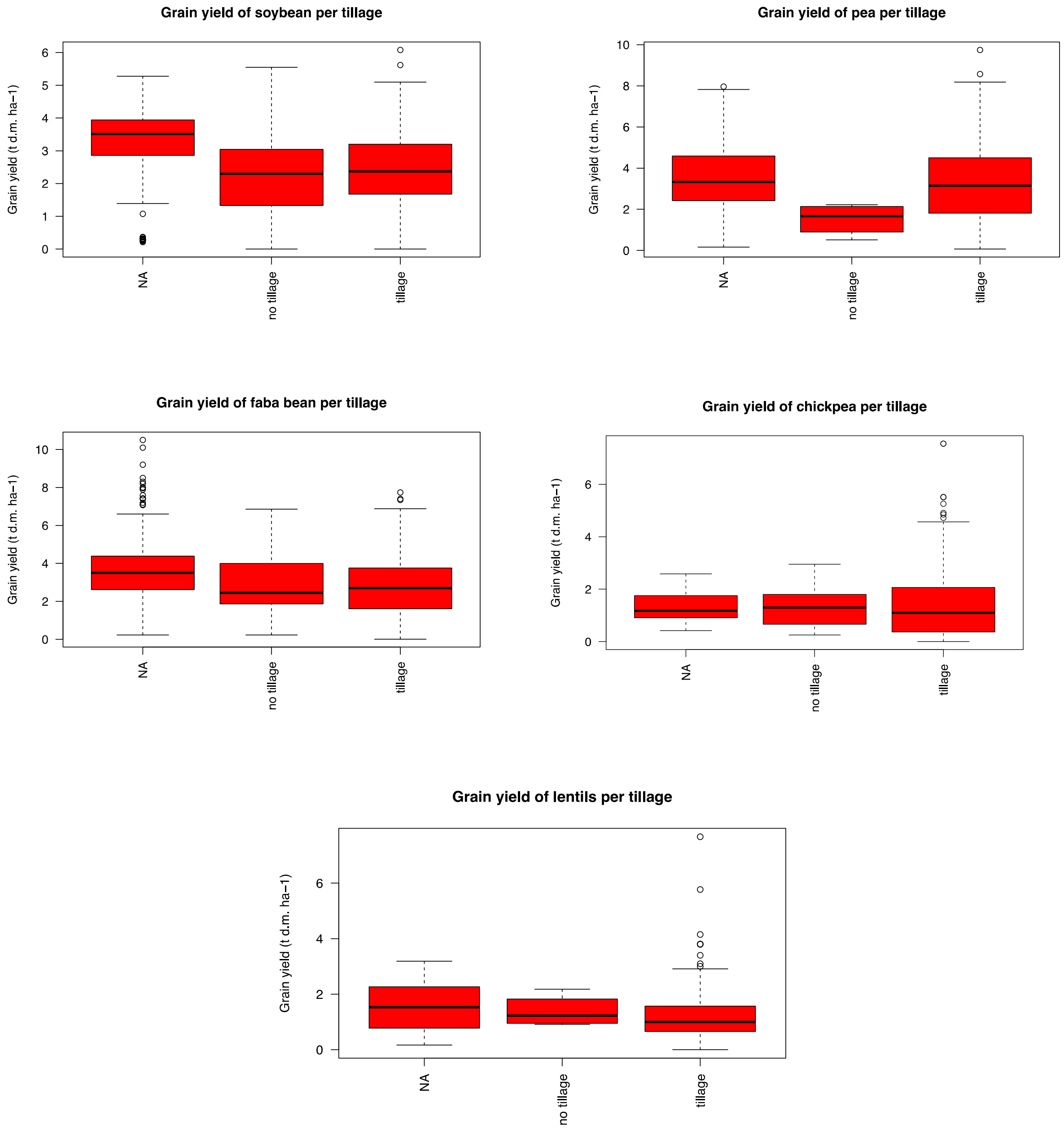
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

Figure 18 - Boxplot of grain yield (t d.m. ha^{-1}) of the 5 legume species as affected by soil texture



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

Figure 19 - Boxplot of grain yield (t d.m. ha^{-1}) of the 5 legume species as affected by soil tillage

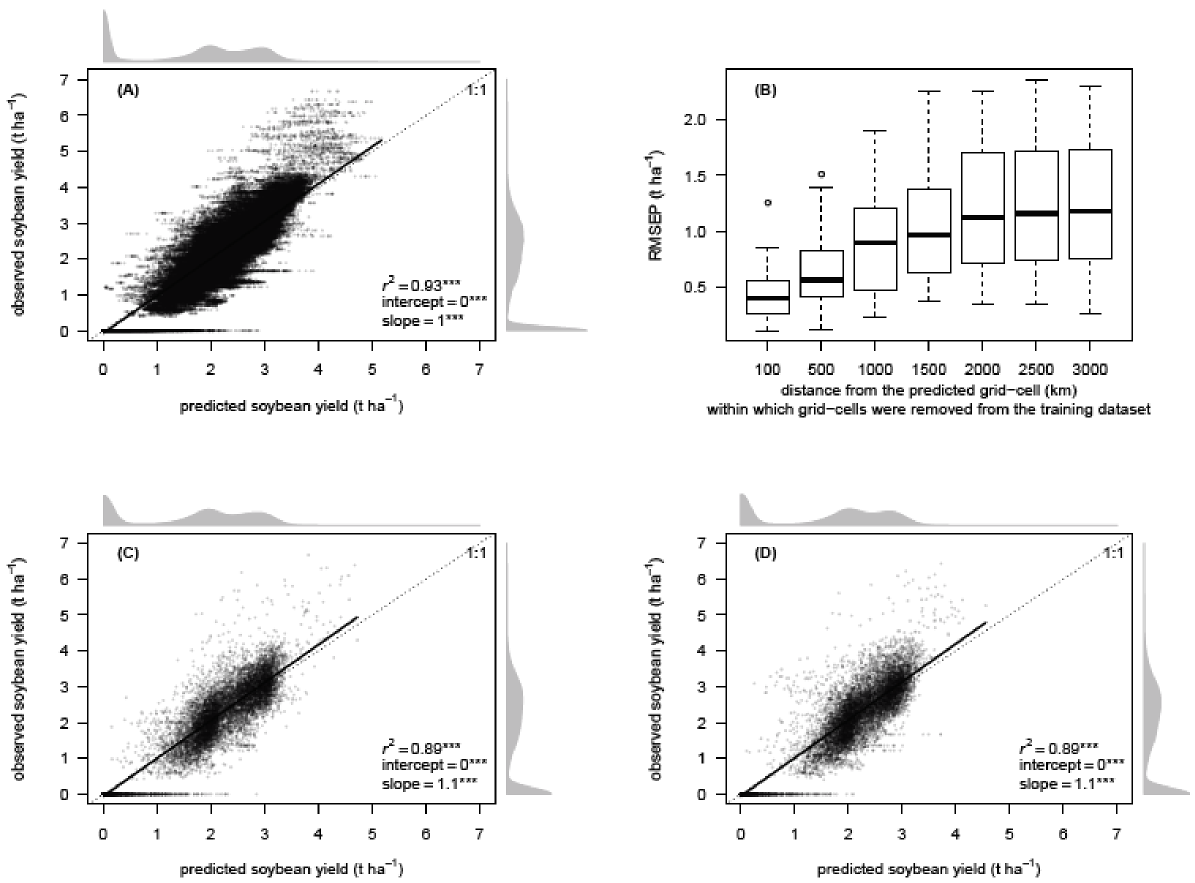


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

4.3. A first example of achievable yield maps for soybean in Europe

Model fitting. Here we present the results obtained for soybean in Europe using the global dataset of historical soybean yield (see section 3.5 for data and methods presentation). Among algorithms tested Random Forest appears to be the most accurate (Figure 20-A) in terms of root-mean-squared error of prediction (RMSEP = 0.35 t ha^{-1}) and Nash–Sutcliffe model efficiency coefficient (MEF = 0.93), as estimated with a classical (unstratified) cross-validation. It also displays the best transferability in time (RMSEP = 0.45 t ha^{-1} when applied to years different from those used for training, Figure 20- C,D) and space (RMSEP = 0.43 t ha^{-1} when applied in locations distant by 500 km, Figure 20-B). Our results reveal that transferability in space decreases with increasing distance between training and test datasets for all models, with a threshold of 1000 km above which the performance of the selected algorithm deteriorates markedly (Figure 20-B).

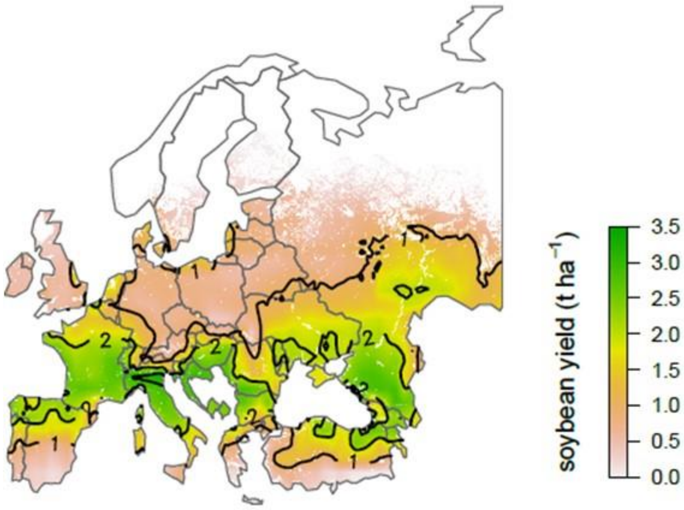
Figure 20 – Assessment of the Random Forest algorithm. (A) The model is first evaluated using a classical bootstrap approach with 25 resamplings. (B) Model transferability in space is then evaluated by ensuring a minimum spatial distance between training and test datasets. Finally, model transferability in time is assessed in (C) where model is fitted on 1981-1995 to predict 1996-2010, and in (D) where model is fitted on 1996-2010 to predict 1981-1995. RMSEP: root mean square error of prediction. Boxplot in panel (B) shows median (center line), 1st and 3rd quartiles (box limits), and 1.5 times the interquartile range (whiskers). Linear regression outputs are shown on panels (A), (C), and (D), as well as marginal distributions of observed and predicted soybean yields (in grey). Dotted lines represent the 1:1 line. In order to extend the range of climate conditions captured by the model and to capture climate conditions leading to zero yield, additional data points were randomly sampled in climate zones known to be unsuitable for soybean production (e.g. deserts and arctic areas) and added to the dataset with their yield value set to zero.



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

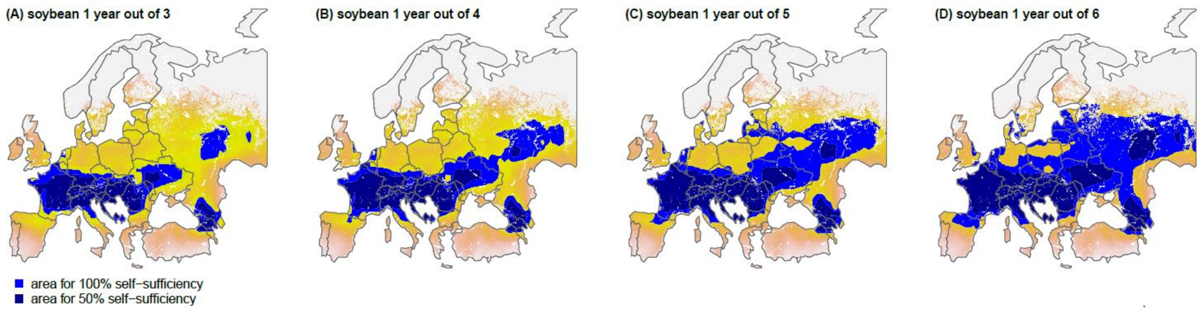
Projections of soybean yield. The projections of the Random Forest algorithm – which assume no irrigation and fixed growing period from April to October – suggest high suitability for soybean under historical climate (Figure 21). About 100 Mha show projected yield equal or higher than 2 t ha^{-1}, while in 2016 the soybean production area in Europe was only 5 Mha with 2 t ha^{-1} of average yield (FAOSTAT, 2019). Therefore, soybean suitable area appears to be much larger than current harvested area in Europe, which suggests that soybean production is not limited by climate conditions.

Figure 21 – Projected soybean yield in Europe under historical climate (1981-2010). Projections are shown only on agricultural area (cropland plus pasture), in the year 2000.



Area requirements for 50% and 100% soybean self-sufficiency in Europe. We estimate the soybean production area required to reach a self-sufficiency level of 50% and 100% based on yield projections presented in Figure 21. A three-step procedure was followed. First, we assumed that soybean could only be grown on current cropland. Under this assumption, soybean cannot be grown in place of permanent pastures, in line with the Common Agricultural Policy of the European Union aiming at their protection. Second, we considered four scenarios for the increase of soybean frequency in crop sequences. In these scenarios, soybean is grown in one year out of three, four, five, or six years, which correspond to 33%, 25%, 20%, and 16% cropland area in a grid-cell under soybean, respectively. These scenarios are consistent with observed and recommended soybean frequencies in crop sequences in Europe. Indeed, a 1-in-3 year or 1-in-4 year soybean cultivation is often recommended to limit the risk of disease occurrence (especially those caused by two fungal pathogens *Sclerotinia sclerotiorum* and *Rhizoctonia solani*), although higher frequencies are observed in Europe and other countries. Third we assumed that soybean is grown preferably in high-yielding grid-cells. Based on this assumption, soybean areas were allocated to grid-cells ranked in decreasing order of projected yield values until the cumulated production (calculated as the product of area and yield) reached 50% and 100% of current annual soybean consumption of Europe (58 Mt in average over 2009-2013[1]). Results suggest that a self-sufficiency level of 50% (100%) would be achievable in Europe under historical climate whatever the frequency of soybean in crop sequences, if ~5% (~10%) of the current European cropland is dedicated to soybean production (Figure 22).

Figure 22 – Area requirements for 50% and 100% soybean self-sufficiency in Europe under historical climate (1981-2000) based on soybean yield projections presented in Figure 21 and assuming various levels of soybean frequency in crop sequences (one year out for three, four, five and six years). Soybean areas were allocated to grid-cells ranked in decreasing order of projected yield values until the cumulated production (calculated as the product of area and yield) reached 50% (light blue) and 100% (dark blue) of the current annual soybean consumption of Europe (58 Mt, average 2009-2013). We assume that soybean can only be grown on current cropland, which excludes permanent pastures in line with the Common Agricultural Policy of the European Union aiming at their protection. Background colors indicate projected soybean yield in t ha^{-1} as in Figure 21.



5. Conclusions

The dataset built under Task 1.2 allowed to consolidate the scientific knowledge on the sources of variability in the yield of grain legumes in Europe, focusing on environmental and agronomical factors. This exercise allowed to confirm the evidence of poor representation of pulses in the existing scientific literature, but also highlighted the existence of grey literature and valuable unpublished results that could add value to the state of the art.

Among the five selected species, soybean and field pea confirmed to be the most studied by scientists, whilst chickpea and lentils are still much constrained by the market dimensions and are not so represented in the scientific production, especially in Northern countries.

Pulses are grown in very contrasting pedoclimatic conditions and this results in quite high variability of grain yields. Although the dataset contributes to identifying some relationships between different species and environmental conditions, additional efforts should be paid to better explore the interactional effects of soils and climates and management. For this aspect, the yield maps targeted to be produced by Task 1.2 will contribute substantially to identify more clear trends.

Among management options, crop genotype is definitively the most present in the studies considered for this activity and in many cases is tested in combination with other agronomic practices. The dataset could represent the baseline for identification of best performing varieties at the European level, thus boosting a wider adoption of cultivars adapted to similar conditions and characterized by desired traits (e.g. pest resistance, weed competitiveness, poor reliance on NP fertilisation, low water demand) and low variability. In this meaning, exploring the dataset addressing the issue of identifying genotypes already tested in different conditions could represent the first step to establish transnational cooperation for genetical improvement of pulses, taking into account not only environmental conditions but also agronomic factors.

Surprisingly, despite the high importance widely recognized to biological N_2 fixation of legumes, N fertilisation is still very present in studies involving also grain legumes. It should be remarked that, although many farmers still consider N fertilisation as a strategy to repair from instability of weather conditions resulting in variable effectiveness of root nodulation by Rhizobia, biological nitrogen fixation could be heavily depressed by availability of mineral forms of N in the soil. Variety trials,

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

seed/root inoculation tests, bioactive compounds and innovation in soil tillage targeted to reduce soil disturbance and improve soil physical quality should be encouraged instead, in order to find out ways to support dinitrogen fixation in legume-based cropping systems.

Finally, the construction of the dataset clearly revealed poor scientific quality or clarity in the existing literature that is constraining a more effective exploitation of the background (e.g. to perform meta-analysis on selected traits). Future research efforts should accomplish for high scientific quality of the results, that is necessary to make significant steps further in the development of innovative and viable solutions to include grain legumes in cropping systems throughout the EU.

The preliminary tests performed on soybean yield estimations were very positive. The selected algorithm (Random Forest) allowed to predict in a reliable way soybean yield under historical climatic conditions. According to preliminary analysis, although not considering the positive effect of irrigation on potential yields, the internal demand of soybean grain seems to be likely met by EU production even if the legume would be present just 1 year over 6 of crop rotations.

6. Acknowledgements

We thank all the LEGVALUE partners that have contributed to this work, in particular by sharing their knowledge and experimental data: INRA, TERIN, WU, PGRO, WR, SSSA, INIAV, AICF, SEGES, LLKC, FH-SWF, CRAN, LAMMC, FiBL.

A special thanks goes to Nicola Guilpart (INRA) who took the lead of modelling activity and yield map generation.

We also acknowledge David Makowski and Nicolas Guilpart (INRA) for their intensive and clever support in the construction of the dataset.

The dataset was built at UNIPI thanks to the work of Silvia Pampana and Lorenzo Gabriele Tramacere who contributed significantly in literature review, data extraction and data entry.

7. References

Cernay, C., E. Pelzer, and D. Makowski. 2016. Data descriptor: A global experimental dataset for assessing grain legume production. Sci. Data 3: 1–20. doi: 10.1038/sdata.2016.84.

Iizumi, T. et al. Historical changes in global yields: Major cereal and legume crops from 1982 to 2006. Glob. Ecol. Biogeogr. 23, 346–357 (2014).

Iizumi, T., Okada, M. & Yokozawza, M. A meteorological forcing data set for global crop modeling: Development, evaluation, and intercomparison. J. Geophys. Res. Atmos. Res. 119, 363–384 (2014).

Magrini, M.-B., Anton, M., Cholez, C., Corre-Hellou, G., Duc, G., Jeuffroy, M.-H., Meynard, J.-M., Pelzer, E., Voisin, A.-S. & Walrand, S. 2016. Why are grain-legumes rarely present in cropping systems despite their environmental and nutritional benefits? Analyzing lock-in in the French agrifood system. Ecological Economics, 126, 152-162.

Pelzer, E., Bourlet, C., Carlsson, G., Lopez-Bellido, R., Jensen, E. & Jeuffroy, M.-H. 2017. Design, assessment and feasibility of legume-based cropping systems in three European regions. Crop and Pasture Science, 68, 902-914

R Core Team, 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.

Schreuder, R. & De Visser, C. 2014. EIP AGRI Focus Group; Protein Crops: final report. European Commission, Brussels, Belgium.

Wickham, H., 2009. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York.

8. Annexes

The grain legume yield dataset (public version without data available only internally to LegValue).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

II. Part – B

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement N°727672

Table of contents

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

1 Index of figures and tables

Tables

Figures

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

2 Summary

Legume crops provide a number of agronomic, environmental and nutritional services. Therefore, increasing their production area has been proposed as a key lever in the agro-ecological transition. But with less than 5% of agricultural land in 2018, legume production area remain very low in the European Union (EU). In this context, identifying legume suitable areas (i.e. regions where it is possible to achieve high and stable yields) appears essential. Here we developed a data-driven approach that combines observed crop yields in field experiments with machine learning techniques to model crop yield from climate inputs in the EU. The fitted models are then apply over the whole EU agricultural area to make yield projections under historical climate (2000-2020), for 5 grain legumes: soybean, field pea, faba bean, chickpea and lentils.

Crop yield data were obtained from an updated version of the dataset presented in Part A of the Deliverable D1.4 of LegValue. This updated version of the dataset is named the "European Grain Legume Dataset" (EGLD) and contains a total of 5217 yield data collected from published and non-published field experiments all over Europe ranging from 1973 to 2020. The EGLD was combined with a global climate dataset designed for crop modelling (JRA55-CDFDM-S14FD dataset). Crop yield was modelled as a function of 4 climate variables (minimum and maximum temperatures, rainfall, and solar radiation) defined at a monthly time step over crop-specific growing seasons. Growing seasons were defined based on observed sowing and harvest months in the EGLD, and winter and spring types were treated as separate crops for field pea, faba bean, chickpea and lentil. The model was fitted using a Random Forest algorithm.

The overall predictive ability of the model was good. R^2 of linear regressions between observed and predicted yield (out-of-bag) were very good for winter chickpea, spring lentil and soybean (R^2>0.8), good for faba bean (spring and winter) and winter pea ($0.7 < R^2 < 0.8$) and medium for spring pea, spring chickpea and winter lentil ($0.59 < R^2 \leq 0.7$). The distributions of model residuals were centred on zero, indicating no systematic bias. Model residuals showed no association with latitude, average in-season *tmax* and total in-season rainfall for any crop, indicating that the model performs equally well under, respectively, low and high latitudes, cool, warm, dry, and wet environments. However, the model over-estimates low yields and under-estimates high yields. Yield projections under historical climate (2000-2020) suggest high suitability for all crops (except winter chickpea) with large areas with high projected yields, e.g. ≤ 2.5 t ha^{-1} for soybean, ≤ 3.5 t ha^{-1} for spring faba bean, and ≤ 4.5 t ha^{-1} for spring pea (all yields at 13% moisture content). However, yield projections for winter lentil and chickpea should be interpreted with caution because of the low amount of data for those crops in the EGLD, which increases the risk of making projections under climate conditions that differ substantially from those under which the model has been calibrated.

Future work include: (i) publication of a data paper describing the public version of the European Grain Legume Dataset, with data hosted on a data repository accessible for download to anyone; (ii) model improvements, including stratified sampling to better handle unbalanced data, adding variables describing evaporative demand from the atmosphere (eg. ETo, VPD) and soil properties (e.g. pH, texture, water-holding capacity) in the model, and using recently developed methods to facilitate model predictions interpretation (partial-dependence plots and Local Interpretable Model-agnostic Explanation); (iii) contribution to the development of a Decision Support Tool aiming at helping farmers to identify how best introducing legume crops into their cropping systems (Legvalue Task 1.4) by integrating the predicted yield values into the Decision Support Tool; (iv) making yield projections under future climate scenarios; (v) publication of the results in scientific journals. Finally, the results of this work are also expected to support future research and development activities on grain legumes in the EU. We believe our results should help breeders to define genetic traits relevant for grain legumes adaption to current and future climatic conditions in the EU, and be of interest to seed companies for estimating the potential seed market for each legume crop (including winter and spring types). By identifying important relationships between climate and crop yield, this work should also provide a useful basis for any further research on the impact of climate change on the development of legumes in the EU.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

3 Introduction

Because of the number of agronomic, environmental and nutritional services they provide, increasing the area under legume crops is often proposed as a key lever in the agro-ecological transition. However, in spite of European and national public policies supporting legumes, their area remains less than 5% of the European Union (EU) agricultural land in 2018 (Food and Agriculture Organization of the United Nations, 2019). Many socio-economic and agronomic factors explain this situation (Magrini et al., 2016; Zander et al., 2016). From an agronomical perspective, the high instability of legume yields has been identified as a key point (Cernay et al., 2015). In order to accompany the development of legumes at the EU level, the identification of areas favourable to their cultivation, i.e. regions where it is possible to achieve high and stable yields, therefore appears essential. Although an initial identification of favourable areas for soybean (*Glycine max*) in Europe has been carried out (Guilpart et al., 2020), this information is not available for several legume species. Moreover, this work has shown the importance of using data collected in Europe to identify favourable areas in a robust manner. However, the databases currently available for legumes (on a global or European scale) contain few data located in Europe. That is why one of the objectives of LegValue WP1 is to try to shed light on these aspects and to determine the actual yield potential of the most common pulse species grown in the EU, namely soybean (*Glycine max*), field pea (*Pisum sativum*), faba bean (*Vicia faba*), chickpea (*Cicer arietinum*), and lentils (*Lens culinaria*).

In Task 1.2 of WP1, we intend to produce EU achievable yield maps for these five major pulse species according to current soil and climate conditions with the following objectives: (i) allow more precise and realistic estimations of protein and starch yield from pulses in current and future scenarios of legume presence in cropping systems; (ii) identify research gaps and technological lock-ins currently hampering a higher share of pulses in cropping systems, (iii) identify sites with high productive potential for pulses that are currently unexplored in the EU, (iv) set for a step up for science on legume crop management by highlighting ways to improve the existing knowledge on the topic.

To generate these maps of achievable yields, a data-driven approach making use of machine-learning techniques to relate observed yields to climate conditions is applied. This approach has been successfully used by (Guilpart et al., 2020) on soybean in Europe, but requires a substantial amount of data reporting observed yields in a range of climate conditions. To this aim, a dataset named the *European Grain Legume Dataset* has been developed by collecting data from: (i) papers published in scientific journals and reporting yields of grain legumes measured in field experiments, (ii) the Legato European research project (http://www.legato-fp7.eu/), (iii) non-published field experiments gathered from LegValue partners. Details about methodology used to develop this dataset, and details descriptive statistics of the final product are available in the Part A of the Deliverable D1.4.

The present report, which is the part B of the deliverable D1.4, contains: (i) a short description of the updated version of the European Grain Legume Dataset that is used here, (ii) details about the methods used to model crop yields from climate inputs, (iii) an assessment of the model quality, and (iv) projections of grain legume yield over the whole European agricultural area based on the fitted model.

4 Materials and methods

○ Data sources

Yield data: The European Grain Legume Dataset

The crop yield dataset used here is an updated version of the dataset described in the deliverable D1.4-Part A. This updated version is named the "European Grain Legumes Dataset" (EGLD) (Antichi et al., 2021). It contains more data and has undergone more detailed quality control. It contains a total of 5217 yield data collected from published and non-published experimentations all over Europe for 5 different pulses: soybean (n=1577), faba bean (n=1474), field pea (n=1331), chickpea (n=451), and lentil (n=384) (Table 1). The dataset covers the 1973-2020 time period, and captures a wide range of yield values, from complete crop failure (yield = 0) to very high yield (> 6 t ha^{-1} of dry matter) (Table 1). The dataset contains a number of other variables in addition to crop yield, including geographic coordinates of experiments (latitude and longitude), sowing and harvest dates, whether irrigation was applied or not (and irrigation quantity if available). Many other variables are available, please refer to Table 5 in appendix for a description of all variables included in the dataset. Each line of the European Grain Legume Dataset corresponds to an experimental unit, defined as a unique combination of year, site, and treatment.

Table 1. Summary statistics of the European Grain Legume Dataset.

Crop	Soybean	Faba bean	Field pea	Chickpea	Lentil	All crops
Number of observations						
Total	1577	1653	1577	451	384	5642
Irrigated	464	171	111	35	10	791
Rainfed	650	793	805	251	220	2719
NA	463	689	661	165	154	2132
Time period						
First year	1973	1981	1980	1988	1993	1973
Last year	2019	2020	2019	2019	2019	2020
Crop yield (t ha^{-1} dry matter)						
Minimum	0.00	0.01	0.07	0.00	0.00	0.00
Maximum	6.08	10.50	9.74	7.55	7.67	10.50
Median	3.14	3.25	3.71	1.14	1.14	3.06
Mean	2.95	3.32	3.70	1.40	1.38	3.04

Historical climate data and climate zones

We used the JRA55-CDFDM-S14FD global retrospective meteorological forcing dataset (Iizumi et al., 2021). This dataset is an updated version of the global retrospective meteorological forcing dataset tailored for agricultural application (GRASP) (Iizumi et al., 2014) with improved spatial resolution and temporal coverage. The JRA55-CDFDM-S14FD dataset has been developed using the bias-corrected Japanese 55-year reanalysis (JRA55), which was bias-corrected for 1958-2020 using the cumulative distribution function-based downscaling method (CDFDM), and the global retrospective meteorological forcing dataset (S14FD) for 1961–2000 as the reference. The JRA55-CDFDM-S14FD dataset contains daily values of five climatic variables relevant to crop growth and yield: maximum (*tmax,* degree Celsius) and minimum (*tmin,* degree Celsius) air temperatures at 2m, precipitation (*rain,* mm day^{-1}), and mean downward shortwave radiation flux (*solar,* W m^{-2}). These variables are available for the period 1958–2020 at a spatial resolution of 0.5 degree. Monthly averages of these variables were calculated, and these monthly values were used to model crop yield from climate inputs. Other meteorological forcing datasets are available (Ruane et al., 2015), but uncertainties associated with different datasets are small at monthly time scale.

To identify in which climate zones the yield data contained in the European Grain Legume Dataset presented in Table 1 were present or absent, we used the climate zonation scheme developed by the Global Yield Gap Atlas (GYGA). This climate zonation scheme has been developed to be relevant to crops and cropping systems (van Wart et al., 2013) and the data are available at http://www.yieldgap.org/download_data. This dataset will be referred to as GYGA-ED (Global Yield Gap Atlas – Extrapolation Domain).

Cropland data

We used the EarthStat data (Monfreda et al., 2008) for total agricultural area (cropland plus pastures), which is representative of agricultural area around the year 2000, and is available at http://www.earthstat.org/cropland-pasture-area-2000/.

○ Model fitting and projections

Data preparation

Data selection. A subset of the European Grain Legume Dataset was used for fitting the model. To get this subset, we first removed all data for which at least one of the following fields was not reported: sowing year, sowing month, latitude, and longitude. When harvest year was missing, it was defined as the year of sowing if sowing occurred before the 1st of September or as the year of sowing +1 if sowing occurred after the 1st of September. Then, all data corresponding to irrigated conditions (indicated as "Y") were removed before modelling. When the information about if irrigation was applied or not was missing (indicated as "NA"), the yield data was kept. Two main reasons underlied this choice: (i) the considered crops are not often irrigated (except soybean), so that the number of irrigated experiments in the dataset is quite low, especially in comparison with rainfed experiments (Table 1), and (ii) the amount of irrigation water applied is not often reported, even when the experiment is indicated as irrigated. We therefore modelled achievable yield under rainfed conditions. All yield data were expressed at a standard moisture content of 13%. The final dataset used for modelling contained a total of 3807 yield data, including 1121 for soybean, 319 for chickpea, 290 for lentil, 1103 for faba bean, and 974 for field pea.

Growing season definition. The distribution of observed sowing and harvest months of selected data are shown in Figure 23. Faba bean, field pea, chickpea and lentil are grown either as winter (sowing occurs from October to December) or spring (sowing occurs mainly from February to April), whereas soybean is always sown between April and May. Based on these results, we defined the growing season from April to October for soybean, which is consistent with (Guilpart et al., 2020). For the other crops, one growing season was defined for winter crops and another one for spring crops (Table 2). These growing seasons were defined to include earliest sowing dates and latest harvest dates observed in Figure 23. Then, all data were classified as winter or spring crops, respectively, if sowing occurred after or before September within a given year.

Linking yield to climate data over the growing season. Each yield data was associated with climate data (from the JRA55-CDFDM-S14FD dataset) over the corresponding crop growing season based on its geographical coordinates and year of sowing.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

Table 2. Growing season used for each crop species. These growing seasons were defined based on Figure 23.

Crop	Spring	Winter
Soybean	April – October	-
Faba bean	February – September	October – August
Field pea	February – September	October – August
Chickpea	January – October	November – July
Lentil	February – September	October – June

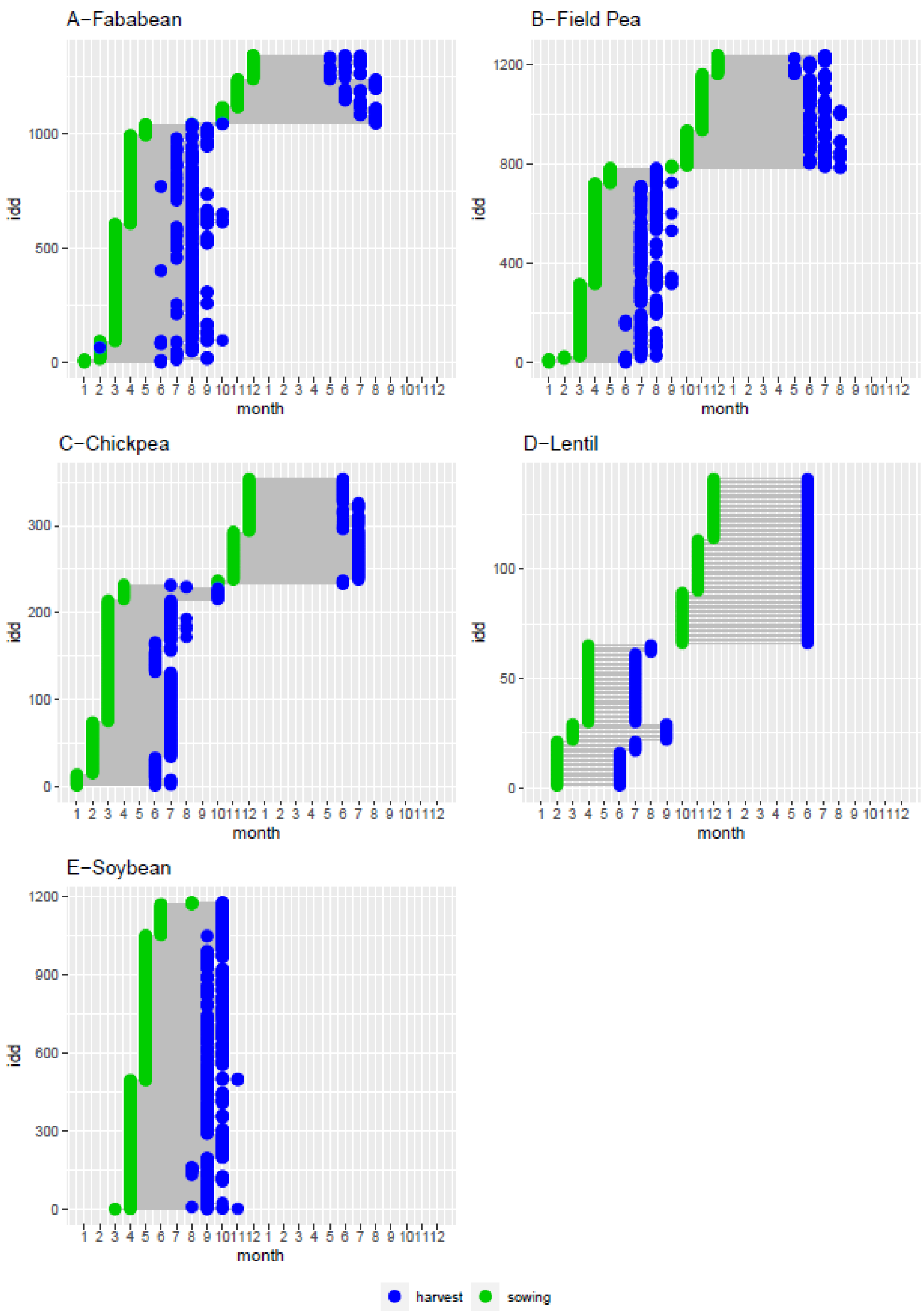


Figure 23. Distribution of sowing and harvest dates by crop in the European Grain Legumes Dataset. Each line is an experimental unit, i.e. a unique combination of experiment, treatment and year. Faba bean (A), field pea (B), chickpea (C), and lentil (D) can be sown as winter or spring crops, whereas soybean (E) is only sown as a spring crop. Growing season of winter crops starts in year *n* and ends in year *n+1*.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

Model fitting and evaluation

Crop yield data was related to the four considered climate variables defined at a monthly time step over the months of the growing season as described in Equation 1, where *tmin* (°C) is the monthly average of daily minimum temperature, *tmax* (°C) is the monthly average of daily maximum temperature, *rain* (mm day^{-1}) is the monthly average of daily rainfall, and *solar* (W m^{-2}) is the monthly average of daily downward shortwave radiation. The number indicated as a suffix indicates the month of the growing season, so that tmin.2 is the average daily minimum temperature in the 2^{nd} month of the growing season. And *n* is the length of the growing season in months. The growing season varied between crops as presented in Table 2. All crops but soybean could be grown either as winter crops or spring crops. Winter and spring types were considered as different crops. We therefore fitted the model described in Equation 1 to the following nine cases: soybean, spring and winter faba bean, spring and winter field pea, spring and winter chickpea, and spring and winter lentil.

Equation 1

$$
\begin{aligned}
yield \sim\ & tmin.1 + tmin.2 + \cdots + tmin.n \\
& + tmax.1 + tmax.2 + \cdots + tmax.n \\
& + rain.1 + rain.2 + \cdots + rain.n \\
& + solar.1 + solar.2 + \cdots + solar.n
\end{aligned}
$$

The model described in Equation 1 was fitted using a Random Forest (RF) algorithm using the *R* software v3.4.0 with the *ranger()* function of the *ranger* package (Wright & Ziegler, 2017) with a number of trees set to 500 and default values for other parameters. Variables importance was measured using the "*impurity*" option of the "*importance*" argument in the *ranger()* function, which corresponds to the variance of the responses for regression. The model predictive ability was assessed using a bootstrap approach with 25 out-of-bag samples generated by bootstrap, using the *train()* function of the *caret* R package, and was measured by computing the R^2 of the linear regression between observed and predicted yields, the root mean square error of prediction (RMSEP, t ha^{-1}), and Nash–Sutcliffe model efficiency (MEF, unitless). A MEF of 1 corresponds to a perfect match of modelled to observed data, a MEF of 0 indicates that model predictions are as accurate as the mean of observed data, whereas a MEF lower than zero occurs when the observed mean is a better predictor than the model. Model residuals (observed yield minus predicted yield) were analysed for their distribution, and relationship with observed yield, latitude, average in-season *tmax* and total in-season rainfall.

Yield projections

Yield projections were performed over the whole European agricultural area using the fitted model for each crop and the JRA55-CDFDM-S14FD climate data. Projections were performed every year from 2000 to 2020. Then the median yield over those years was computed and mapped. All projections assumed no irrigation and the growing seasons presented in Table 2. A common challenge when doing such projections, is to ensure that the combination of environmental conditions under which the model is calibrated are similar to the environmental conditions to which the model is projected, although a reasonable degree of extrapolation might be acceptable (Fitzpatrick & Hargrove, 2009). We addressed this challenge in two ways. First, yield projections are shown only on existing agricultural area (cropland + pastures) (Ramankutty et al., 2008). Second, to identify the climatic conditions captured in the training dataset of our model, we used the GYGA-ED climate zonation scheme (van Wart et al., 2013). For each crop, all climate zones containing at least one experiment of the European Grain Legumes Dataset were selected and mapped. Then, these maps were used to identify areas

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

where projections should be interpreted with caution because they likely involved some extrapolation out of the environmental conditions under which the model was calibrated.

5 Results and discussion

○ Assessment of model performances

Comparison of observed and predicted yields. Values of R^2 for the linear regressions between observed and predicted yield range from 0.59 (winter pea) to 0.91 (winter chickpea) (Table 3). Spring chickpea, winter lentil and winter pea have the lower R^2 (between 0.52 and 0.62) while other crops all have R^2 values higher than 0.7. MEF values are all positive and higher than 0.5 (except for winter pea with MEF=0.43), showing that model predictions are better than the average of observed data. The comparison of observed and predicted yields presented in Figure 24 shows the model has no systematic bias as points align along the 1:1 line for all crops. The model is also able to reproduce the wide range of observed yields for all crops. Based on those results, the model predictive ability can be considered as (i) very good for winter chickpea, spring lentil and soybean, (ii) good for faba bean (spring and winter) and spring field pea, (iii) medium for spring chickpea, winter lentil and winter field pea.

Analysis of model residuals. The distribution of model residuals is centered on zero for all crops (see insets in Figure 24). Model residuals show no association with latitude (Figure 30), average in-season *tmax* (Figure 31) and total in-season rainfall (Figure 32) for any crop. This demonstrates that the model performs equally well under, respectively, low and high latitudes, cool/warm and dry/wet environments. However, model residuals are positively associated with observed yields for all crops (Figure 33). The model therefore over-estimate low yields and under-estimate high yields. This conservative behaviour of the model has already been observed with Random Forest used for crop yield predictions in previous studies (Guilpart et al., 2020; Jeong et al., 2016).

Table 3. Predictive ability metrics of the Random Forest algorithm for the different crops. For each crop, the model is evaluated using a classical bootstrap approach with 25 resamplings, and out-of-bag predictions are compared to observed yields. Then R^2, Root Mean Square Error of Prediction (RMSEP, in t ha^{-1}), and Nash–Sutcliffe model efficiency (MEF, ranges between 0 and 1) are calculated. A MEF of 1 corresponds to a perfect match of modelled to observed data, a MEF of 0 indicates that model predictions are as accurate as the mean of observed data, whereas a MEF lower than zero occurs when the observed mean is a better predictor than the model. Crops are ordered by decreasing value of R^2. Yields values are expressed at a standard moisture content of 13%.

Crop	*n**	Average yield (t ha^{-1})	*R²* (no unit)	RMSEP (t ha^{-1})	MEF (no unit)
Winter chickpea	130	1.33	0.90	0.42 (32%)**	0.84
Spring lentil	149	1.87	0.90	0.43 (23%)	0.83
Soybean	951	3.29	0.84	0.55 (17%)	0.79
Spring faba bean	1006	4.09	0.78	0.84 (21%)	0.75
Winter field pea	420	4.18	0.75	1.91 (46%)	0.73
Winter faba bean	276	2.95	0.73	1.29 (44%)	0.67
Spring field pea	800	4.43	0.70	0.98 (22%)	0.67
Spring chickpea	189	1.11	0.63	0.57 (51%)	0.58
Winter lentil	141	1.17	0.59	0.52 (44%)	0.57

** number of observations in training dataset*

*** value in parenthesis represents RMSEP as a percentage of average yield in training dataset*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

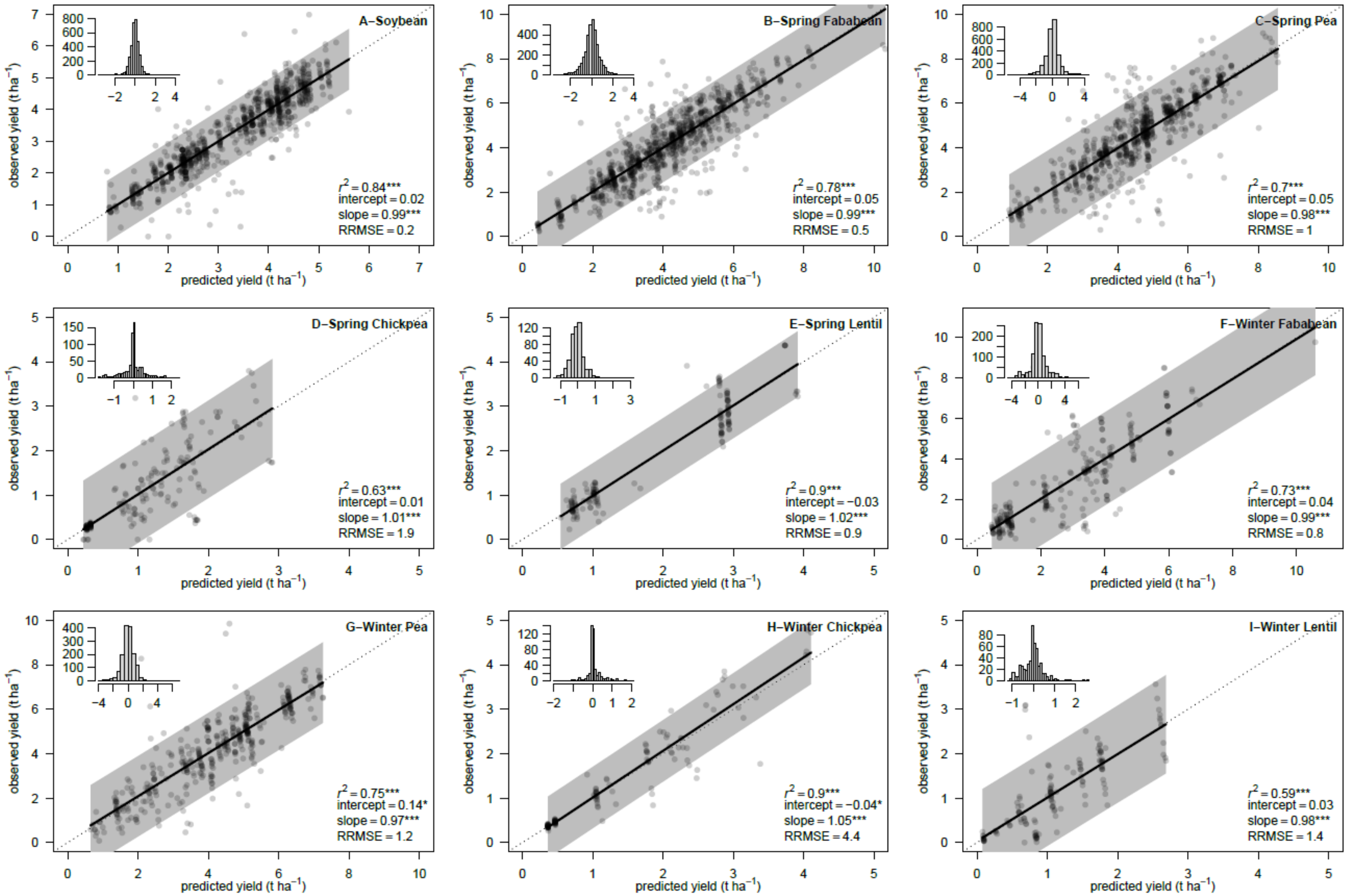


Figure 24. Assessment of the Random Forest algorithm for the different crops considered in this study. For each crop, the model is evaluated using a classical bootstrap approach with 25 resamplings, and out-of-bag predictions are compared to observed yields. Black lines represent the linear regression between observed and predicted yields. Linear regression outputs are shown on the bottom right in each panel. The 95% prediction interval is shown in grey. Dotted lines represent the 1:1 line. Histograms of model residuals are shown as insets. Yield values are expressed at a standard moisture content of 13%.

Variable importance. Variables importance is presented in Figure 34 for all crops and all variables, and

Table 4 gives the 5 most important variables for each crop. The following points can be highlighted.

1. Soybean is the only crop for which a climate variable at the beginning of the growing season appears in the top-5 most influent climate variables. This variable is tmax_1, which suggests soybean is very sensitive to temperature at and right after sowing. This is probably linked with temperature requirements for emergence.
2. As springs crops, lentil, faba bean and pea have the same growing season (from February to September). For these 3 crops, only climate variables from the last 3 months of the growing season (i.e. July to September) appear in the top-5, which suggest a very important role of these last 3 months for yield formation. For spring lentil, 3 climate variables in the top-5 are in month 6 (July), which suggests July is a key month for yield formation.
3. As winter crops, lentil, faba bean and pea have a similar growing season (around October to August). For these three crops, months at the beginning and end of the growing season don't appear in the top-5. For winter pea, 4 variables in the top-5 correspond to rainfall, whereas for winter faba bean no variable in the top-5 correspond to rainfall.
4. For chickpea (both spring and winter), it is mostly temperature variables that appear in the top-5, and no rain.

Table 4. Top 5 most important climate variables for each pulse as identified by the Random Forest algorithm. *tmin* (°C) is the monthly average of daily minimum temperature. *tmax* (°C) is the monthly average of daily maximum temperature. *rain* (mm day^{-1}) is the monthly average of daily rainfall. *solar* (W m^{-2}) is the monthly average of daily downward shortwave radiation. The number indicated as a suffix indicates the month of the growing season. For example, *tmax_5* is the average daily maximum temperature in the 5^{th} month of the growing season which is August for soybean. For winter crops, the growing starts in year *n* and ends in year *n+1*.

Crop	***n****	**Growing season**	**Top 5 most important variables** (by order of importance)				
			1	**2**	**3**	**4**	**5**
Soybean	7	April – October	tmax_1	rain_4	tmax_6	tmax_5	solar_4
Faba bean – spring	8	February – September	tmin_7	rain_7	tmin_8	rain_8	rain_6
Faba bean – winter	11	October – August	tmin_7	solar_5	solar_4	tmax_10	solar_10
Field pea – spring	8	February – September	rain_6	solar_7	tmin_7	rain_8	rain_7
Field pea – winter	11	October – August	rain_7	rain_8	rain_5	solar_7	rain_9
Chickpea – spring	10	January – October	tmax_3	tmin_9	solar_8	tmin_3	tmax_9
Chickpea – winter	9	November – July	tmax_5	solar_9	tmax_3	tmin_3	tmin_7
Lentil – spring	8	February – September	rain_6	solar_6	rain_8	tmin_6	solar_8
Lentil – winter	9	October – June	solar_8	tmax_5	tmax_6	rain_6	solar_7

** length of the growing season in month*

○ Yield projections under historical climate

Assessing where projections are reliable. Yield projections under historical climate (2000-2020) are presented in Figure 25 for spring crops, and in Figure 26 for winter crops. These projections suggest a high climatic suitability (i.e. large areas with high yields) of European agricultural area for the cultivation of all crops, except for winter chickpea. However, caution is needed to interpret these results where projections are made far away from the locations of experiments used to train the model, because climate conditions under which the model has been calibrated might differ substantially from the environmental conditions to which the model is projected (Fitzpatrick & Hargrove, 2009). To address this issue, we identified and mapped the climate zones that contained at

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

least one experiment from the training dataset. These climate zones can be considered as a proxy of the climate validity domain of the model and are presented in Figure 27. Analysis of Figure 27 reveals three groups of crops according to the areal extent of selected climate zones: (i) high coverage (soybean, spring faba bean, and spring field pea), (ii) medium coverage (winter pea, winter faba bean, spring lentil), (iii) low coverage (winter lentil, spring and winter chickpea). Therefore we discuss below projections for the high coverage group as well as options to interpret and improve projections of the two other groups.

Yield projections under historical climate suggest high suitability for most crops. Yield projections suggest high suitability for soybean, spring faba bean, and spring field pea with large areas showing high projected yields (Figure 25 A-C). This confirms that the current extent of harvested areas for these crops is not limited by climate. Projected high-yielding areas appear quite similar for soybean and spring pea, with highest-yielding areas located over an arch from south-western France to southern Germany, Hungary and Romania. For faba bean, highest projected yields (at 13% moisture) concentrate in the north-west of Europe, especially in the UK and Ireland. Projected soybean yield is higher than 2.5 t ha^{-1} almost everywhere in Europe. Spring pea yield is projected to exceed 4.5 t ha^{-1} in the centre of Europe. Spring faba bean yield is projected to be higher than 3.5 t ha^{-1} in a large north-west part of Europe. Although the comparison is difficult because of different spatial scale, these general patterns appear consistent with observed actual yields at national levels from official statistics (Figure 28). As mentioned above, projections must be interpreted with caution for winter pea, winter faba bean, and spring lentil (medium coverage group) and especially for winter lentil, spring and winter chickpea (low coverage group). Yield projections for winter pea, winter faba bean, and spring lentil also suggest high climatic suitability over Europe, especially for winter pea with large areas with projected yield higher than 4 t ha^{-1} (at 13% moisture). For the other crops for which the risk of unjustified extrapolation of the model is high (winter lentil, spring and winter chickpea), we don't discuss the yield projections at this stage, but options to do so are discuss in the "next steps and perspective" section below.

Other factors that may prevent from reaching the projected yield values. We highlight that maps of projected yields shown in Figure 25 and Figure 26 should be interpreted as a kind of "yield potential" maps. We don't refer to yield potential as defined by (Van Ittersum et al., 2013) where water and nutrients are non-limiting and biotic stresses effectively controlled, because we don't know if experiments gathered in the European Grain Legume Dataset fulfil those conditions. However, it is widely recognized that growing conditions in experimental plots are not always similar to the conditions experienced by the crops in commercial farmer's fields, with crop yields measured in experimental plots being often higher than in farmer's fields (Lobell et al., 2009). Moreover, timely sowing is required to ensure a good yield level can be achieved, and this depends at least on two factors that are not taken into account by our models: (i) rainfall distribution within a month, and (ii) constraints on sowing date imposed at the cropping system level, especially by the preceding crop in the crop sequence (Ballot et al., 2019; Rizzo et al., 2021). In addition to these agronomic considerations, we highlight that economic context is likely to influence the feasibility of legume crops as well. This is especially the case where growing another crop (e.g. wheat or maize) is more profitable than the five legume crops considered here. In this case, despite a high projected yield, the probability of growing a legume crop might still be low relatively to other more profitable crops.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

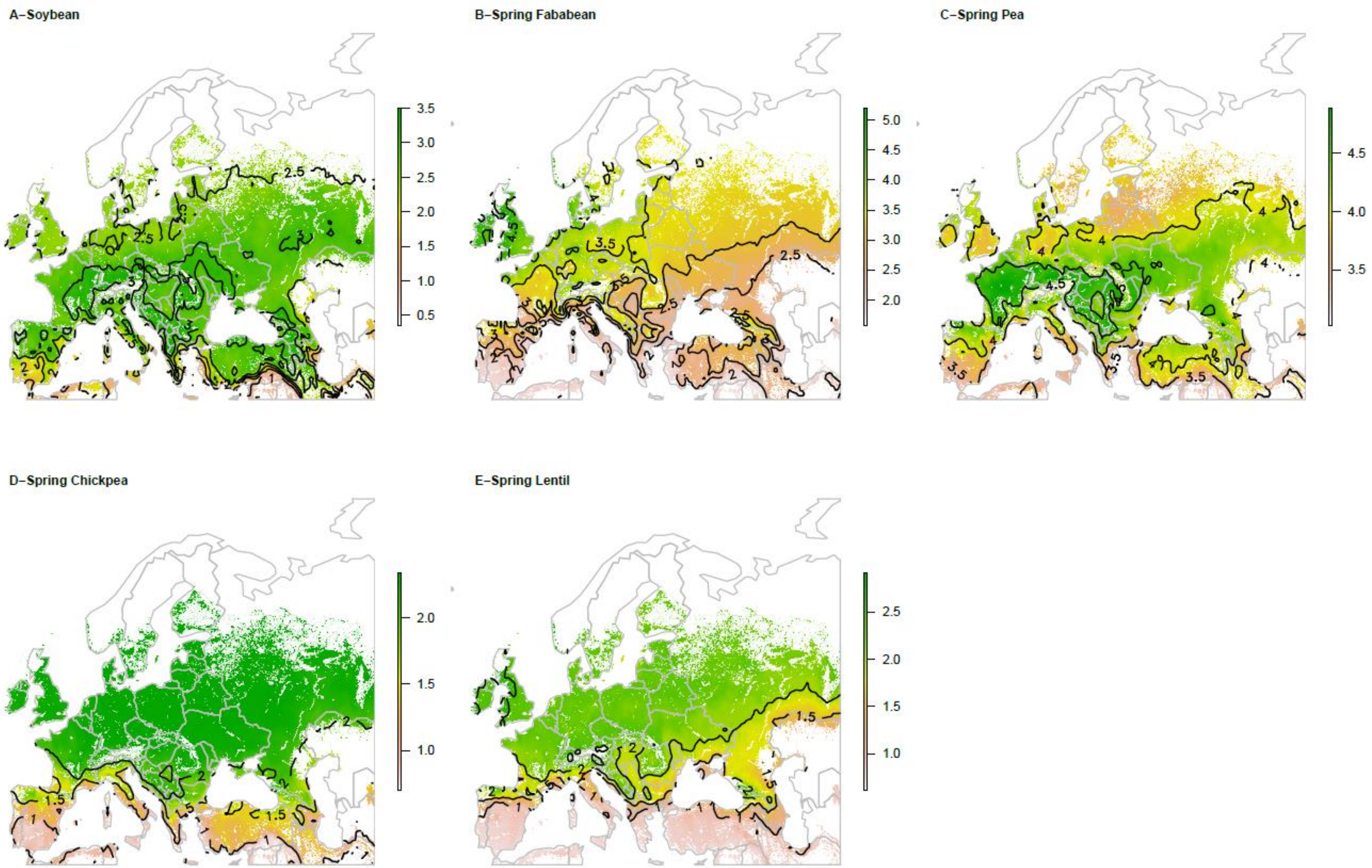


Figure 25. Projected yields (t ha^{-1}) for spring pulses under historical climate (2000-2020). Maps show the median projected yield over years. Projections are shown only on agricultural area (cropland plus pastures) in the year 2000. Yield is expressed at a standard moisture content of 13% for all crops.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

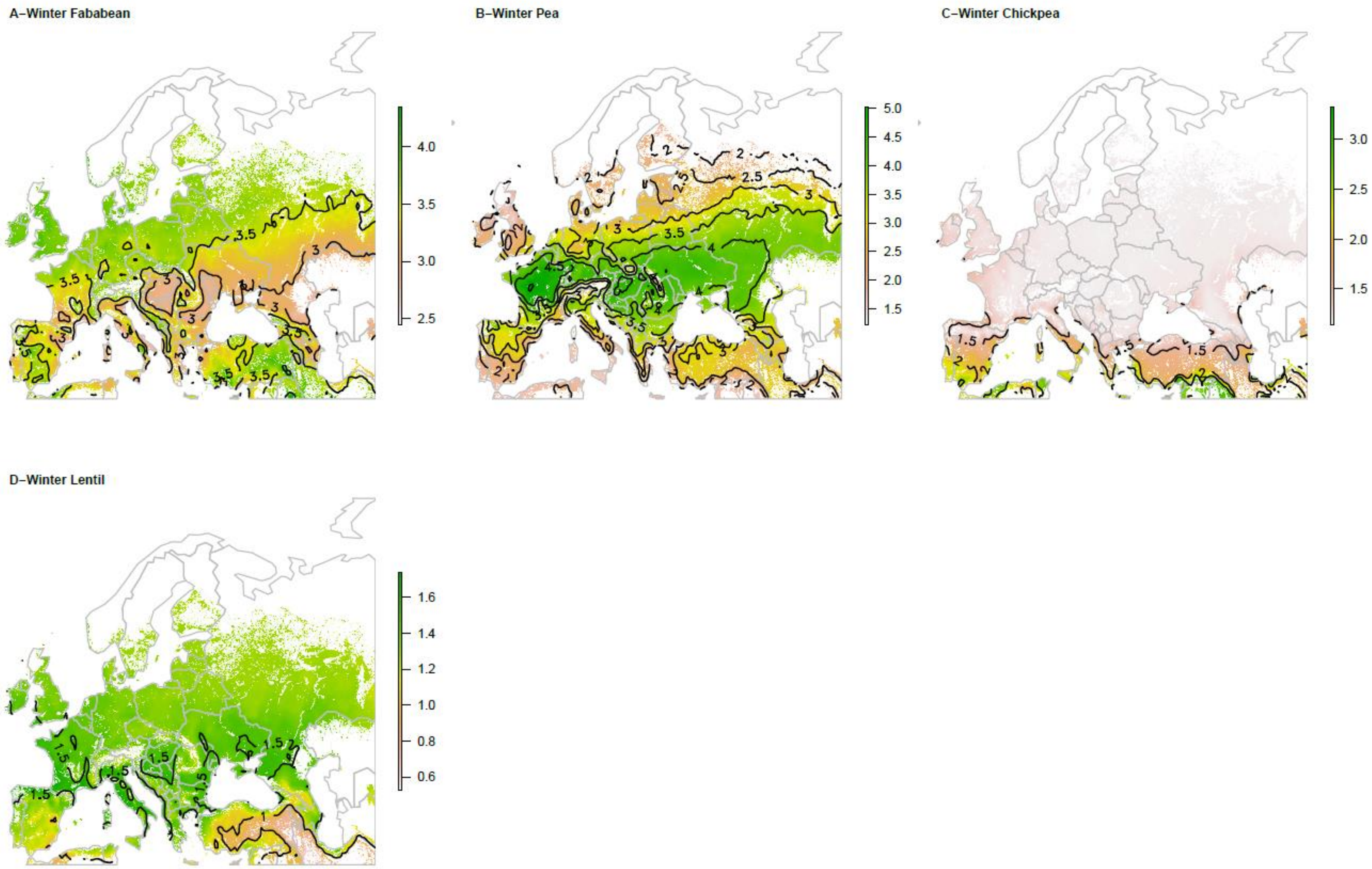


Figure 26. Projected yields (t ha^{-1}) for winter pulses under historical climate (2000-2020). Maps show the median projected yield over years. Projections are shown only on agricultural area (cropland plus pastures) in the year 2000. Yield is expressed at a standard moisture content of 13% for all crops.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

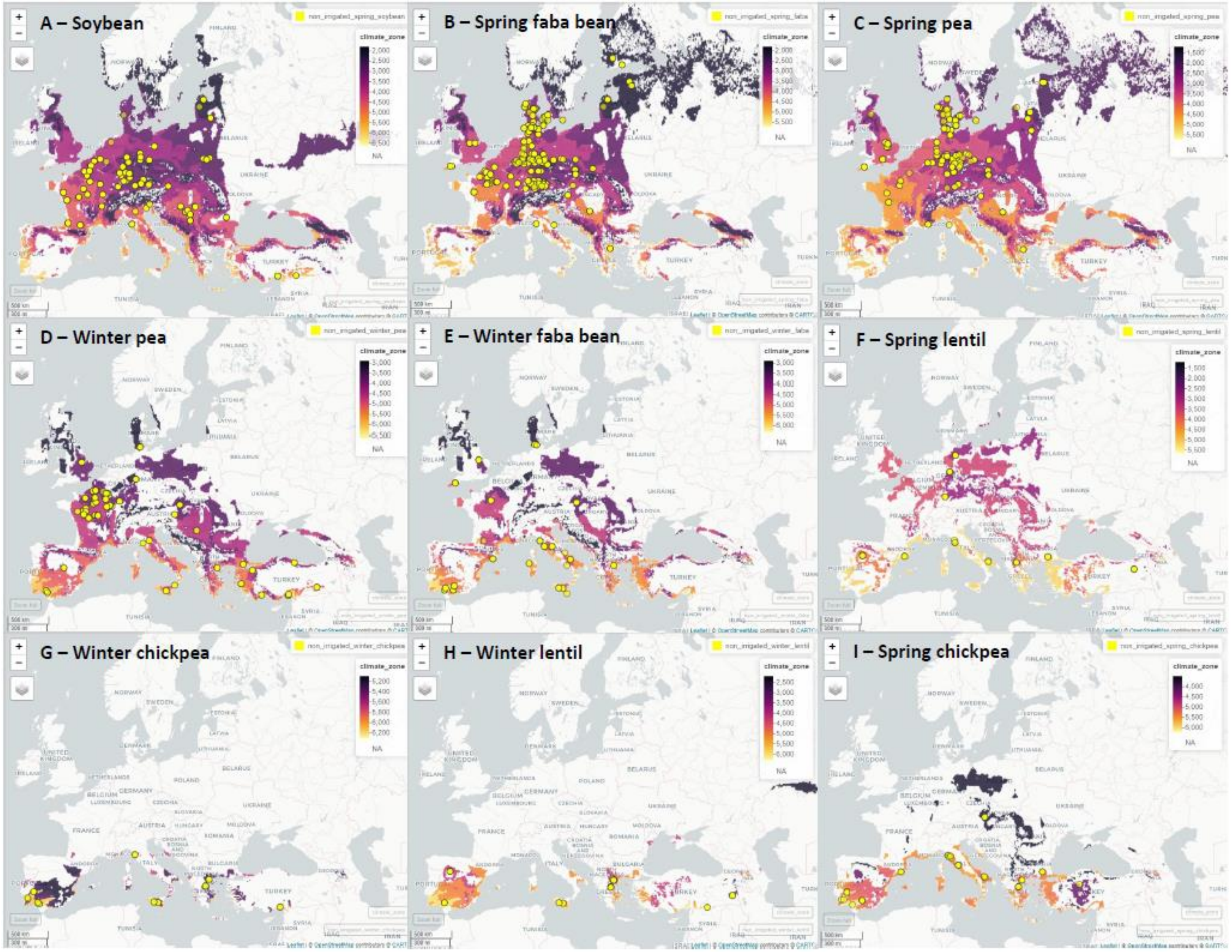


Figure 27. Maps of locations of experiments used for model fitting (yellow dots) and climate zones containing at least one experiment (filled polygons). Climate zones are the Global Yield Gap Atlas Extrapolation Domain (GYGA-ED) available at www.yieldgap.org .

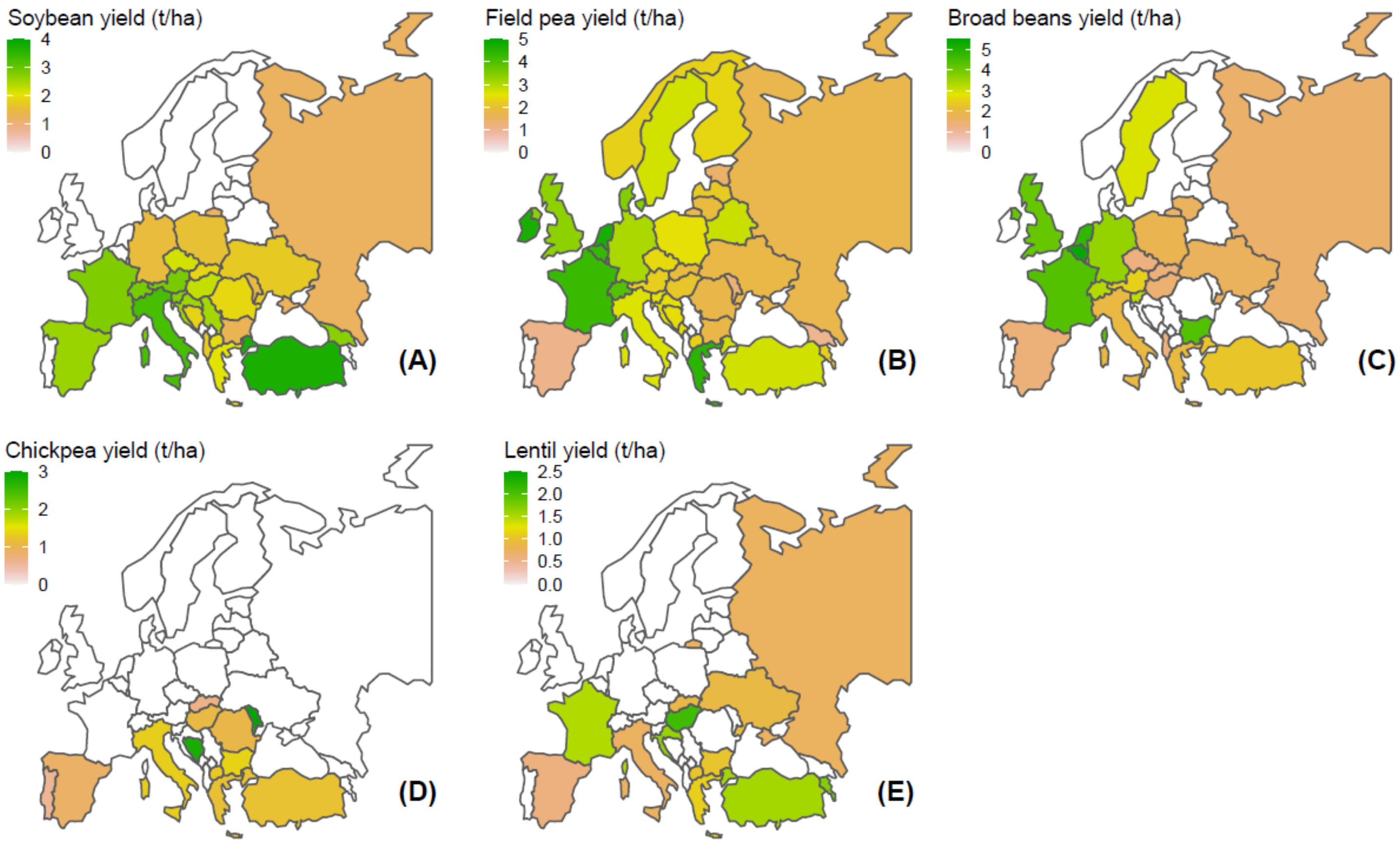


Figure 28. Actual yield (standard moisture content) at the country level (average 2008-2012) for the 5 crops considered here. (A) Soybean, (B) Field pea, (C) Faba bean, (D) Chickpea, (E) Lentil. *Source: FAOSTAT.*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

- ### Next steps and perspectives

Model improvements

More predictors. The model(s) developed and presented in this report use four climate variables (*tmin, tmax, rain, solar*) defined at a monthly time step (Equation 1) and relate those variables to crop yield with a Random Forest algorithm. The model(s) used for yield projections have then been trained over the whole dataset. The results presented in section o show this approach give good results. However, we think the model can be improved in two ways. First, more variables can be included in the predictors. Indeed, in addition to temperature, solar radiation and rainfall, crop growth is also known to be sensitive to the evaporative demand of the atmosphere (Grossiord et al., 2020). This can be characterized by variables like reference evapotranspiration (ETo), relative humidity (RH), vapor pressure (VP), or vapor pressure deficit (VPD). The JRA55-CDFDM-S14FD climate data do include relative humidity, specific humidity and wind (T. Iizumi et al., 2021), which will allow us to add new variables related to air humidity in the model. Crop growth is also known to be sensitive to soil type, either through the soil water-holding capacity in the root zone (Guilpart et al., 2017), or through physico-chemical properties like pH, cation exchange capacity (CEC), soil organic matter content (SOC) and texture (Toshichika Iizumi & Wagai, 2019; Islam et al., 1980). Global and regional maps of soil properties do exist, which provide some of the above-mentioned variables (Batjes, 2016; de Sousa et al., 2020; Orgiazzi et al., 2018). We believe including more climate variable and some soil properties in the predictors should improve the model(s).

Handling unbalanced data with stratified sampling to train random forest. As shown in **Figure 27. Maps of locations of experiments used for model fitting (yellow dots) and climate zones containing at least one experiment (filled polygons)**. Climate zones are the Global Yield Gap Atlas Extrapolation Domain (GYGA-ED) available at www.yieldgap.org .Figure 27, the geographic distribution of field experiments that were used to train the Random Forest algorithm is not homogeneous: some countries are over-represented in the dataset while others are under-represented. Because of this geographical bias, the training dataset of the model is considered as unbalanced. The effect of unbalanced datasets due to sampling bias on the predictive ability of species distribution models in ecology has been well studied, and some methods have been proposed to deal with it (Fourcade et al., 2014; Gaul et al., 2020). Among those methods, the stratified sampling approach proposes to create an artificially balanced training dataset by performing a stratified sampling of the initial biased dataset. The stratification can be made based on geographical space (e.g. systematic sampling) (Fourcade et al., 2014) or based on predictors values (Gaul et al., 2020). Even if data quantity has been shown to be more important than its spatial bias for predictive species distribution modeling, we argue that a stratified sampling approach could improve the predictive ability of our models.

Develop a model for irrigated conditions. The models presented in this report have been developed for purely rainfed conditions only. Although experiments that applied irrigation are available in the European Grain Legume Dataset, they were removed before model training. The main reasons underlying this choice are: (i) the considered crops are not often irrigated, so that the number of irrigated experiments in the dataset is quite low, especially in comparison with rainfed experiments (Table 1), and (ii) the amount of irrigation water applied is not often reported, even when the experiment is indicated as irrigated. However, developing a model for irrigated conditions, at least for soybean (which is the most often irrigated crop considered here), appear as an interesting perspective to possibly highlight higher yield potential levels when irrigation is also applied.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

Improved model evaluation and interpretation

Assessing model transferability in time and space. Recent papers have highlighted the importance of rigorous cross-validation strategies to ensure that the predictive capacity of a given algorithm is evaluated on data as independent as possible from the data used to train that algorithm (Fourcade et al., 2018; Roberts et al., 2017). Following (Guilpart et al., 2020) we will run two cross-validation strategies to assess transferability of our models in time and space. Transferability in time will be assessed by splitting the dataset into two periods in order to assess the ability of each algorithm to predict a period of time different from the one used for the training, while transferability in space will be assessed by ensuring a minimum spatial distance between training and test datasets as in (Guilpart et al., 2020).

Toward interpretable machine learning models. Machine learning models are often considered as black-boxes because the reasons underlying their predictions are not easy to identify. However, recent advances in the so-called field explainable Artificial Intelligence are providing some tools to overcome this difficulty. Three of them can be mentioned: (i) measures of variable importance over the whole training dataset (presented in Figure 34), (ii) partial-dependence plots that allows analysing the effect of one single variable on yield, (iii) estimation of variables contributions to an individual prediction. Partial-dependence plots are interesting because they allow to check whether a variable has an impact on yield that is consistent with the current knowledge of the crop's physiology. For example, Guilpart et al. (2020) show that *tmax* in the first month of the growing season had a positive impact on soybean yield, especially above a threshold of 4°C that corresponds to the base temperature for germination. This kind of findings reinforces greatly the confidence in the model and therefore in its projections. We will look into partial-dependence plots for selected variables for all crops considered in this report to check whether their impact on yield is consistent with our current knowledge of their physiology. Then we will use recently developed methods to estimate variables contributions to an individual prediction, such as the LIME method (Local Interpretable Model-agnostic Explanation) (Ryo et al., 2020) which is already implemented into the *Lime* R Package. This will allow to identify climatic drivers of yield projections at a specific location. We believe this will be helpful to (i) analyse consistency with crop physiology, (ii) discuss with local agronomists of the plausibility of the model outputs. Organizing a workshop with LegValue partners who provided data to the European Grain Legume Dataset to discuss this kind of modelling outputs might be an interesting and valuable option.

Identify where yield projections are reliable. We mentioned earlier that caution is needed to interpret yield projections where they are made far away from the locations of experiments used to train the model, because climate conditions under which the model has been calibrated might differ substantially from the environmental conditions to which the model is projected (Fitzpatrick & Hargrove, 2009). Results in Figure 25 and Figure 26 show this is the case for winter lentil, spring and winter chickpea. We see two options to deal with this situation: (i) show projections only in climate zones where experiments have already being carried out, (ii) show projections only in areas where climate similarity would be higher than a given threshold (e.g. 80%) to at least one experimental site. However this requires further research.

Yield projections under climate change

Similarly to Guilpart et al. (2020), projections under climate change scenarios will be made using 16 climate change scenarios consisting of bias-corrected data of eight Global Circulation Models (GCM; GFDL-ESM2M, HadGEM2-ES, IPSL-CM5A-LR, MIROC5, MIROC-ESM, MIROC-ESM-CHEM, MRI-CGCM3, and NorESM1-M) used in the Coupled Model Intercomparison phase 5 (CMIP5) (Taylor et al., 2012) and two Representative Concentration Pathways (RCPs; 4.5 and 8.5 W m^{-2}) (Van Vuuren et al., 2011). Details on the bias-correction method used is available in (Minamikawa et al., 2016). Although daily data are available in the bias-corrected GCMs outputs, we will compute and use monthly data in our analysis. We will consider three time periods for projections: 1981-2010 (historical), 2050-2059 (mid-century), and 2090-2099 (end of the century). We will present the median predicted yield over the eight GCMs.

- **Data accessibility**

The European Grain Legume Dataset. As mentioned earlier, the European Grain Legume Dataset contains data from: (i) papers published in scientific journals, (ii) the Legato European research project, and (iii) non-published field experiments from LegValue partners. If data from (i) and (ii) are already publicly available, data from (iii) may be publicly available or not depending on the decision of the institution who owns the data. In line with open science principles, we will make publicly available as much data as possible. Therefore, a data paper will be published that will include a description of the public version of the European Grain Legume Dataset and the data will be hosted on a data repository (e.g. www.zenodo.org) accessible for download to anyone. The full EGLD (public and non-public version) will however be available only on request from LegValue partners to Daniele Antichi (daniele.antichi@unipi.it) for internal use only within the context of the LegValue project.

Maps of yield projections under historical and future climate scenarios. The maps of yield projections generated for soybean, pea, faba bean, chickpea, and lentils under historical climate and future climate scenarios will be made available for download to anyone in geoTIFF or netCDF format. They will be posted on a data repository (e.g. www.zenodo.org) when the corresponding paper(s) will be published in appropriate scientific journals. They will also be available on request to Nicolas Guilpart (nicolas.guilpart@agropatistech.fr) before publication.

6 Acknowledgements

We thank all the LEGVALUE partners that have contributed to this work, in particular by sharing their knowledge and experimental data: INRAE, TERIN, UNIPI, WU, PGRO, WR, SSSA, INIAV, AICF, SEGES, LLKC, FH-SWF, CRAN, LAMMC, FiBL. We acknowledge David Makowski (INRAE) for his support in designing the final structure of the European Grain Legume Dataset and his helpful statistical insights for modelling.

7 References

Antichi, D., Jeuffroy, M.-H., Makowski, D., Tramacere, L. G., Pampana, S., Bertin, I., Biarnès, V., & Guilpart, N. (2021). "The European Grain Legume Dataset » : un jeu de données expérimentales pour prédire le rendement des légumineuses à graines en Europe. *3e Rencontres Francophones Des Légumineuses*.

Ballot, R., Guilpart, N., Pelzer, E., & Jeuffroy, M.-H. (2019). Current dominant crop sequences across EU: a typology based on LUCAS dataset. *European Conference on Crop Diversification (ECCD)*. https://doi.org/https://doi.org/10.5281/zenodo.3492238

Batjes, N. H. (2016). Geoderma Harmonized soil property values for broad-scale modelling (WISE30sec) with estimates of global soil carbon stocks. *Geoderma*, *269*, 61–68. https://doi.org/10.1016/j.geoderma.2016.01.034

Cernay, C., Ben-Ari, T., Pelzer, E., Meynard, J.-M., & Makowski, D. (2015). Estimating variability in grain legume yields across Europe and the Americas. *Scientific Reports*, *5*, 11171.

de Sousa, L. M., Poggio, L., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Riberio, E., & Rossiter, D. (2020). SoilGrids 2.0: producing quality-assessed soil information for the globe. *Soil*, *under revi*.

Fitzpatrick, M. C., & Hargrove, W. W. (2009). The projection of species distribution models and the problem of non-analog climate. *Biodiversity and Conservation*, *18*(8), 2255–2261. https://doi.org/10.1007/s10531-009-9584-8

Food and Agriculture Organization of the United Nations. (2019). *FAOSTAT Statistics Database*. http://www.fao.org/faostat/en/#data

Fourcade, Y., Besnard, A. G., & Secondi, J. (2018). Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics. *Global Ecology and Biogeography*, *27*(2), 245–256. https://doi.org/10.1111/geb.12684

Fourcade, Y., Engler, J. O., Rödder, D., & Secondi, J. (2014). Mapping species distributions with MAXENT using a geographically biased sample of presence data: A performance assessment of methods for correcting sampling bias. *PLoS ONE*, *9*(5), 1–13. https://doi.org/10.1371/journal.pone.0097122

Gaul, W., Sadykova, D., White, H. J., Leon-Sanchez, L., Caplat, P., Emmerson, M. C., & Yearsley, J. M. (2020). Data quantity is more important than its spatial bias for predictive species distribution modelling. *PeerJ*, *8*, 1–27. https://doi.org/10.7717/peerj.10411

Grossiord, C., Buckley, T. N., Cernusak, L. A., Novick, K. A., Poulter, B., Siegwolf, R. T. W., Sperry, J. S., & McDowell, N. G. (2020). Plant responses to rising vapor pressure deficit. *New Phytologist*, *226*(6), 1550–1566. https://doi.org/10.1111/nph.16485

Guilpart, N., Grassini, P., van Wart, J., Yang, H., van Ittersum, M. K., van Bussel, L. G. J., Wolf, J., Claessens, L., Leenaars, J. G. B., & Cassman, K. G. (2017). Rooting for food security in Sub-Saharan Africa. *Environmental Research Letters*, *12*, 114036.

Guilpart, N., Iizumi, T., & Makowski, D. (2020). Data-driven yield projections suggest large opportunities to improve Europe's soybean self-sufficiency under climate change. *BioRxiv*, 2020.10.08.331496. https://doi.org/10.1101/2020.10.08.331496

Iizumi, T., Ali-Babiker, I.-E. A., Tsubo, M., Tahir, I. S. A., Kurosaki, Y., Kim, W., Gorafi, Y. S. A., Idris, A. A. M., & Tsujimoto, H. (2021). Rising temperatures and increasing demand challenge wheat supply in Sudan. *Nature Food*.

Iizumi, Toshichika, Okada, M., & Yokozawza, M. (2014). A meteorological forcing data set for global crop modeling: Development, evaluation, and intercomparison. *Journal of Geophysical Research: Atmospheres RESEARCH*, *119*, 363–384. https://doi.org/10.1002/2013JD020222.Received

Iizumi, Toshichika, & Wagai, R. (2019). Leveraging drought risk reduction for sustainable food, soil and climate via soil organic carbon sequestration. *Scientific Reports*, *9*(1), 1–8. https://doi.org/10.1038/s41598-019-55835-y

Islam, A. K. M. S., Edwards, D. G., & Asher, C. J. (1980). pH optima for crop growth. *Plant and Soil*, *54*(3), 339–357. https://doi.org/10.1007/bf02181830

Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., Timlin, D. J., Kyo-Moon, S., Gerber, J. S., Reddy, V. R., & Kim, S.-H. (2016). Random Forests for Global and Regional Crop Yield Predictions. *PLoS One*, *11*(6), e0156571. https://doi.org/10.1371/journal.pone.0156571

Lobell, D., Cassman, K., & Field, C. (2009). Crop yield gaps: their importance, magnitudes, and causes. *Annual Review of Environment and Resources*, *34*. https://doi.org/10.1146/annurevfienviron.041008.093740

Magrini, M. B., Anton, M., Cholez, C., Corre-Hellou, G., Duc, G., Jeuffroy, M. H., Meynard, J. M., Pelzer, E., Voisin, A. S., & Walrand, S. (2016). Why are grain-legumes rarely present in cropping systems despite their environmental and nutritional benefits? Analyzing lock-in in the French agrifood system. *Ecological Economics*, *126*, 152–162. https://doi.org/10.1016/j.ecolecon.2016.03.024

Minamikawa, K., Fumoto, T., Iizumi, T., Cha-un, N., & Pimple, U. (2016). Prediction of future methane emission from irrigated rice paddies in central Thailand under different water management practices. *Science of the Total Environment*, *566–567*, 641–651. https://doi.org/10.1016/j.scitotenv.2016.05.145

Monfreda, C., Ramankutty, N., & Foley, J. A. (2008). Farming the planet : 2. Geographic distribution of crop areas

, yields , physiological types , and net primary production in the year 2000. *Global Biogeochemical Cycles*, *22*, 1–19. https://doi.org/10.1029/2007GB002947

Orgiazzi, A., Ballabio, C., Panagos, P., Jones, A., & Fernández-Ugalde, O. (2018). LUCAS Soil, the largest expandable soil dataset for Europe: a review. *European Journal of Soil Science*, *69*(1), 140–153. https://doi.org/10.1111/ejss.12499

Ramankutty, N., Evan, A. T., Monfreda, C., & Foley, J. A. (2008). Farming the planet : 1. Geographic distribution of global agricultural lands in the year 2000. *Global Biogeochemical Cycles*, *22*(August 2007), 1–19. https://doi.org/10.1029/2007GB002952

Rizzo, G., Monzon, J. P., & Ernst, O. (2021). Cropping system-imposed yield gap: Proof of concept on soybean cropping systems in Uruguay. *Field Crops Research*, *260*(December 2019), 107944. https://doi.org/10.1016/j.fcr.2020.107944

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), 913–929. https://doi.org/10.1111/ecog.02881

Ruane, A. C., Goldberg, R., & Chryssanthacopoulos, J. (2015). Climate forcing datasets for agricultural modeling: Merged products for gap-filling and historical climate series estimation. *Agricultural and Forest Meteorology*, *200*, 233–248. https://doi.org/10.1016/j.agrformet.2014.09.016

Ryo, M., Angelov, B., Mammola, S., Kass, J. M., Benito, B. M., & Hartig, F. (2020). Explainable artificial intelligence enhances the ecological interpretability of black-box species distribution models. *Ecography*, 1–7. https://doi.org/10.1111/ecog.05360

Taylor, K. e., Stouffer, R. J., & Meehl, G. A. (2012). An Overview of CMIP5 and experiment design. *American Meteorological Society*, *93*, 485–498. https://doi.org/10.1175/BAMS-D-11-00094.1

Van Ittersum, M. K., Cassman, K. G., Grassini, P., Wolf, J., Tittonell, P., & Hochman, Z. (2013). Yield gap analysis with local to global relevance-A review. *Field Crops Research*, *143*, 4–17. https://doi.org/10.1016/j.fcr.2012.09.009

Van Vuuren, D. P., Edmonds, J., Kainuma, M., Riahi, K., Thomson, A., Kathy, H., Hurtt, G. C., Kram, T., Krey, V., Lamarque, J.-F., Masui, T., Meinshausen, M., Nakicenovic, N., Smith, S. J., & Rose, S. K. (2011). The representative concentration pathways : an overview. *Climatic Change*, *109*, 5–31. https://doi.org/10.1007/s10584-011-0148-z

van Wart, J., van Bussel, L. G. J., Wolf, J., Licker, R., Grassini, P., Nelson, A., Boogaard, H., Gerber, J., Mueller, N. D., Claessens, L., van Ittersum, M. K., & Cassman, K. G. (2013). Use of agro-climatic zones to upscale simulated crop yield potential. *Field Crops Research*, *143*, 44–55. https://doi.org/10.1016/j.fcr.2012.11.023

Wright, M. N., & Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal Of Statistical SoftwareSoftware*, *77*(1), 1–17. https://doi.org/10.18637/jss.v077.i01

Zander, P., Amjath-Babu, T. S., Preissel, S., Reckling, M., Bues, A., Schläfke, N., Kuhlman, T., Bachinger, J., Uthes, S., Stoddard, F., Murphy-Bokern, D., & Watson, C. (2016). Grain legume decline and potential recovery in European agriculture: a review. *Agronomy for Sustainable Development*, *36*(2). https://doi.org/10.1007/s13593-016-0365-y

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement N°727672

8 Appendix

Table 5. List and description of variables included in the European Grain Legume Dataset.

Variable name	Variable description
id	Entry ID
source	Experiment/Paper
publicly_available	Y/N (YES if this data could be part of an open-access publication of the dataset on a Data Journal; N if for internal use in LegValue)
experiment_ID	Experiment acronym_Surname of responsible or FirstAuthorSurnameYEAR
site_country	Name of the country in full
site_region	NUTS 3. "NA" if not available
site_name	Name of the site in full. "NA" if not available
lat	Decimal degrees of Latitude (XX.xx). "NA" if not available
lat_cardinal	N/S. . "NA" if not available
lon	Decimal degrees of Longitude (XX.xx). "NA" if not available
lon_cardinal	W/E . "NA" if not available
site_soil_classification_name	Soil classification type (USDA). "NA" if not available
site_soil_texture_name	Soil texture class (e.g. loam, sandy, sandy loam). "NA" if not available
soil_texture_anomaly	Soil texture reported is not standard (Y/N)
site_rain	Total rainfall (mm) in the period considered. "NA" if not available
site_rain_period	annual/growing season. "NA" if not available
site_rain_period_month	Initial Final month of the period for which precipitations are reported (e.g. Jan Dec). "NA" if not available
site_rain_period_year	Years of registration of the precipitations (e.g. 1993). "NA" if not available
site_temp	Average temperature (°C) in the considered period. "NA" if not available
site_temp_period	annual/growing season. "NA" if not available
site_temp_period_month	Initial Final month of the period for which temperature is reported (e.g. Jan Dec). "NA" if not available
site_temp_period_year	Years of registration of the temperature (e.g. 1993). "NA" if not available
organic_farming	Y/N
management_evaluated	e.g. tillage, irrigation, variety
treatment_name	Report the name of the treatment or (in case of factorial combination) the name of the combination
scientific_name	Latin name (without author initials) of the legume crop species (e.g. Glycine max)
previous_crop	Latin name (without author initials) of the crop species grown before the legume (e.g. Triticum aestivum). "NA" if not available

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

crop	Common name of the legume crop species (e.g. Soybean)
crop_type	Field pea: "green" or "dry". Faba bean: "horse" for var. equina, "pigeon" for var. minor, "broad" for var. major
cultivar	Name of the legume crop variety. "NA" if not available
precocity_group	Only for soybean. Report here the precocity group (000 to 10)
gm	Genetically modified variety? Y/N
sow_dd	Day of sowing as originally reported in the source document. If more than one, the range is reported. "NA" if not available
sow_mm	Month of sowing as originally reported in the source document. If more than one, the range is reported. "NA" if not available
sow_yy	Year of sowing as originally reported in the source document. If more than one, the range is reported. "NA" if not available
sow_date	mm/dd/yyyy. "NA" if not available
har_dd	Day of harvest as originally reported in the source document. If more than one, the range is reported. "NA" if not available
har_mm	Month of harvest as originally reported in the source document. If more than one, the range is reported. "NA" if not available
har_yy	Year of harvest as originally reported in the source document. If more than one, the range is reported. "NA" if not available
har_date	mm/dd/yyyy. "NA" if not available
cycle_length	Length of crop cycle in the experimental year (nr. of days from sowing to harvest). "NA" if not available
tillage	"Y" if a tillage operation is performed before legume sowing, or "N" if sod-seeding legume
plant_density	nr of legume plants per m2 (alternative to sowing density). "NA" if not available
sowing_density	nr of legume seeds per m2 (alternative to plant density). "NA" if not available
row_spacing	inter-row space in meters. "NA" if not available
N_rate	Total amount of N (kg ha^{-1}) supplied to the crop
N_fertiliser type_1	Name(s) of the first N fertiliser applied to the crop with its level of N application rate (kg N ha-1) (e.g. Poultry manure -30-). "NR" if not relevant (if fertilisation is not applied)
N_fertiliser type_2	Name(s) of the second N fertiliser applied to the crop with its level of N application rate (kg N ha-1) (e.g. Poultry manure -30-). "NR" if not relevant (if fertilisation is not applied)
N_fertiliser type_3	Name(s) of the third N fertiliser applied to the crop with its level of N application rate (kg N ha-1) (e.g. Poultry manure -30-). "NR" if not relevant (if fertilisation is not applied)
N_nb_application	Nr. of applications of N fertilisers. "NR" if not relevant (if fertilisation is not applied)
N_perc_from_organic_fert	% of total N supplied to the crop coming from organic fertilisers or amendments. "NR" if not relevant (if fertilisation is not applied)
P_rate	Total amount of P (kg ha^{-1}) supplied to the crop
P_fertiliser_type_1	Name(s) of the first P fertiliser applied to the crop with its level of P application rate (kg P ha-1) (e.g. Poultry manure -30-). "NR" if not relevant (if fertilisation is not applied)
P_fertiliser_type_2	Name(s) of the second P fertiliser applied to the crop with its level of P application rate (kg P ha-1) (e.g. Poultry manure -30-). "NR" if not relevant (if fertilisation is not applied)

P_nb_application	Nr. of applications of P fertilisers. "NR" if not relevant (if fertilisation is not applied)
P_perc_from_organic_fert	% of total P supplied to the crop coming from organic fertilisers or amendments. "NR" if not relevant (if fertilisation is not applied)
K_rate	Total amount of K (kg ha^{-1}) supplied to the crop
K_fertiliser_type_1	Name(s) of the first K fertiliser applied to the crop with its level of K application rate (kg K ha-1) (e.g. Poultry manure -30-). "NR" if not relevant (if fertilisation is not applied)
K_fertiliser_type_2	Name(s) of the second K fertiliser applied to the crop with its level of K application rate (kg K ha-1) (e.g. Poultry manure -30-). "NR" if not relevant (if fertilisation is not applied)
K_nb_application	Nr. of applications of K fertilisers. "NR" if not relevant (if fertilisation is not applied)
K_perc_from_organic_fert	% of total K supplied to the crop coming from organic fertilisers or amendments. "NR" if not relevant (if fertilisation is not applied)
irrigation	Y/N_partial/full (Y if irrigation was applied or N if not; PARTIAL if irrigation did not cover the full water need of the crop or FULL if it did)
irrigation_quantity	Mean amount of irrigation water (mm) applied (exact amount or MIN-MAX value if a range is reported). "NA" if not available. "NR" if not relevant (if irrigation is not applied)
herbicide_application	Were chemical herbicides applied or not (Y/N)
mechanical_weed_control	Was mechanical weeding applied or not (Y/N)
crop_protection	Were crop protection products, including natural or biocontrol agents, applied to the crop (Y/N)
replicate_nb	Number of replicates concurring to the mean yield value reported in a single site x year combination (e.g. number of blocks or spatial replicates)
site_nb	Number of different sites considered as spatial replicates for computing the mean yield value reported, if mean yield values for each site are not available
year_nb	Number of years concurring to the mean yield value reported, if single year mean yield values are not available
moisture_at_harvest	Moisture percentage of the marketable yield (e.g. "13" for 13%) as reported in the source material
yield	Yield of the grain of the legume crop in t d.m. ha^{-1} (the humidity reported in the previous column, when available, is removed from the grain yield reported)
yield_se	Value of the standard error of the mean of the yield, if available. If not available, "NA"
yield_sd	Value of the standard deviation of the mean of the yield, if available. If not available, "NA"
yield_cv	Value of the coefficient of variation of the mean of the yield, if available. If not available, "NA"
yield_var	Value of the variance of the mean of the yield, if available. If not available, "NA"

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

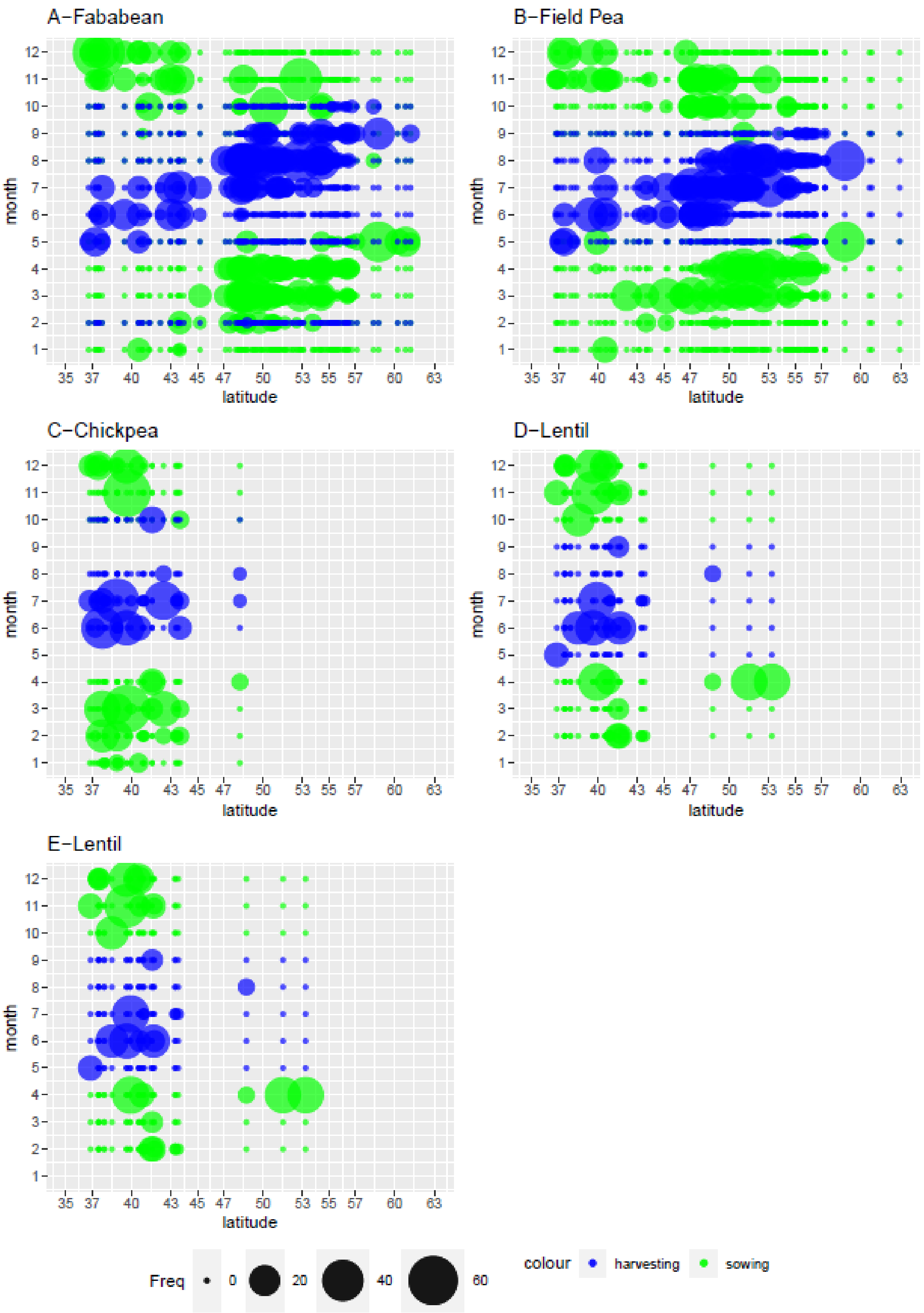


Figure 29. Frequency of sowing and harvesting months based on the latitude of the experimentations for each pulse. Each point represents experimentations with the same sowing or harvesting months. The size of the point increase with the number of experimentations sown or harvested on the same month. The green points are the sowing months and the blue points, the harvesting months. The experiments are conducted between a latitude from 36.73 to 62.94.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

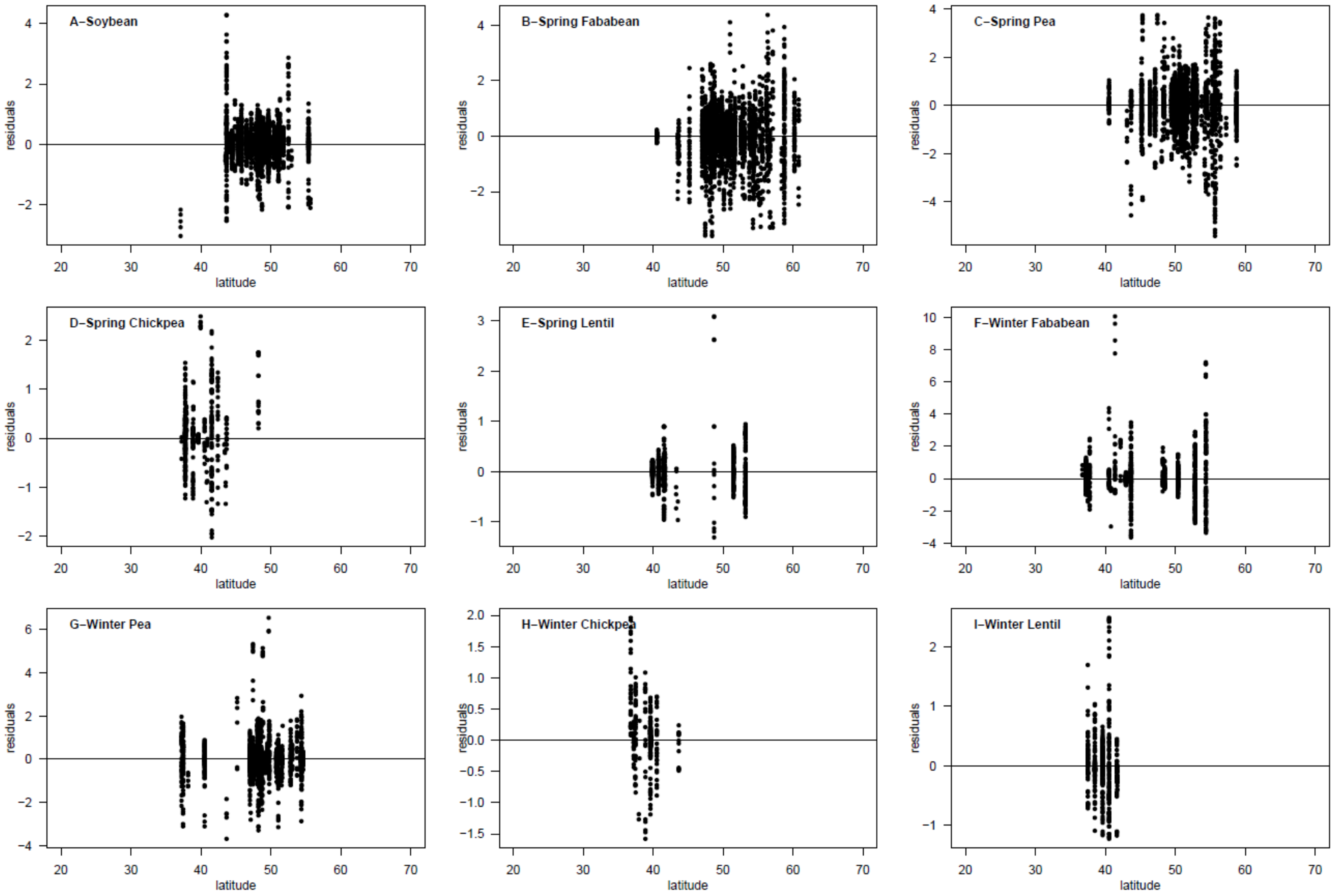


Figure 30. Analysis of model residuals: residuals as a function of latitude.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

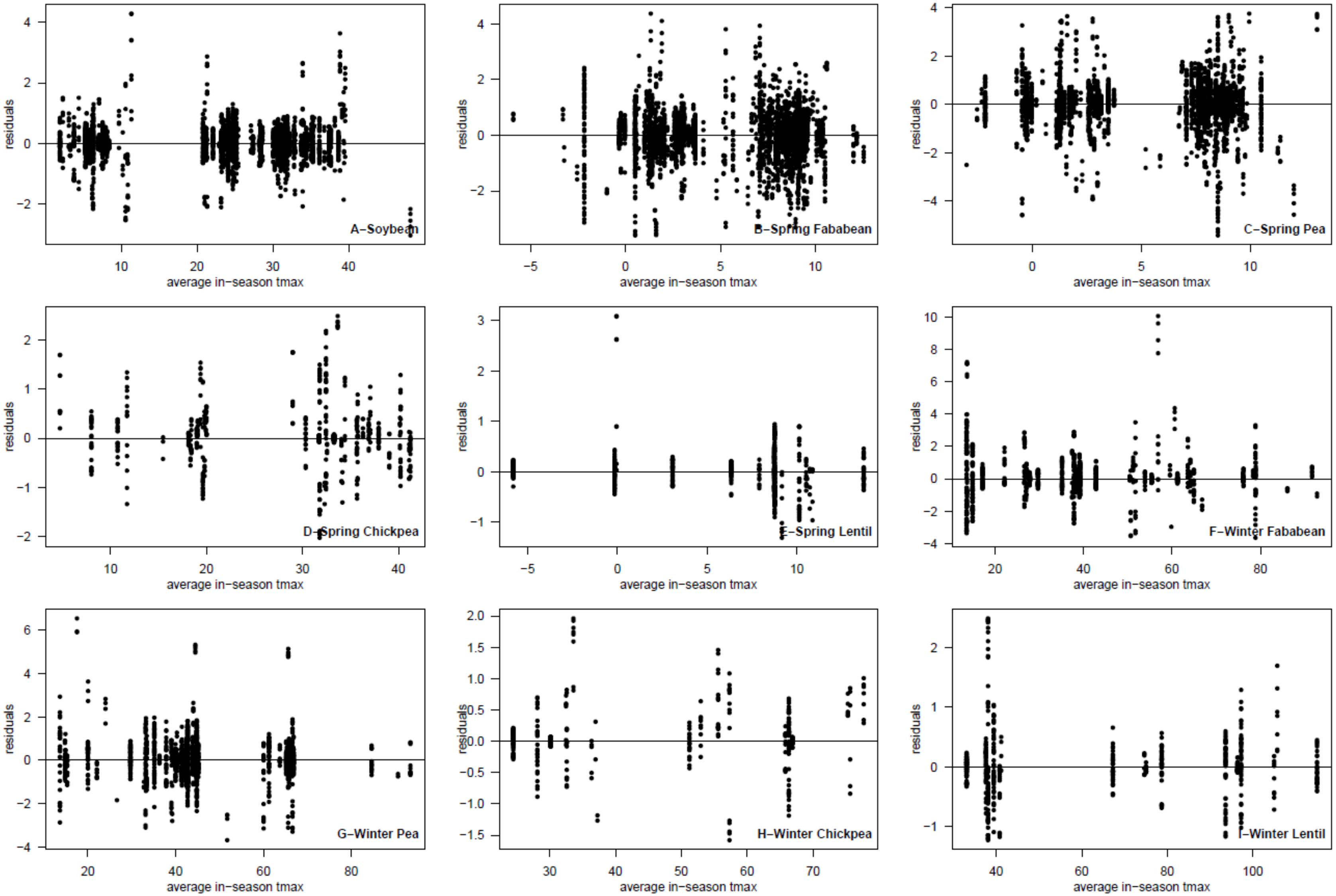


Figure 31. Analysis of model residuals: residuals as a function of average in-season tmax.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

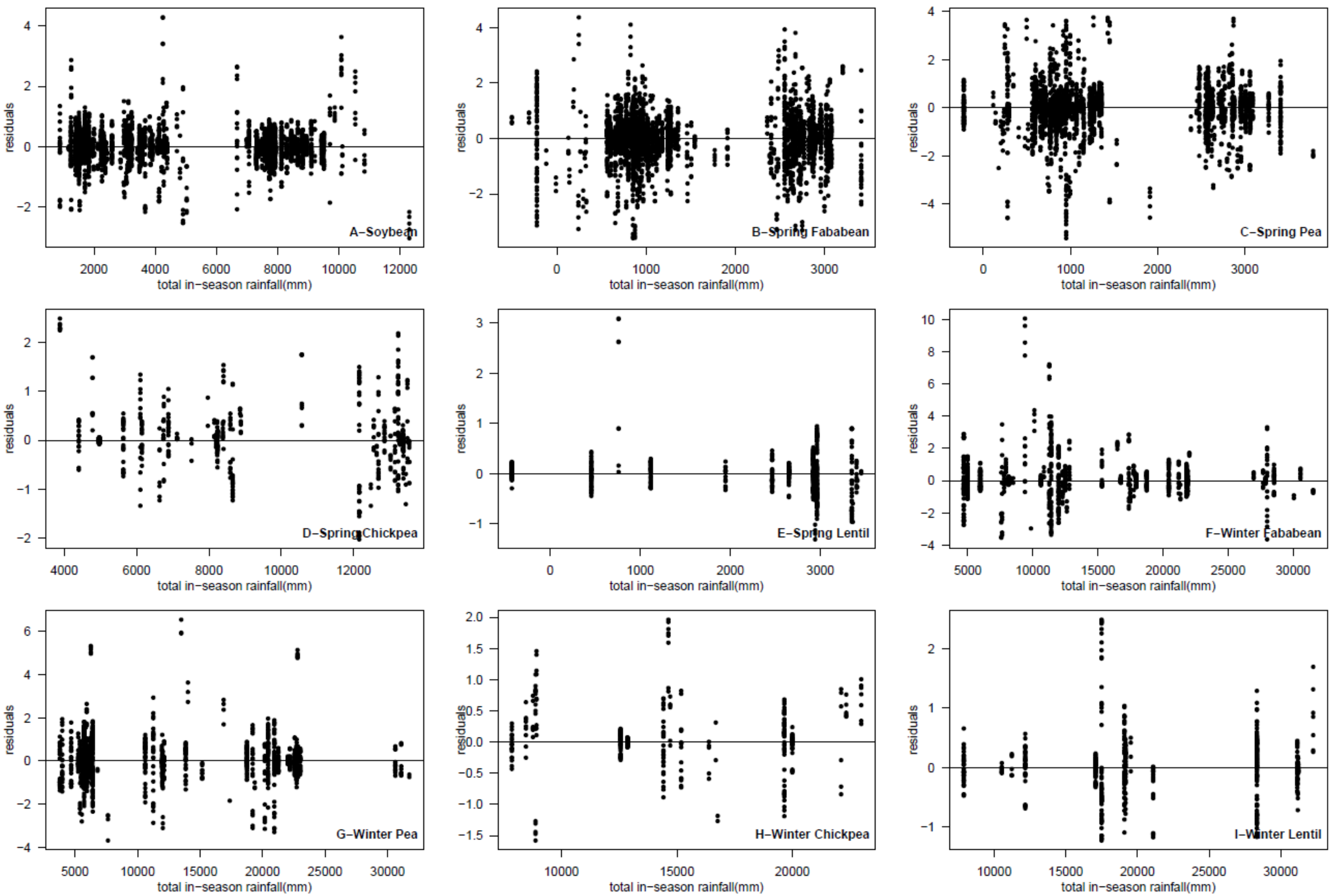


Figure 32. Analysis of model residuals: residuals as a function of total in-season rainfall.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

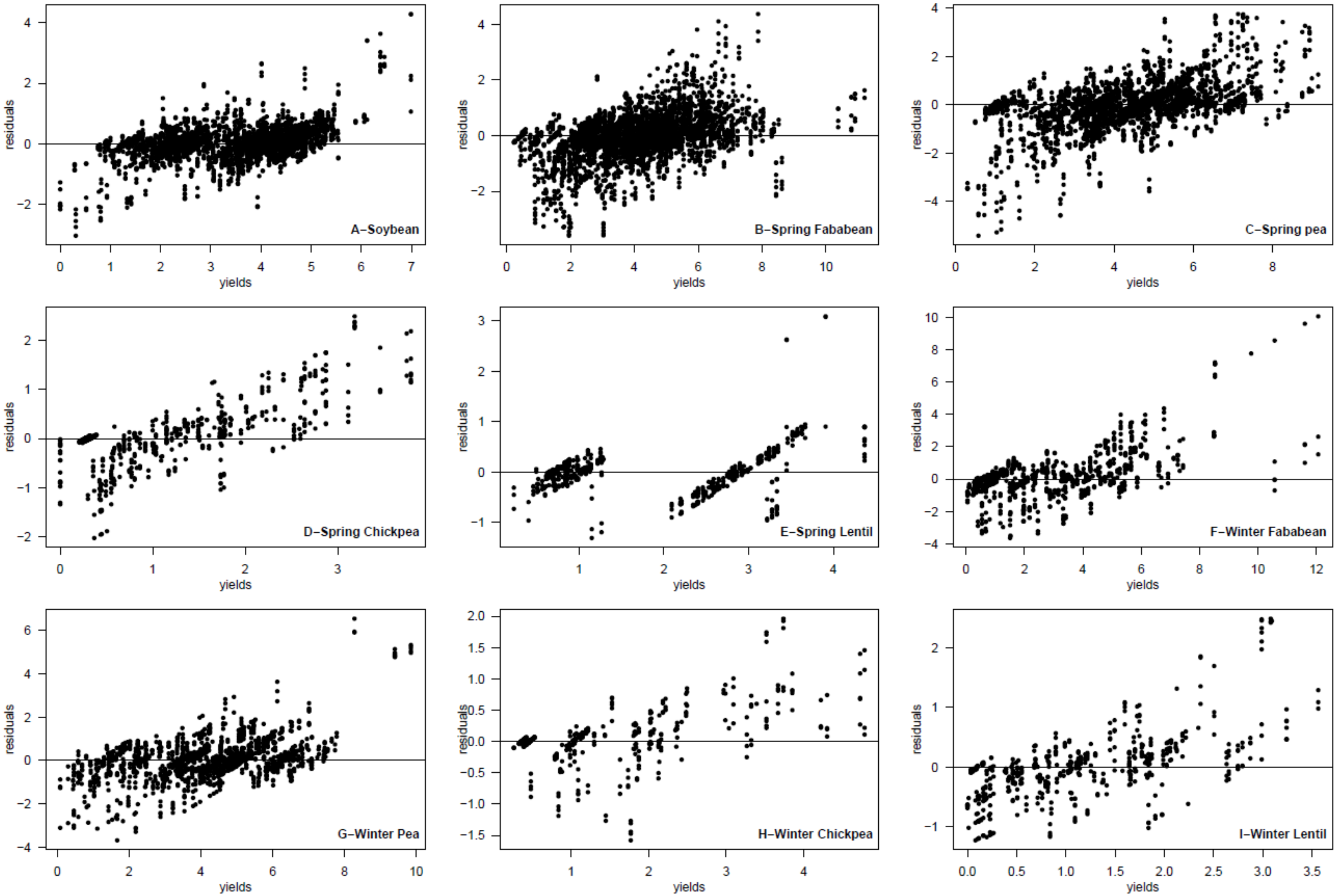


Figure 33. Analysis of model residuals: residuals as a function of observed yields

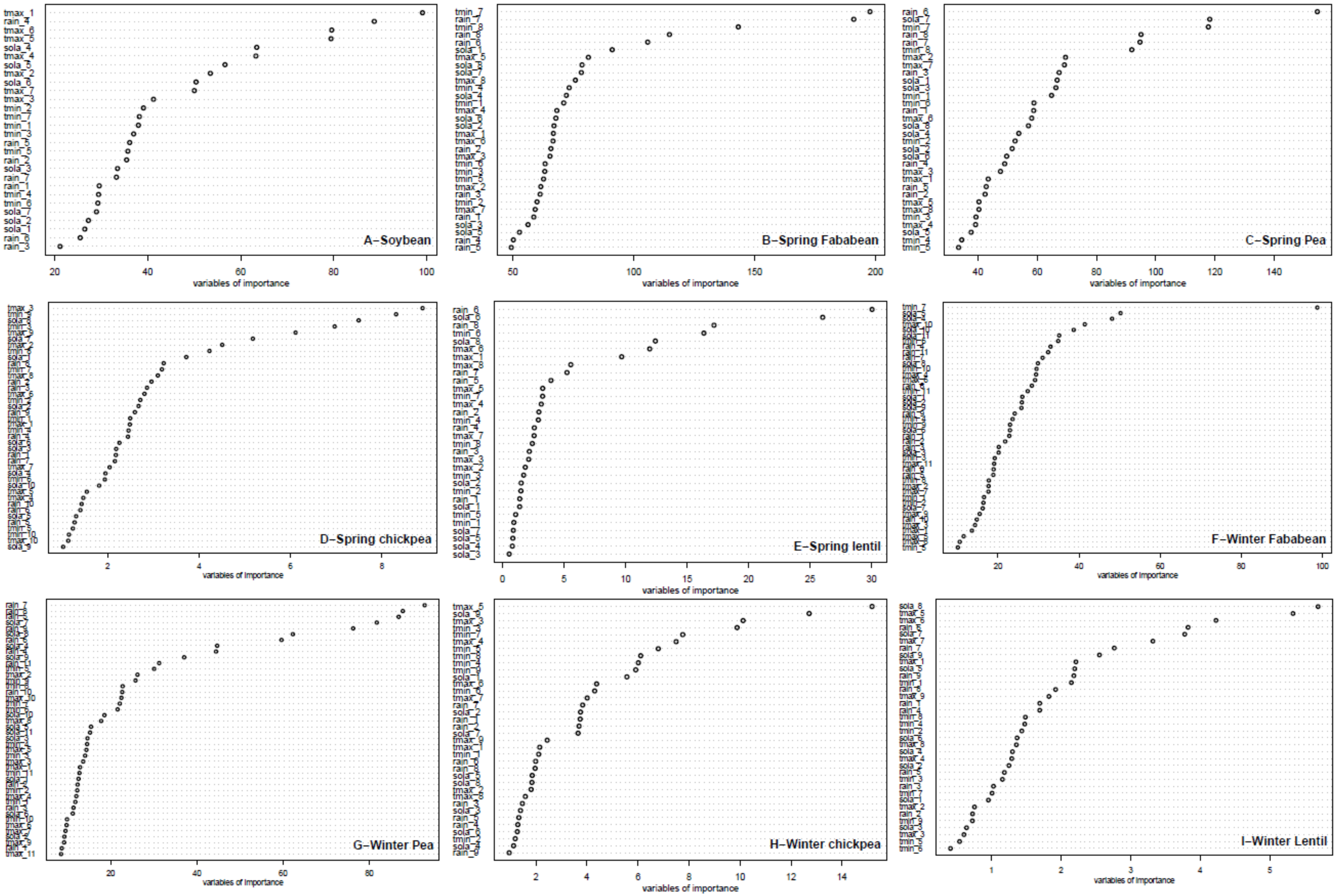


Figure 34. Variables importance plots derived from the Random Forest algorithm. *tmin* (°C) is the monthly average of daily minimum temperature, *tmax* (°C) is the monthly average of daily maximum temperature, *rain* (mm day^{-1}) is the monthly average of daily rainfall, and *solar* (W m^{-2}) is the monthly average of daily downward shortwave

radiation. The number indicated as a suffix indicates the month of the growing season, so that tmin_2 is the average daily minimum temperature in the 2nd month of the growing season.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

A – Soybean

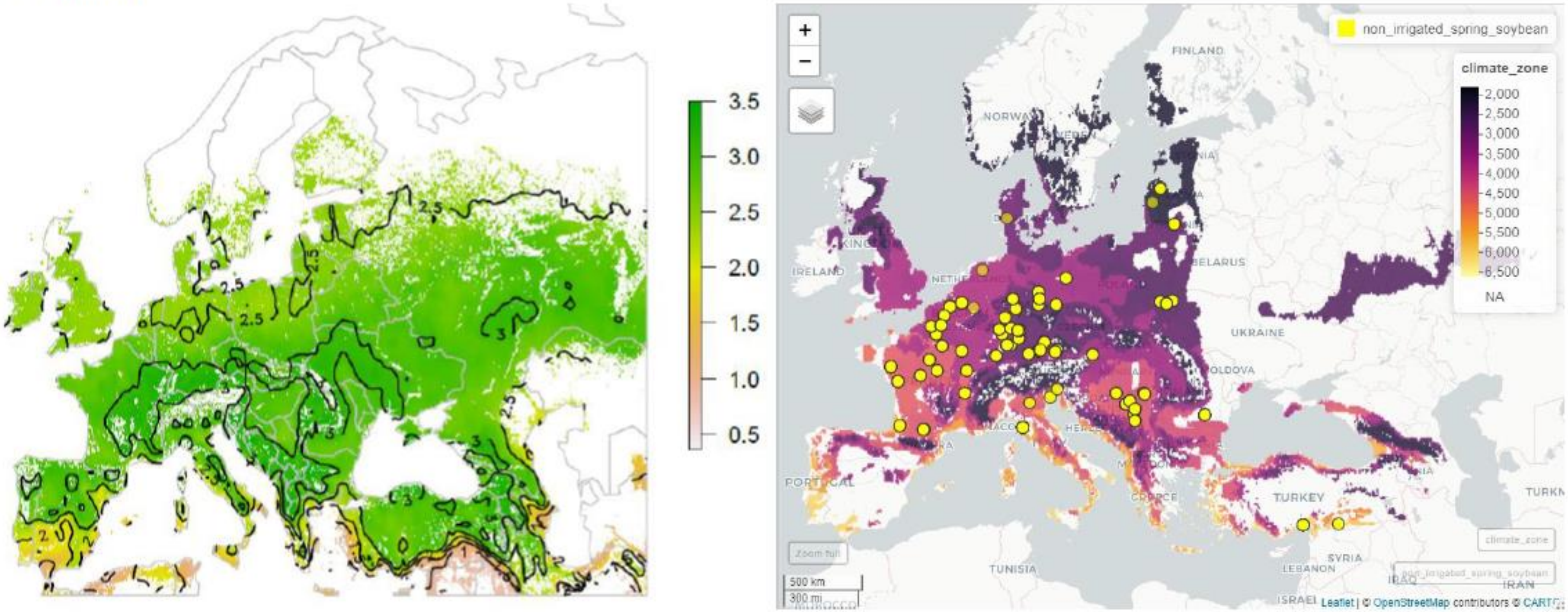


Figure 35. Soybean projected yields (t ha^{-1} at 13% moisture) under historical climate (2000-2020) and maps of locations of experiments used for model fitting (yellow dots) and climate zones containing at least one experiment (filled polygons). The first map shows the median projected yield over years. Projections are shown only on agricultural area (cropland plus pastures) in the year 2000. Climate zones are the Global Yield Gap Atlas Extrapolation Domain (GYGA-ED) available at www.yieldgap.org .

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement N°727672

B – Spring faba bean

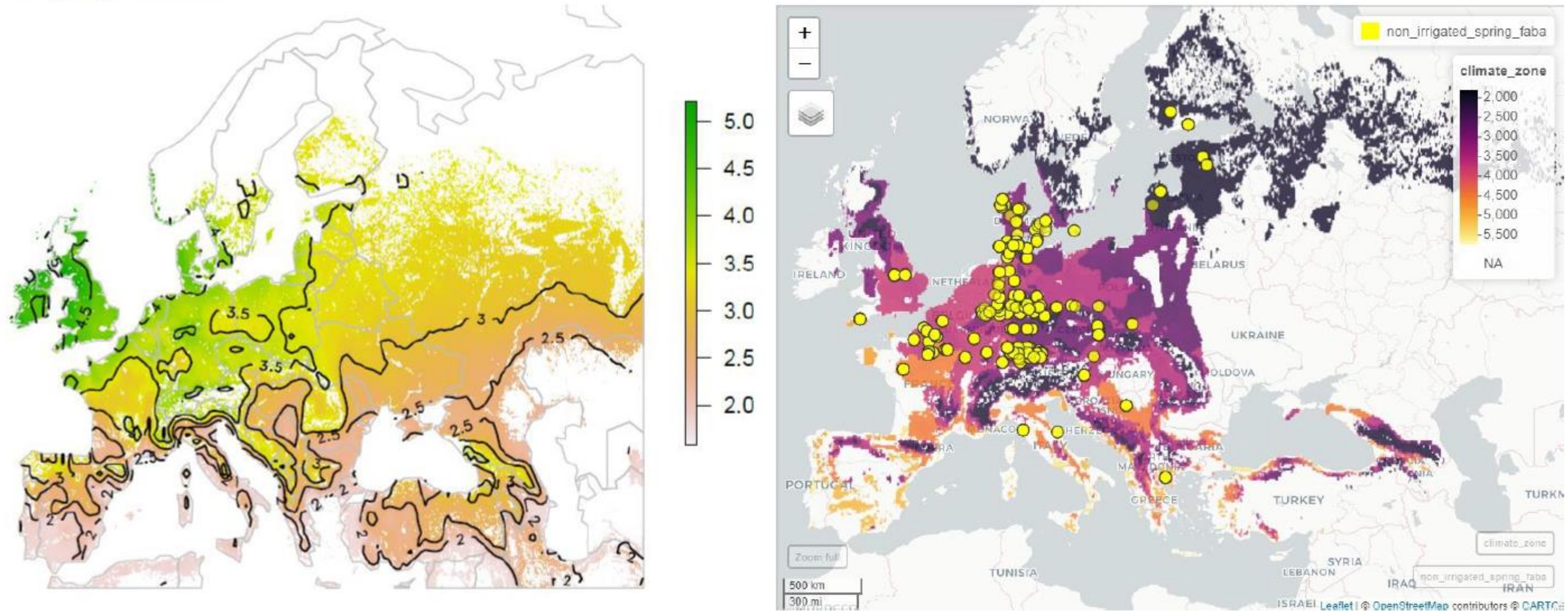


Figure 36. Spring faba bean projected yields (t ha^{-1} at 13% moisture) under historical climate (2000-2020) and maps of locations of experiments used for model fitting (yellow dots) and climate zones containing at least one experiment (filled polygons). The first map shows the median projected yield over years. Projections are shown only on agricultural area (cropland plus pastures) in the year 2000. Climate zones are the Global Yield Gap Atlas Extrapolation Domain (GYGA-ED) available at www.yieldgap.org .

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

C – Spring pea

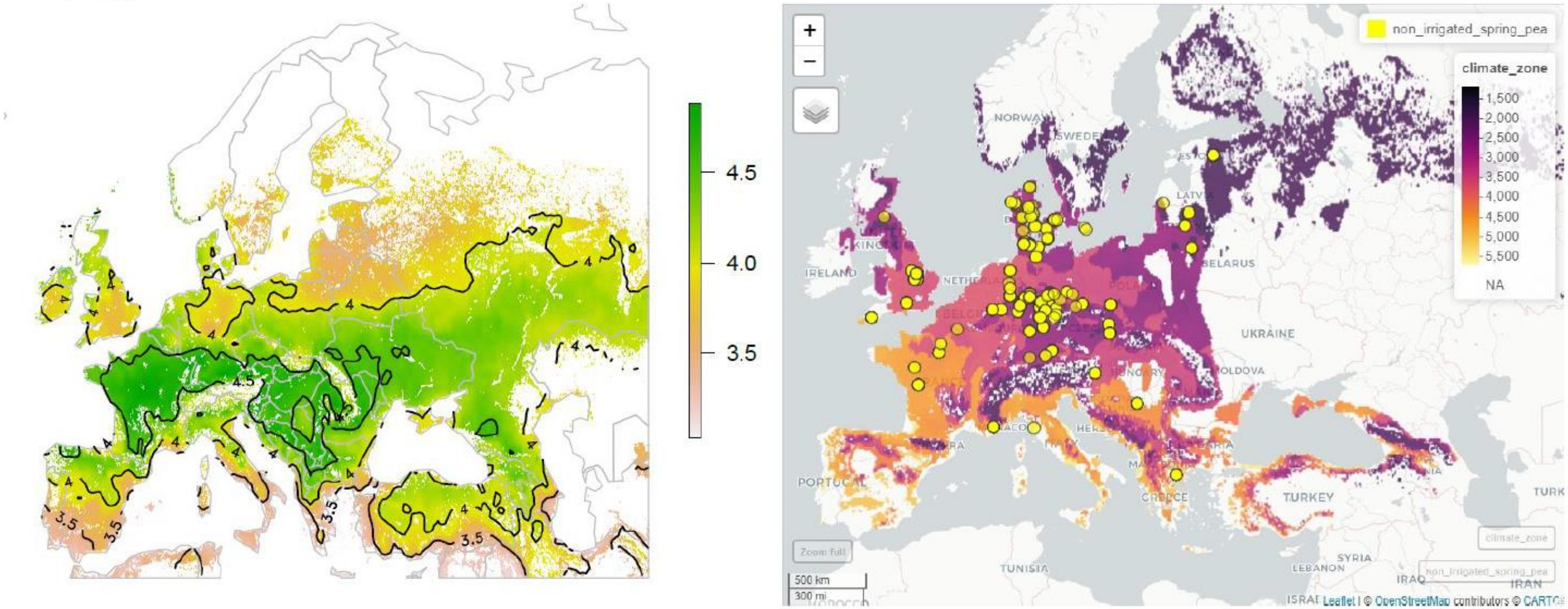


Figure 37. Spring field pea projected yields (t ha^{-1} at 13% moisture) under historical climate (2000-2020) and maps of locations of experiments used for model fitting (yellow dots) and climate zones containing at least one experiment (filled polygons). The first map shows the median projected yield over years. Projections are shown only on agricultural area (cropland plus pastures) in the year 2000. Climate zones are the Global Yield Gap Atlas Extrapolation Domain (GYGA-ED) available at www.yieldgap.org .

D – Winter pea

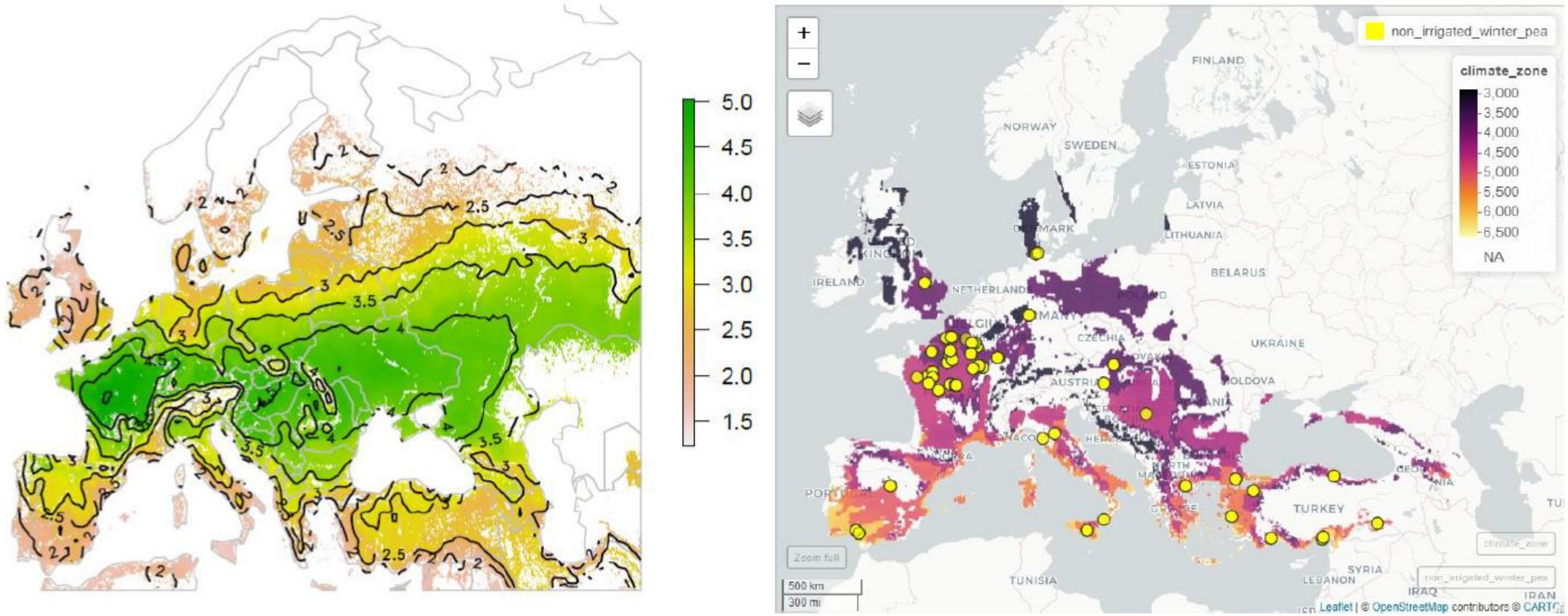


Figure 38. Winter field pea projected yields (t ha^{-1} at 13% moisture) under historical climate (2000-2020) and maps of locations of experiments used for model fitting (yellow dots) and climate zones containing at least one experiment (filled polygons). The first map shows the median projected yield over years. Projections are shown only on agricultural area (cropland plus pastures) in the year 2000. Climate zones are the Global Yield Gap Atlas Extrapolation Domain (GYGA-ED) available at www.yieldgap.org .

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

E – Winter faba bean

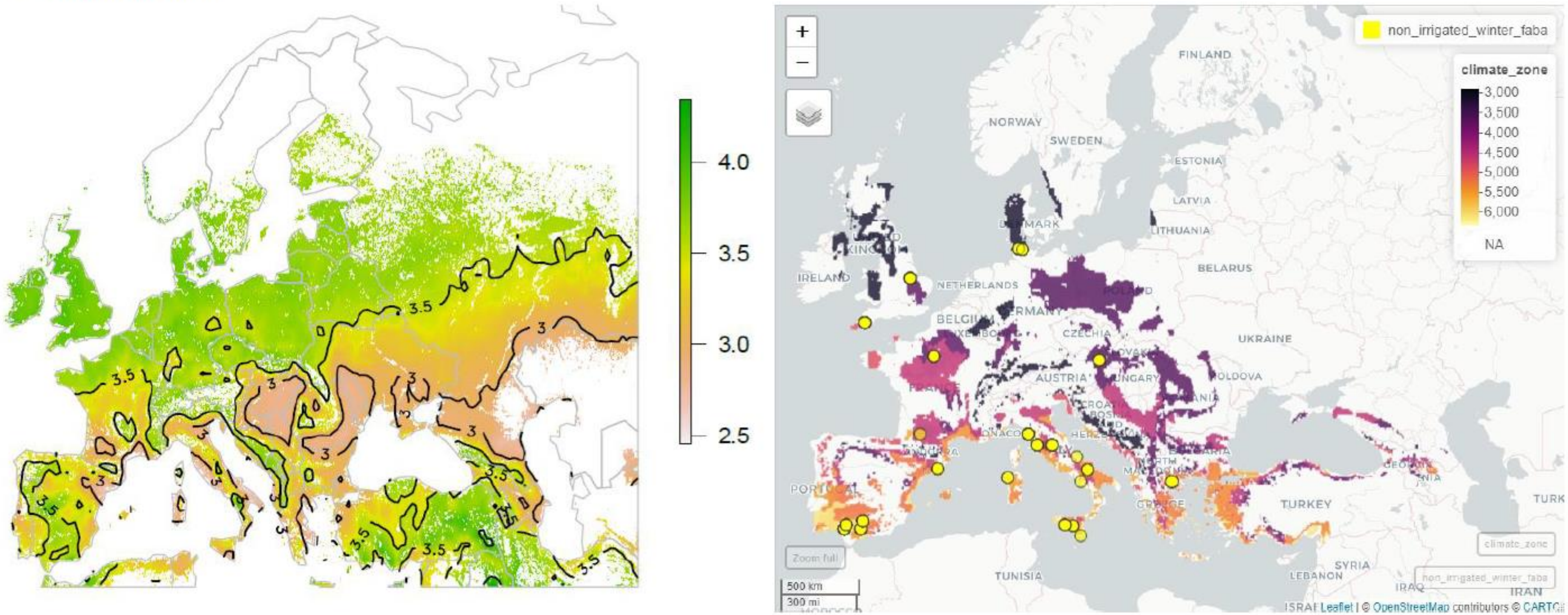


Figure 39. Winter faba bean projected yields (t ha^{-1} at 13% moisture) under historical climate (2000-2020) and maps of locations of experiments used for model fitting (yellow dots) and climate zones containing at least one experiment (filled polygons). The first map shows the median projected yield over years. Projections are shown only on agricultural area (cropland plus pastures) in the year 2000. Climate zones are the Global Yield Gap Atlas Extrapolation Domain (GYGA-ED) available at www.yieldgap.org .

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

F – Spring lentil

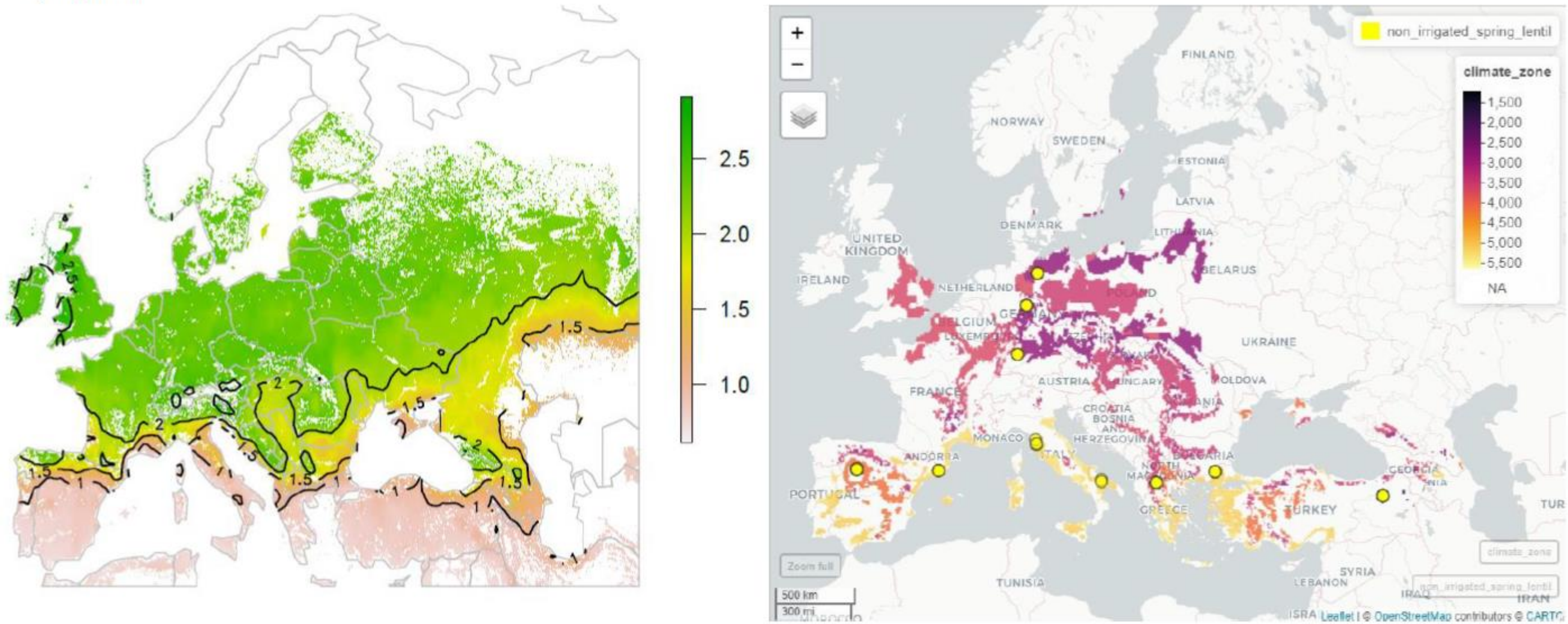


Figure 40. Spring lentil projected yields (t ha^{-1} at 13% moisture) under historical climate (2000-2020) and maps of locations of experiments used for model fitting (yellow dots) and climate zones containing at least one experiment (filled polygons). The first map shows the median projected yield over years. Projections are shown only on agricultural area (cropland plus pastures) in the year 2000. Climate zones are the Global Yield Gap Atlas Extrapolation Domain (GYGA-ED) available at www.yieldgap.org .

G – Winter chickpea

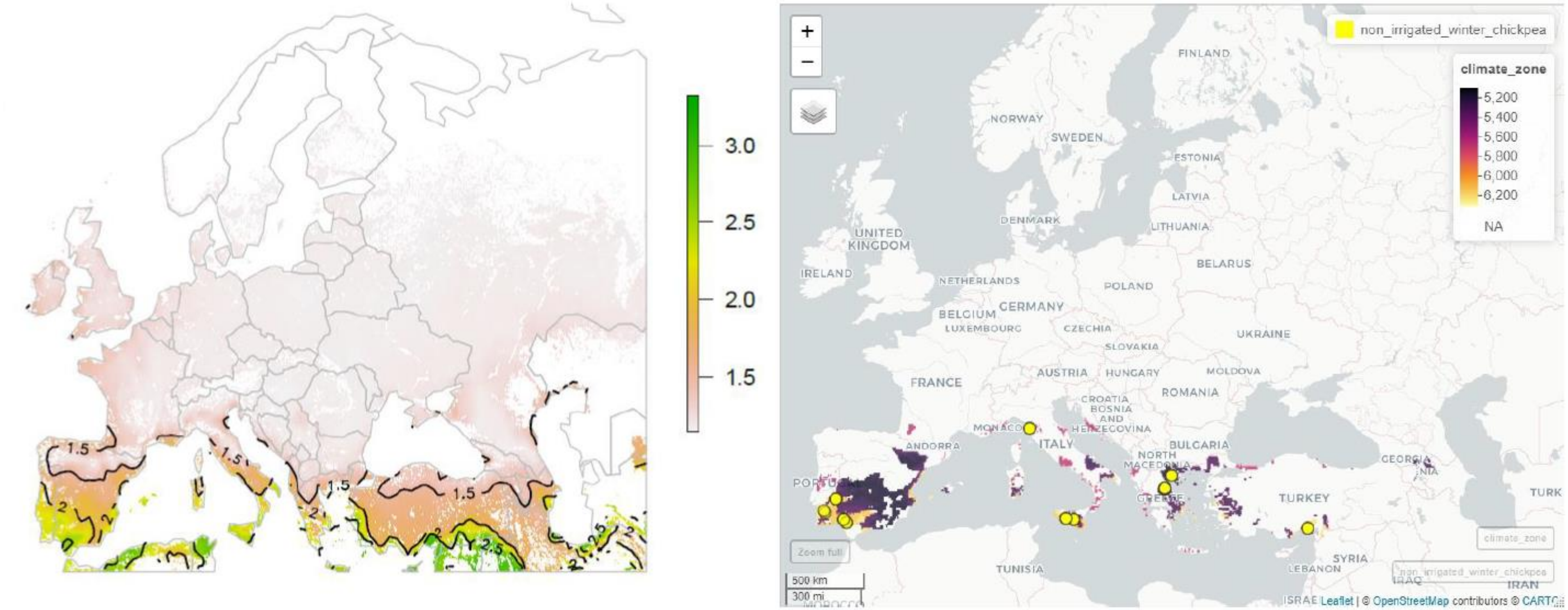


Figure 41. Winter chickpea projected yields (t ha^{-1} at 13% moisture) under historical climate (2000-2020) and maps of locations of experiments used for model fitting (yellow dots) and climate zones containing at least one experiment (filled polygons). The first map shows the median projected yield over years. Projections are shown only on agricultural area (cropland plus pastures) in the year 2000. Climate zones are the Global Yield Gap Atlas Extrapolation Domain (GYGA-ED) available at www.yieldgap.org .

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement N°727672

H – Winter lentil

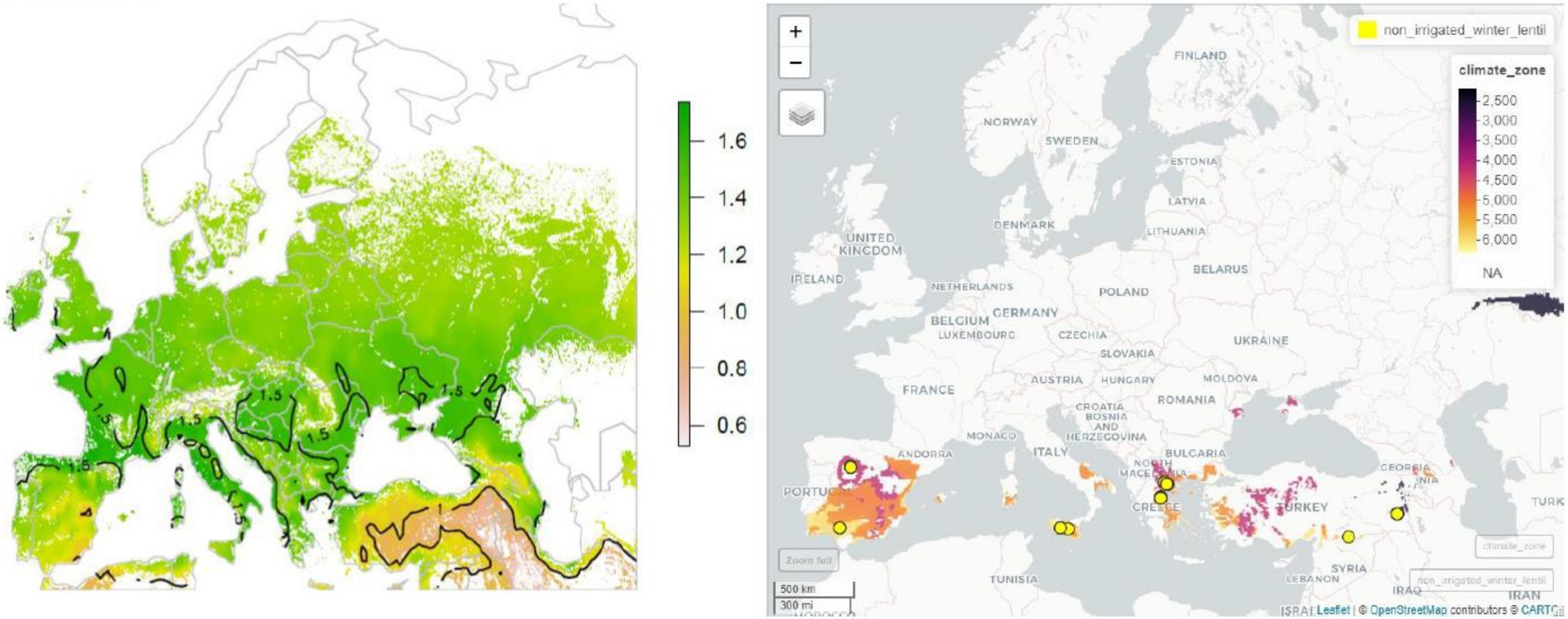


Figure 42. Winter lentil projected yields (t ha^{-1} at 13% moisture) under historical climate (2000-2020) and maps of locations of experiments used for model fitting (yellow dots) and climate zones containing at least one experiment (filled polygons). The first map shows the median projected yield over years. Projections are shown only on agricultural area (cropland plus pastures) in the year 2000. Climate zones are the Global Yield Gap Atlas Extrapolation Domain (GYGA-ED) available at www.yieldgap.org .

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

I – Spring chickpea

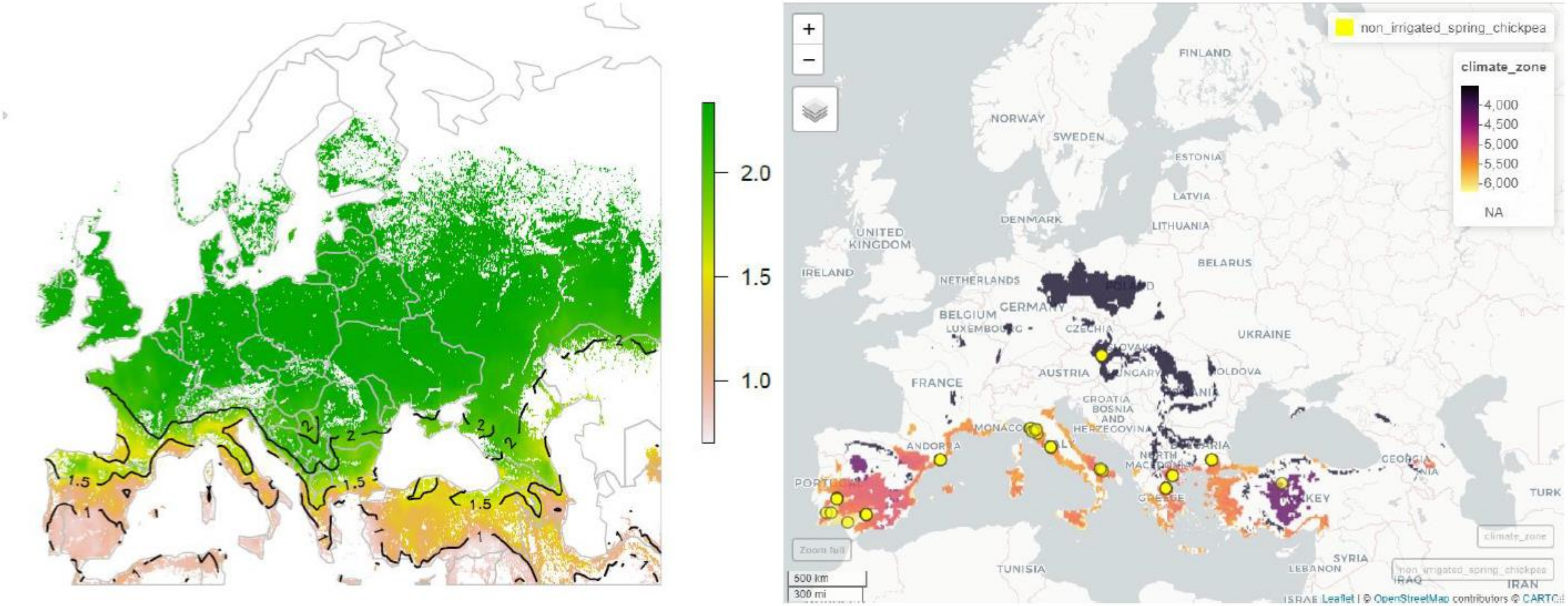


Figure 43. Spring chickpea projected yields (t ha^{-1} at 13% moisture) under historical climate (2000-2020) and maps of locations of experiments used for model fitting (yellow dots) and climate zones containing at least one experiment (filled polygons). The first map shows the median projected yield over years. Projections are shown only on agricultural area (cropland plus pastures) in the year 2000. Climate zones are the Global Yield Gap Atlas Extrapolation Domain (GYGA-ED) available at www.yieldgap.org .

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672