LEGVALUE

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 (Part B)

Planned delivery date: M36

Actual submission date: M43

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

Workpackage: WP1

Workpackage leader: INRAE Deliverable leader: INRAE

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

Version: V1

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

Table of contents

Authors: Nicolas Guilpart (AgroParisTech, nicolas.guilpart@agroparistech.fr), Iris Bertin (INRAE), Daniele Antichi (UNIPI), Véronique Biarnès (Terres Inovia), Rémy Ballot (INRAE), Paul Belleville (INRAE), Marie-Hélène Jeuffroy (INRAE)

Date : June 2021

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 (main text)

Tables (Appendix)

Figures (Main text)

Figures (Appendix)

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 remains 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 applied over the whole EU agricultural area to make yield projections under historical climate (2000-2020), for 5 grain legumes: soybean, field pea, fababean, 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 6488 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 5 climate variables (minimum and maximum temperatures, rainfall, solar radiation, and reference evapotranspiration) 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, fababean, chickpea and lentil. The model was fitted using a Random Forest algorithm.

The overall predictive ability of the model is good, with an R^2 of 0.85 between observed and predicted yields across all crops based on cross-validation. At the individual crop level, the predictive ability of the model is very good for winter chickpea, spring lentil and soybean ($R^2 \geq 0.85$), good for field pea and fababean (spring and winter) ($R^2 \geq 0.75$), and medium for winter lentil and spring chickpea ($R^2 \geq 0.60$). 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 climatic suitability for most crops, as projections reveal large areas where projected yields are higher than the actual average national yield based on official statistics. High-yielding areas in Europe are identified for each crop.

Future work will 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 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 Europe.

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 achievable yields 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 intended to produce EU achievable yield maps for these five major pulse species according to current soil and climate conditions with the following objectives: (i) allowing more precise and realistic estimations of protein and starch yield from pulses in current and future scenarios of legume presence in cropping systems; (ii) identifying research gaps and technological lock-ins currently hampering a higher share of pulses in cropping systems, (iii) identifying sites with high productive potential for pulses that are currently unexplored in the EU, (iv) setting 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 field trials performed in 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 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 model quality, and (iv) projections of grain legume yield over the whole European agricultural area based on the fitted model.

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

4 Materials and methods

4.1 Data sources

4.1.1 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 6488 yield data collected from published and non-published experimentations all over Europe for 5 different pulses: soybean (*n*=1577), faba bean (*n*=1653), field pea (*n*=2423), 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	2423	451	384	6488
Irrigated	464	171	111	35	10	791
Rainfed	650	793	805	251	220	2719
NA	463	689	1507	165	154	2978
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	3.78	1.14	1.14	3.06
Mean	2.95	3.32	3.75	1.40	1.38	3.04

4.1.2 Historical climate data and climate zones

Historical climate data. 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 several climatic variables relevant to crop growth and yield: maximum (*tmax,* degree Celsius) and minimum (*tmin,* degree Celsius) air temperatures at 2m, total precipitation (*rain,* mm day^{-1}), mean downward shortwave radiation flux (*solar,* W m^{-2}), mean relative humidity at 2m (*RH*, %), mean wind speed at 10m (*wind*, m s^{-1}). These variables are available for the period 1958–2020 at a spatial resolution of 0.5 degree. Five variables were selected to model crop yield: *tmin*, *tmax*, *rain*, *solar*, and *ETo* (reference evapotranspiration). Daily *ETo* was not available in the JRA55-CDFDM-S14FD dataset, so it was calculated using available variables. To this aim, clear

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

sky transmissivity (*cst*) was calculated as a function of solar radiation, date, and latitude using the *cst()* function of the *sirad* R package; and daily actual vapor pressure (*ea*, kPa) was calculated using Equation 1 from relative humidity and *tmin* and *tmax* following FAO recommendations (Allen et al., 1998).

Equation 1

$$ea = \frac{RH}{100} \times \frac{0.6108 \times e^{\left(\frac{17.27 \times tmax}{tmax+237.3}\right)} + 0.6108 \times e^{\left(\frac{17.27 \times tmin}{tmin+237.3}\right)}}{2}$$

Then, daily total reference evapotranspiration (*ETo*, mm) was calculated as a function of *tmax*, *tmin*, *solar*, *ea*, *wind*, *cst,* latitude, and elevation using the *et0()* function of the *sirad* R package, which uses the FAO Penman-Monteith evapotranspiration equation (Allen et al., 1998). Finally, monthly averages of *tmin*, *tmax*, *solar*, *rain*, and *ETo* 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.

Climate zones. 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).

4.1.3 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/.

4.2 Model fitting and projections

4.2.1 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" in the referred column of the dataset) 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 underlined 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 4960 yield data, including 951 for soybean, 319 for chickpea, 290 for lentil, 1334 for faba bean, and 2066 for field pea.

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 1.

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

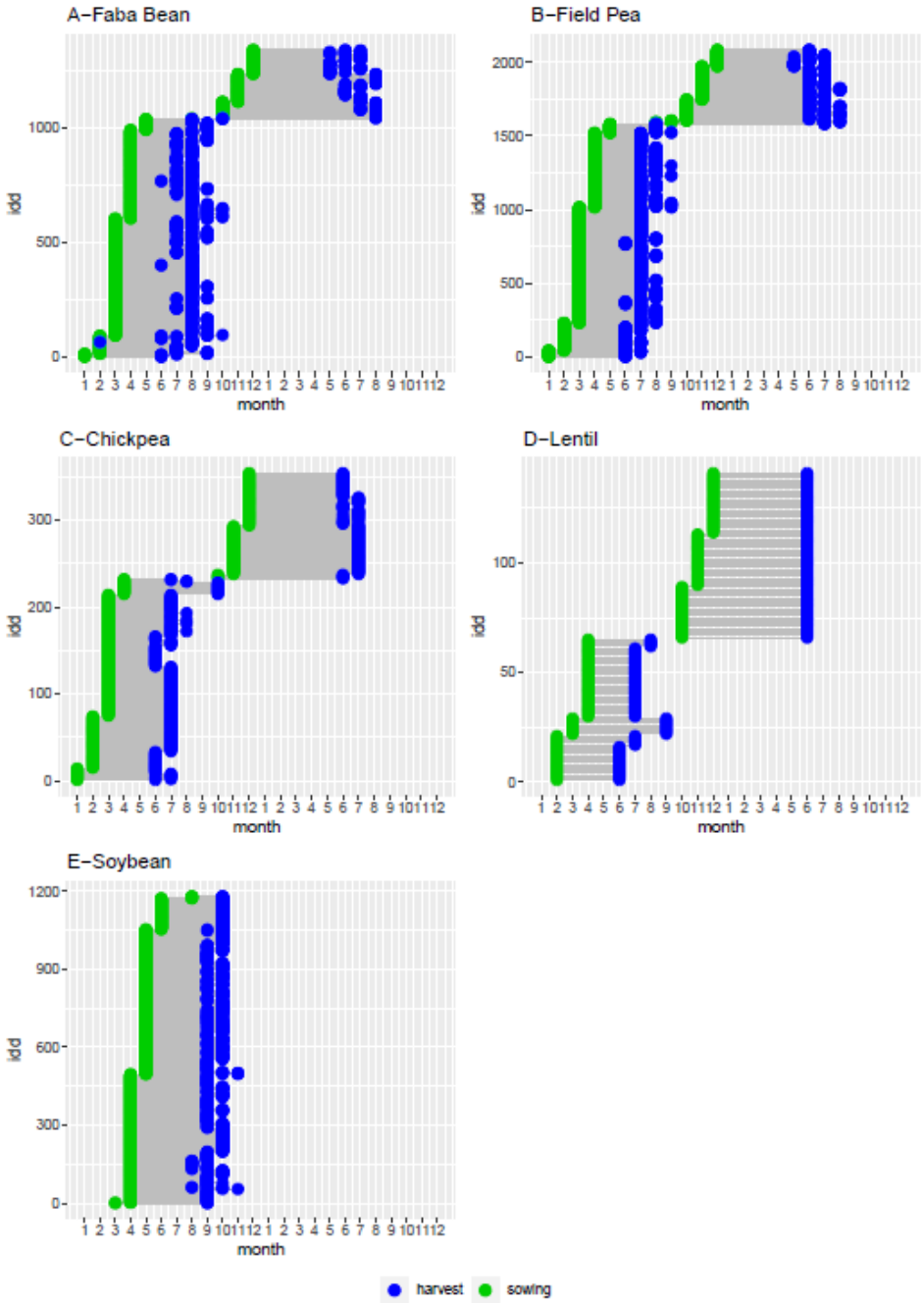

Figure 1. Distribution of sowing and harvest dates by crop in the European Grain Legumes Dataset. Each line is an experimental unit, i.e. a unique combination of experiment, treatment and year. Faba bean (A), field pea (B), chickpea (C), and lentil (D) can be sown as winter or spring crops, whereas soybean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

Growing season definition. The distribution of observed sowing and harvest months of selected data is shown in Figure 1. 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). Sowing and harvest dates change with latitude (see Figure 7 in appendix), but to keep the model as simple as possible, fixed growing seasons were defined. The defined growing seasons were designed to include earliest sowing dates and latest harvest dates observed Figure 1. 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.

4.2.2 Model fitting and evaluation

Crop yield data were related to the five considered climate variables defined at a monthly time step over the months of the growing season as described in Equation 2, 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, *solar* (W m^{-2}) is the monthly average of daily downward shortwave radiation, and *ETo* is the monthly average of daily reference evapotranspiration (mm day^{-1}). 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. 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 2 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 2

$$yield \sim tmin.1 + tmin.2 + \cdots + tmin.n + tmax.1 + tmax.2 + \cdots + tmax.n + rain.1 + rain.2 + \cdots + rain.n + solar.1 + solar.2 + \cdots + solar.n + ETo.1 + ETo.2 + \cdots + ETo.n$$

The model described in Equation 2 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, and the root mean square error of prediction (RMSEP, t ha^{-1}). 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.

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

4.2.3 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 average 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 climatic conditions captured in the training dataset of our model, we retrieved minimum and maximum values of each climatic variable as observed in the training dataset (see examples in Figure 14 to Figure 18 for spring crops in appendix). Then, every single value of every climatic variable in the projection dataset (i.e. whole Europe) was classified as *in-range* of training data (i.e. higher or equal than minimum and lower or equal than maximum) or *out-of-range* of training data (i.e. lower than minimum or higher than maximum). This was done on a pixel basis, for every crop and every year from 2000 to 2020, taking into account crop-specific growing seasons as defined in Table 2. Then the frequency of *out-of-range* events over all years and climatic variables was calculated over the 2000-2020 period and mapped (see Figure 19 and Figure 20 in appendix). This allowed to identify areas in Europe where climate is frequently *out-of-range* of training data. Then, yield projections were shown only in areas where the frequency of *out-of-range* events did not exceed 20% for soybean, pea, and fababean, and 40% for chickpea and lentils. This procedure allowed to ensure that the combination of environmental conditions under which the model was calibrated are similar to the environmental conditions to which the model is projected, while accepting a reasonable degree of extrapolation. We also used the GYGA-ED climate zonation scheme (van Wart et al., 2013) to map, for each crop, all climate zones containing at least one experiment of the European Grain Legumes Dataset.

5 Results and discussion

5.1 Assessment of model performances

Comparison of observed and predicted yields. The predictive ability of the model is good, with an overall R^2 of 0.85 between observed and predicted yields across all crops based on cross-validation (Figure 2). The comparison of observed and predicted yields 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. At the individual crop level, R^2 values range from 0.60 (winter lentil) to 0.91 (winter chickpea) (Table 3 and Figure 8 in appendix). Winter chickpea, spring lentil and soybean have $R^2 \geq 0.85$; field pea and fababean (both spring and winter) have R^2 between 0.75 and 0.80; and spring chickpea and winter lentil have R^2 between 0.60 and 0.65. 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 field pea and fababean (spring and winter), (iii) medium for winter lentil and spring chickpea.

Analysis of model residuals. The distribution of model residuals is centered on zero for all crops (see insets in Figure 8 in appendix)). Model residuals show no association with latitude (Figure 9), average in-season *tmax* (Figure 10) and total in-season rainfall (Figure 11) 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 12). The model therefore over-estimate low yields and under-estimate high yields. This

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

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 and Root Mean Square Error of Prediction (RMSEP, in t ha^{-1}) are calculated. 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})	R^2 (no unit)	RMSEP (t ha^{-1})
Winter chickpea	130	1.33	0.91	0.38 (29%)**
Spring lentil	149	1.87	0.91	0.41 (22%)
Soybean	951	3.29	0.85	0.50 (15%)
Winter field pea	454	4.33	0.80	0.85 (20%)
Spring faba bean	1058	4.04	0.78	0.83 (21%)
Winter faba bean	276	2.95	0.75	1.26 (43%)
Spring field pea	1612	4.39	0.75	0.71 (16%)
Spring chickpea	189	1.11	0.64	0.59 (53%)
Winter lentil	141	1.17	0.60	0.56 (48%)

** number of observations in training dataset*

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

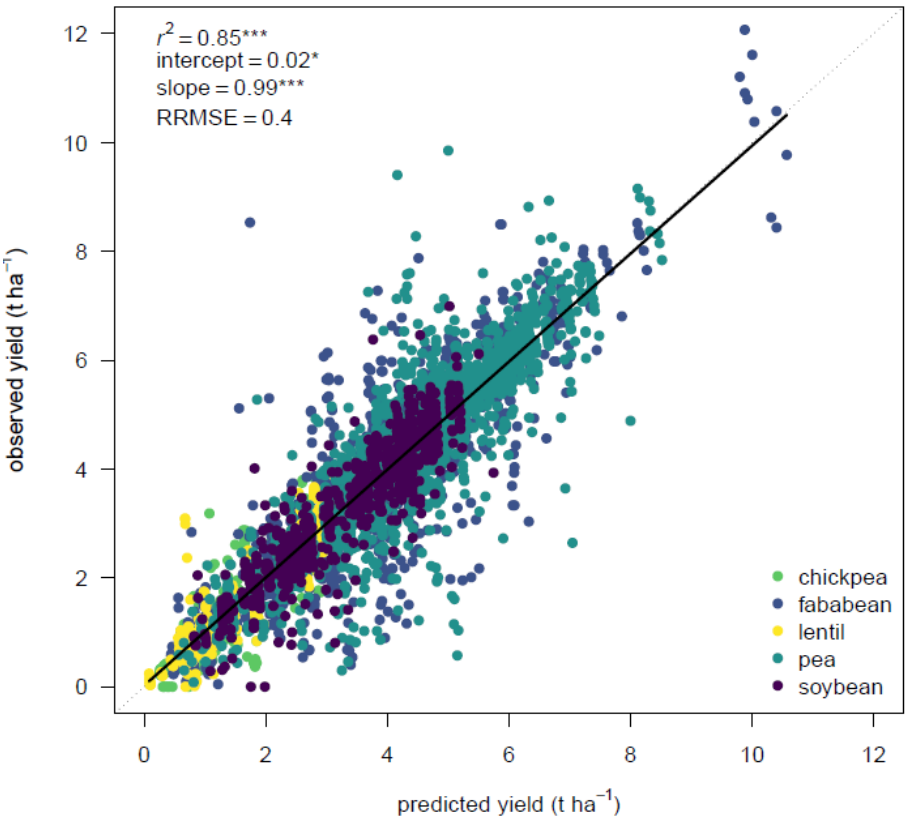

Figure 2. Assessment of the Random Forest algorithm for the different crops considered in this study. For each crop, the model is evaluated using cross-validation (unstratified bootstrap approach with 25 resamplings), and out-of-bag predictions are compared to observed yields. The black line represents the linear regression between observed and predicted yields. Linear regression outputs are shown on the top left of the panel. The dotted line represents the 1:1 line. Each coloured dot shows the comparison of observed and predicted yields for one experimental data of one peculiar crop. Yield values are expressed at a standard moisture content of 13%.

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

Variable importance. Variables importance is presented in Figure 13 in appendix for all crops and all variables. Table 4 gives the 5 most important variables for each crop. These results provide interesting insights about key climatic variables for yield formation of these crops. However, further analysis is required here to compare these results with ecophysiological knowledge on these crops, which is out of the scope of the present report.

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. *ETo* (mm day^{-1}) is the monthly average of daily reference evapotranspiration. 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 5th 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	rain_4	rain_3	tmin_1	tmin_3	tmax_1
Faba bean – spring	8	February – September	ETo_7	solar_11	solar_2	ETo_6	solar_7
Faba bean – winter	11	October – August	tmin_2	tmin_11	tmax_2	solar_8	tmax_8
Field pea – spring	8	February – September	ETo_5	solar_2	rain_4	rain_7	solar_3
Field pea – winter	11	October – August	solar_9	rain_10	solar_11	ETo_9	solar_10
Chickpea – spring	10	January – October	tmin_4	solar_6	ETo_4	tmin_1	tmin_3
Chickpea – winter	9	November – July	ETo_9	solar_9	tmin_3	ETo_5	ETo_1
Lentil – spring	8	February – September	solar_3	solar_5	ETo_6	solar_6	rain_7
Lentil – winter	9	October – June	solar_1	rain_2	tmax_3	ETo_1	tmin_3

** length of the growing season in month*

5.2 Yield projections under historical climate

Assessing where projections are reliable. Yield projections under historical climate (2000-2020) are presented in Figure 3 for spring crops, and in Figure 4 for winter crops. Projections are restricted to areas where climate conditions are similar enough to the climate conditions under which the model was trained (see methods). This approach reveals two groups of crops: (i) a "high coverage" group for which projections could almost be made over the whole European agricultural area, including soybean, fababean (spring and winter), pea (spring and winter), and spring lentil; and (ii) a "low coverage" group for which projections could be made only on a small part of the European agricultural area, including winter lentil and chickpea (spring and winter). These two groups reflect the amount and location data available in the European Grain Legume Dataset to train the model. Indeed, as compared to crops in the low coverage group, crops in the high coverage group have a much larger amount of data available in the European Grain Legume Dataset (see Table 1), and corresponding experiments are located in a much wider variety of places (see Figure 21 to Figure 29 in appendix). Therefore, climatic conditions captured in the model are much wider for crops in the high coverage group. This highlights a need for data collection for crops in the low coverage group to increase the range of climate conditions captured in the model.

Yield projections under historical climate suggest high climatic suitability for pulses in Europe. Yield projections presented in Figure 3 and Figure 4 suggest high climatic suitability for pulses in Europe. Indeed, these projections reveal large areas where projected yields are higher than the actual average national yield (Figure 5, Figure 6, and Table 6). This is especially true for soybean, pea and fababean. This confirms that the current extent of harvested areas for these crops is not limited by climate. This conclusion holds for lentil and chickpea, although to a lower extent because projections for these crops

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

are limited to some parts of Europe to prevent from making extrapolation of the model outside of the climatic conditions captured in the training dataset.

Identification of high-yielding areas. Yield projections presented in Figure 3 and Figure 4 allows to identify high-yielding areas in Europe. For soybean, highest yield areas (≥ 3 t ha^{-1}) are located between the south of France and the south of Belarus and northern Ukraine, including southern Germany, Czech Republic, Poland, Hungary, northern Romania and northern Italy. Spring fababean display a strong north-south gradient with highest yielding areas (≥ 4 t ha^{-1}) in the north, including the UK, Ireland, Belgium, the Netherlands, Denmark, northern Germany, northern Poland and Baltic states. In contrast, high yielding areas of winter fababean are located in Spain, Turkey, and the UK. High-yielding areas for spring pea (≥ 4 t ha^{-1}) are located in the north-west of Europe, including France, Belgium, The Netherlands, Denmark, Germany, the UK and Ireland, while winter pea high-yielding areas (≥ 4.5 t ha^{-1}) are concentrated in western France. High-yield areas for winter chickpea are located in the south of Europe, including southern France, Spain, Italy, Romania, Bulgaria, Greece and Turkey. Spring chickpea high-yielding areas appear to be located a bit more in the north, but with no major differences. Lentil displays contrasted projections for the spring type, with high yielding (≥ 2.5 t ha^{-1}) areas located in the north of Europe, including France, Germany, Belgium, the Netherlands, Denmark, Poland, Belarus, and Baltic States, and winter type with high-yielding areas mostly located in western France. Although the comparison is difficult because of different spatial scales, these general patterns appear consistent with observed actual yields at national levels from official statistics (Figure 5). Across all crops, spring fababean displays the highest projected yield level (5 t ha^{-1}), followed by winter pea (4.5 t ha^{-1}), spring pea (4 t ha^{-1}), winter fababean (3.5 t ha^{-1}), soybean (3 t ha^{-1}), lentils and chickpea.

Other factors that may prevent from reaching the projected yield values. We highlight that maps of projected yields shown in Figure 3 and Figure 4 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 farmers' fields, with crop yields measured in experimental plots being often higher than in farmers' 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

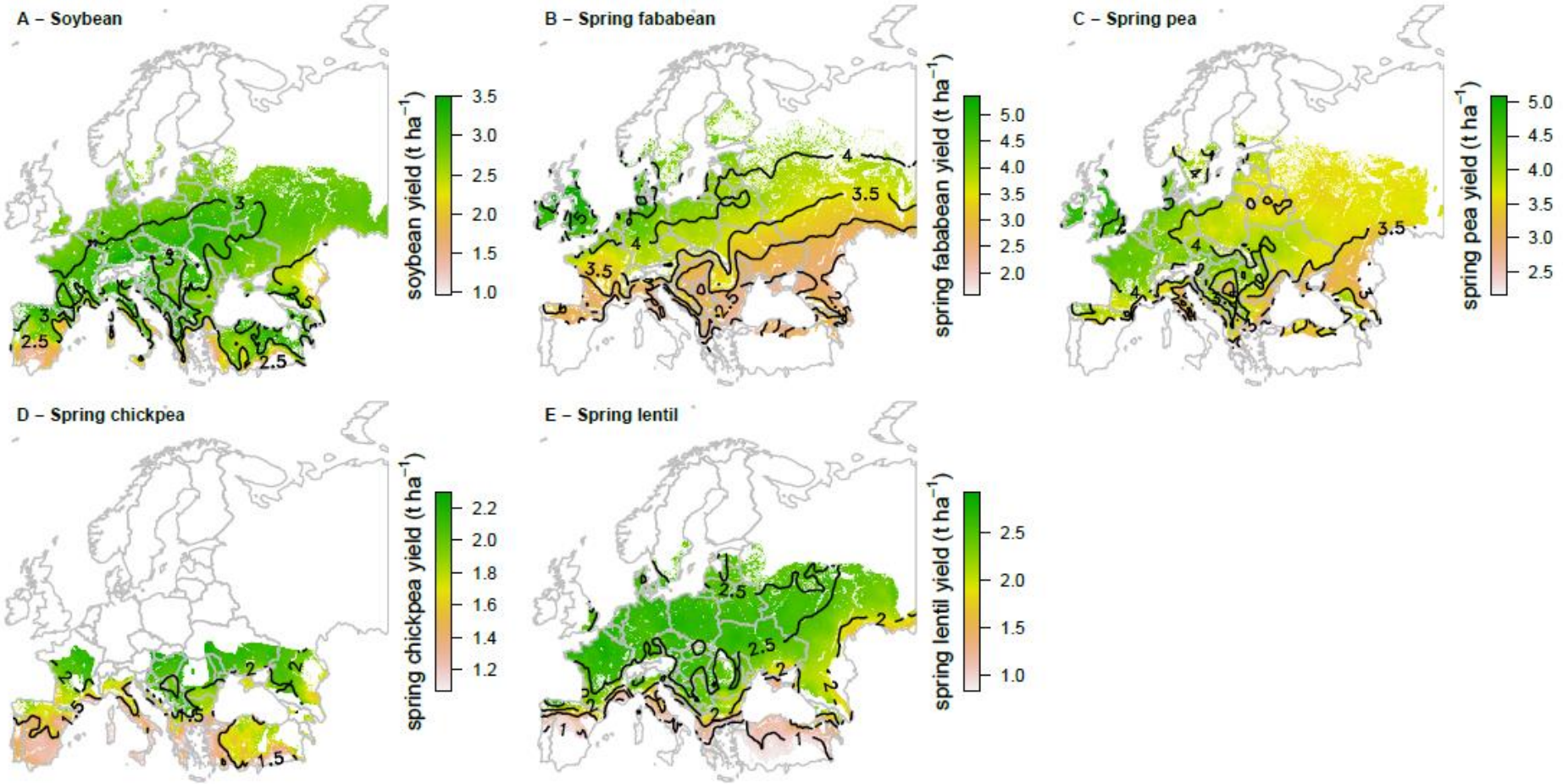

Figure 3. Average projected yields (t ha^{-1}) for spring pulses under historical climate (2000-2020). Projections are shown only on agricultural area (cropland plus pastures) in the year 2000 and where climate conditions are similar enough to climate conditions in the training dataset (see methods for details). 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

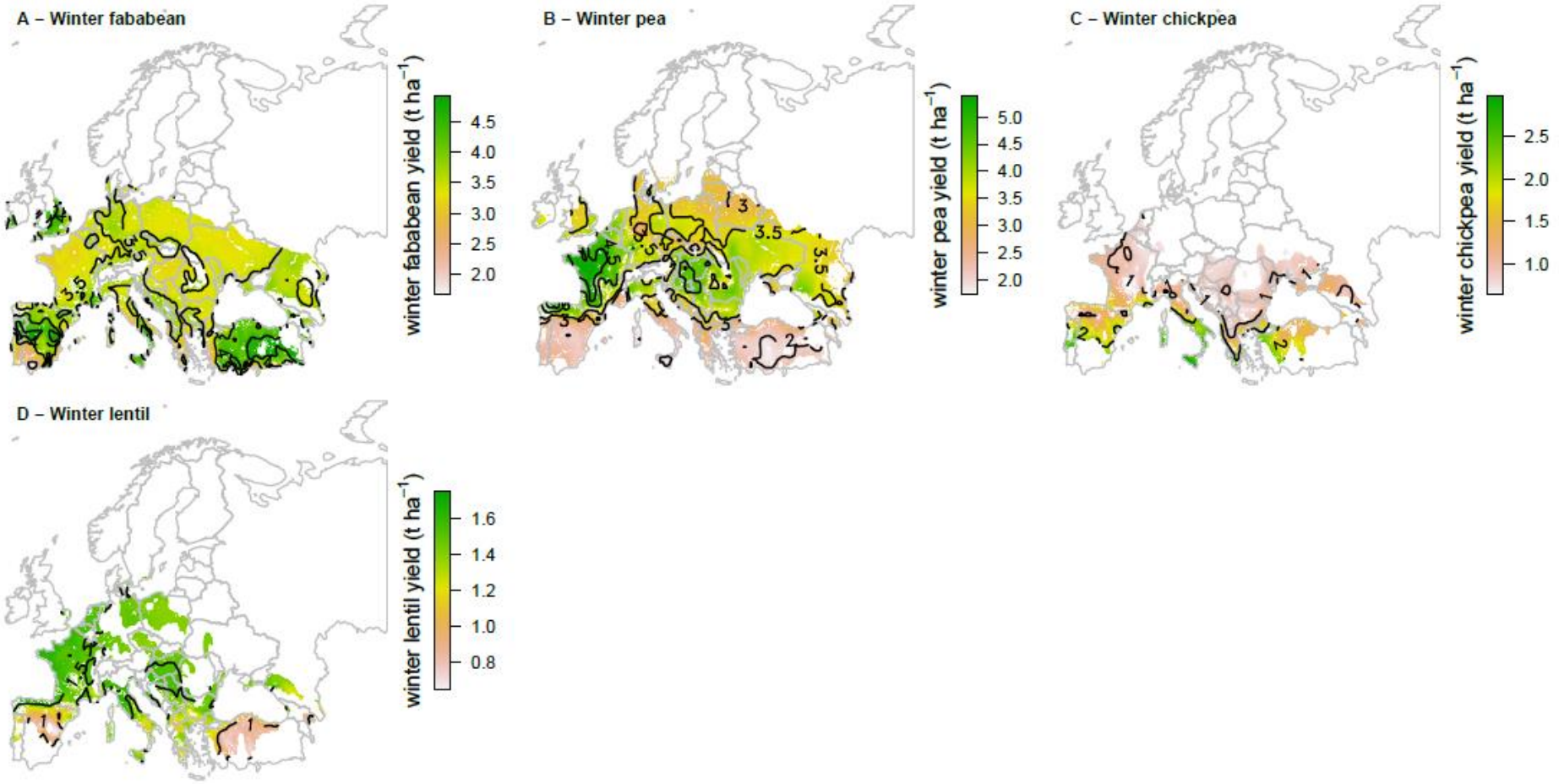

Figure 4. Average projected yields (t ha^{-1}) for winter pulses under historical climate (2000-2020). Projections are shown only on agricultural area (cropland plus pastures) in the year 2000 and where climate conditions are similar enough to climate conditions in the training dataset (see methods for details). 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

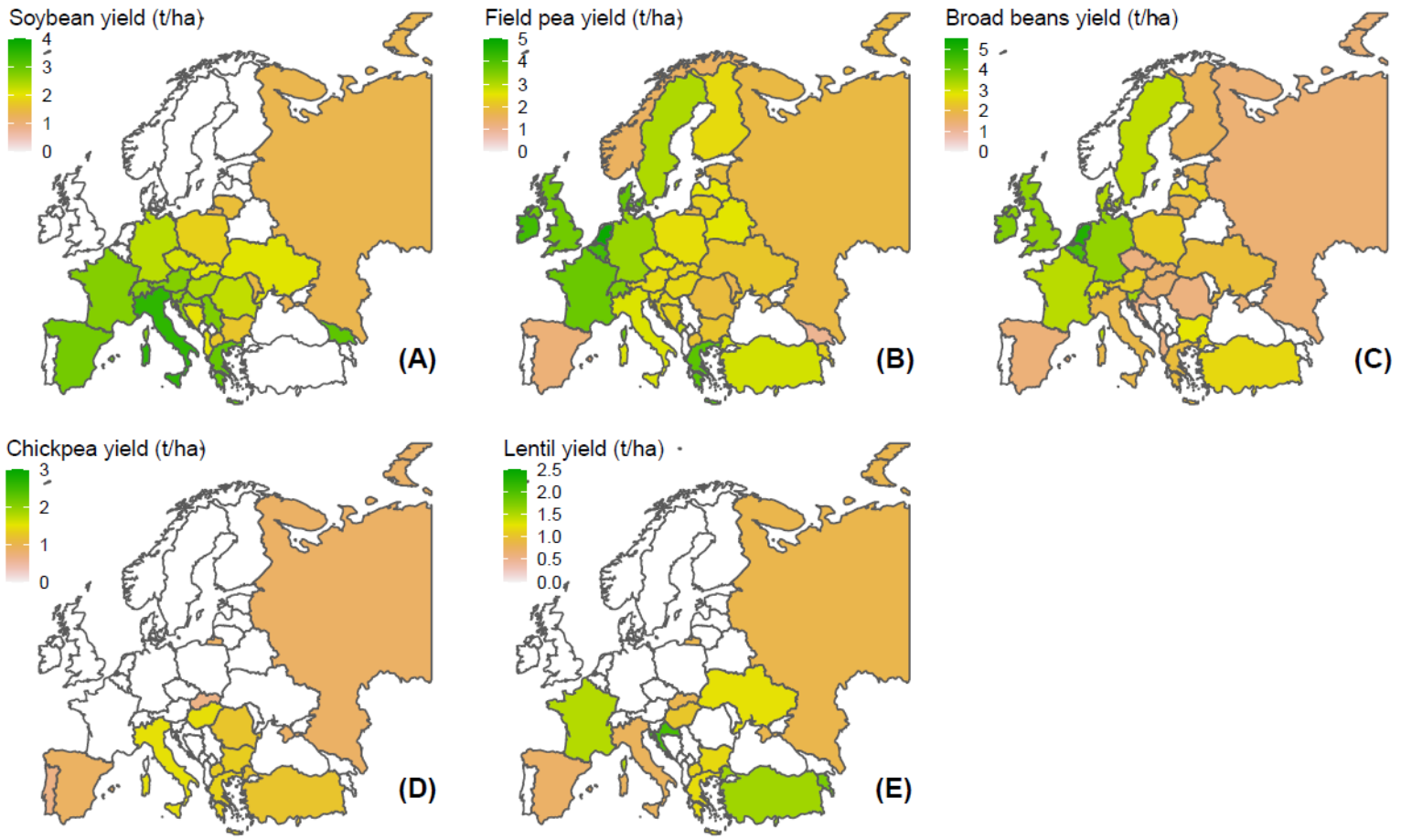

Figure 5. Actual yield (standard moisture content) at the country level (average 2010-2019) for the 5 crops considered here. (A) Soybean, (B) Field pea, (C) Faba bean, (D) Chickpea, (E) Lentil. *Source: FAOSTAT.*

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

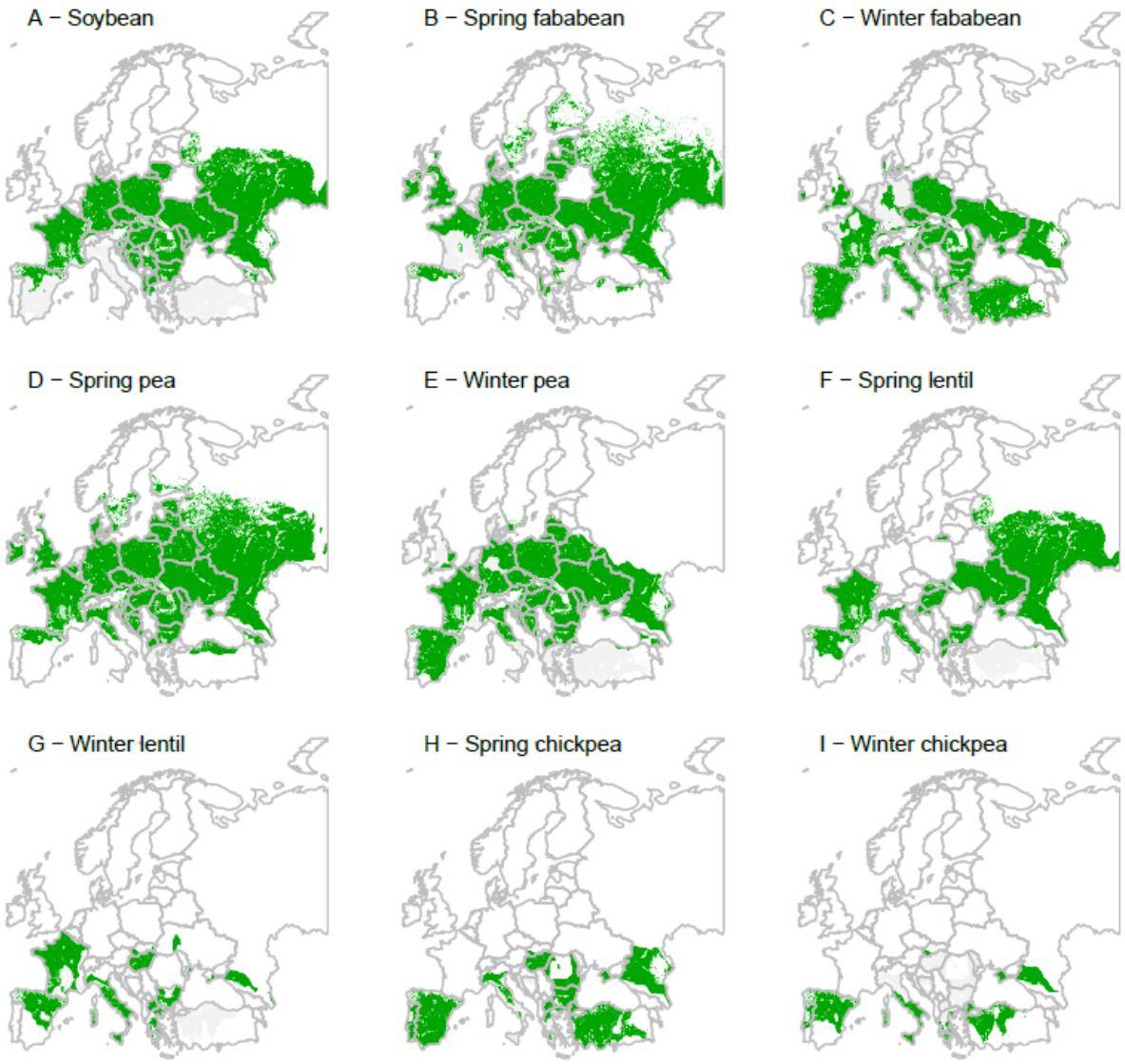

Figure 6. Maps showing where projected yield is higher than actual average national yield by pulse. Green areas show where projected yield is higher than actual average national yield. Grey areas show where projected yield is lower than average national yield. No projection was made in white areas because climate was not similar enough to climate conditions captured in the training dataset of the model. Projected yields are the average yields over 2000-2020 presented in Figure 3 and Figure 4, and actual yield where retrieved from FAOSTAT at the national scale and averaged over 2010-2019.

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

5.3 Next steps and perspectives

5.3.1 Model improvements

Including soil properties in model predictors. The model(s) developed and presented in this report use five climate variables (*tmin, tmax, rain, solar,* and *ETo*) defined at a monthly time step (Equation 2) 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 5.1 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, 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 (Iizumi & Wagai, 2019; Islam et al., 1980). Moreover, soil texture may also affect field trafficability and timely implementation of tillage and sowing or harvesting operations, thus indirectly affecting crop yields. 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 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 21 to Figure 29 (in appendix), 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 (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. 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.

Matching climate predictors with crop phenology. The models developed in this work are based on monthly climate data. Although attention was paid to select months corresponding to crop-specific growing seasons, climate data are not defined based on crop phenology. Previous research has shown the relevance of considering weather data at different crop phenological stages, like temperature and rainfall from sowing to emergence, or during critical periods for yield formation (e.g. flowering). This approach allows to make connections between climate conditions and well-known physiological processes of crops like emergence, seed number formation, seed set, or grain filling. This might represent interesting perspectives for improving the models developed here.

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

Accounting for protein yield. As shown in the results section (Figure 3 and Figure 4), projected yield levels and location of high-yielding areas differ between crops. However, in addition to grain yield, the protein content of harvested grains and therefore the total protein production are of interest. Grain protein content varies (i) between crops, with average values of 40% for soybean, 30% for fababean, 27% for lentils, 24% for pea, and 22% for chickpea (crude protein in % of dry matter according to www.feedipedia.org), and (ii) within crops with important effects of environmental conditions during crop growth, for example observed ranges of crude protein content are 35.3-43.8 % for soybean and 25.2-33.5 % for fababean (www.feedipedia.org). Prediction of protein content and total protein production would require further research, and the modelling approach presented here might be of interest to this aim.

5.3.2 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 of 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 13), (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 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.

5.3.3 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)

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

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.

5.4 Data accessibility

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

Allen, R., Pereira, L., Raes, D., & Smith, M. (1998). Crop evapotranspiration-Guidelines for computing crop water requirements. *FAO Irrigation and Drainage, Rome*, *56*.

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

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

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

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)

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

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

Table 6. Pulses actual yield by country. Data are in t ha^{-1} and represent average yield over the 2010-2019 time period. Source: *FAOSTAT*.

Country	Broad beans	Chickpeas	Lentils	Peas, dry	Soybeans
Albania	1.3	-	-	-	2
Armenia	-	-	1.8	2.3	-
Austria	2.4	-	-	2.4	2.7
Belarus	-	-	-	2.6	-
Belgium	4.4	-	-	4.1	-
Bosnia and Herzegovina	-	3.1	-	2.3	2
Bulgaria	2.8	1.3	1.2	2.1	1.7
Croatia	1.2	-	2	2.4	2.6
Czech Republic	1.3	-	-	2.6	2.1
Denmark	3.2	-	-	3.8	-
Estonia	1.8	-	-	1.9	-
Finland	1.7	-	-	2.4	-
France	3.2	-	1.5	3.7	2.7
Georgia	-	-	-	1	3
Germany	3.6	-	-	3.2	2.3
Greece	2.1	1.4	1.2	3.9	3
Hungary	1.5	1.5	1	2.4	2.4
Ireland	3.8	-	-	4.1	-
Italy	1.9	1.5	0.8	2.6	3.5
Latvia	2.5	-	-	2.5	-
Lithuania	1.8	-	-	2.3	1.5
Luxembourg	2.3	-	-	2.9	-
Moldova	-	3.7	-	1.7	1.5
Montenegro	-	-	-	2.8	-
Netherlands	5	-	-	4.9	-
Norway	-	-	-	1.5	-
Poland	2.4	-	-	2.4	1.8
Portugal	8.4	0.7	-	-	-
Republic of Macedonia	-	1.3	1.1	2	1.7
Romania	1.3	1.2	-	1.9	2.3
Russia	1.4	0.9	0.9	1.8	1.4
Serbia	-	-	-	-	2.8
Slovakia	1.6	0.7	0.8	2.3	2.1
Slovenia	3.4	-	-	2.4	2.6
Spain	1.4	0.9	0.7	1.3	2.9
Sweden	3.1	-	-	3	-
Switzerland	2.9	-	-	3.5	2.7
Turkey	2.6	1.2	1.6	2.7	4.1
Ukraine	2.1	-	1.2	2.1	2.1
United Kingdom	3.6	-	-	3.6	-
Europe (all countries)	**2.9**	**1.0**	**0.9**	**2.1**	**1.9**

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

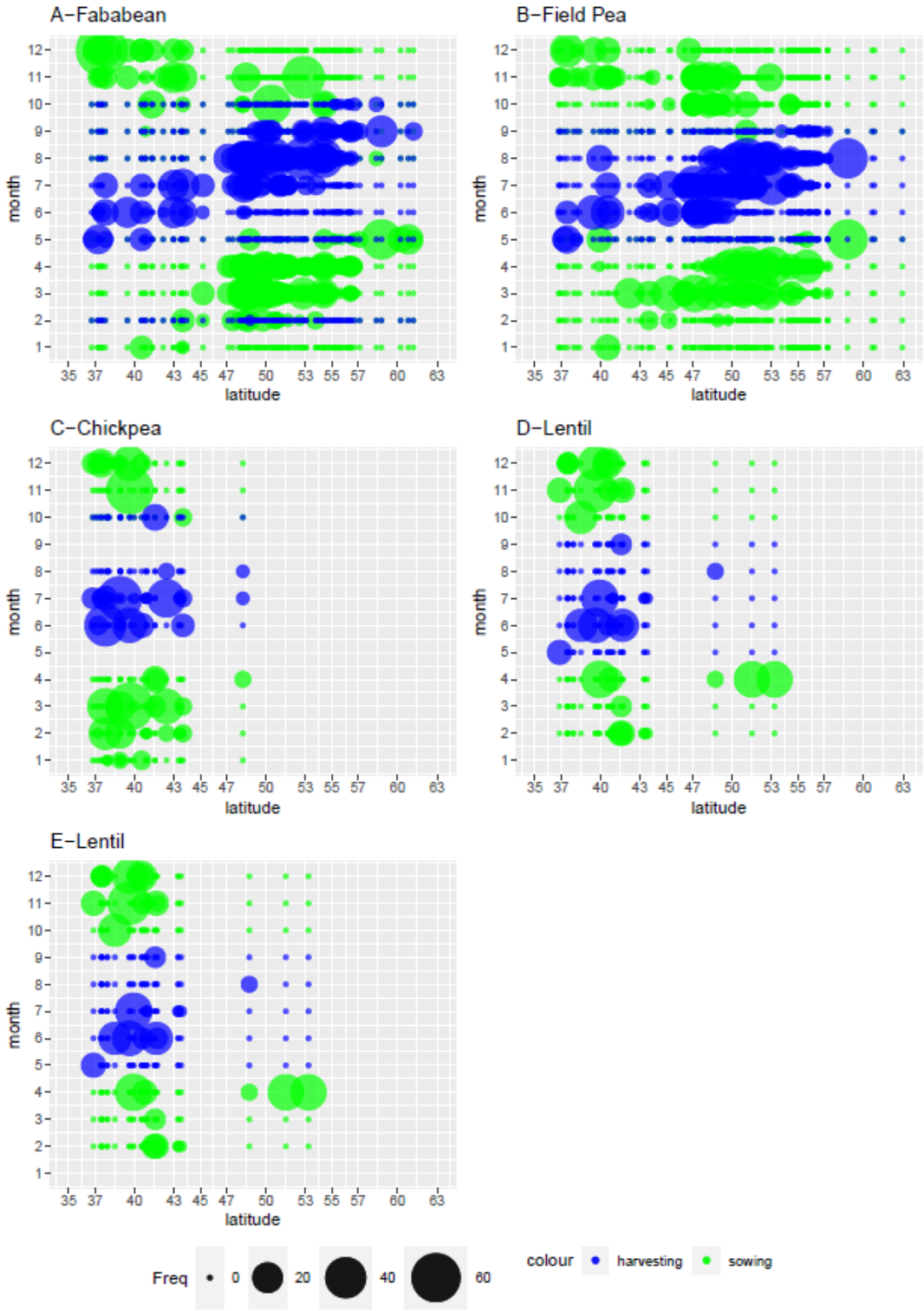

Figure 7. Frequency of sowing and harvesting months based on the latitude of the experimentations for each pulse. Each point represents experimentations with the same sowing or harvesting months. The size of the point increase with the number of experimentations sown or harvested on the same month. The 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

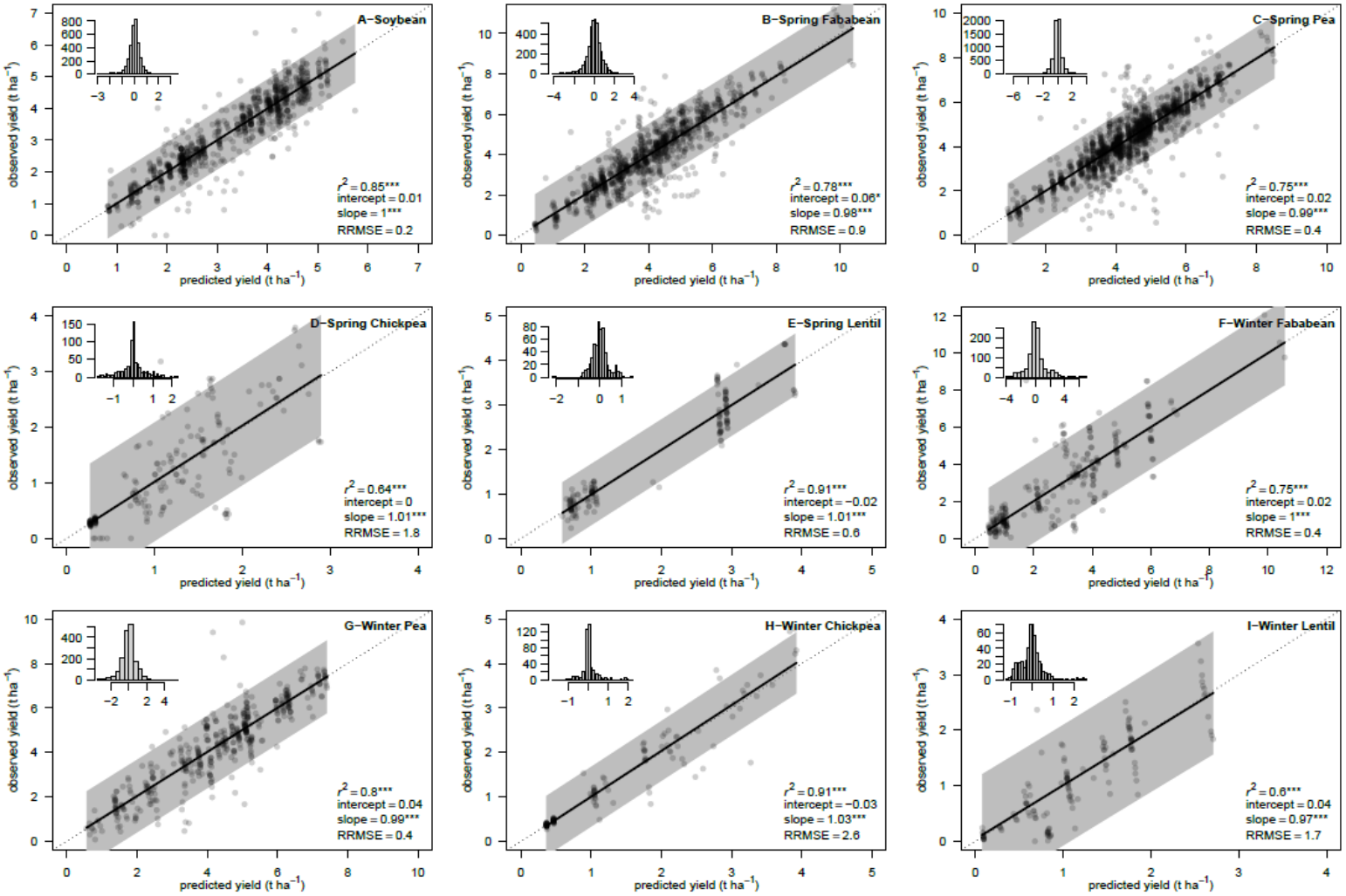

Figure 8. Assessment of the Random Forest algorithm for the different crops considered in this study. For each crop, the model is evaluated using a classical bootstrap approach with 25 resamplings, and out-of-bag predictions are compared to observed yields. Black lines represent the linear regression 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%.

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

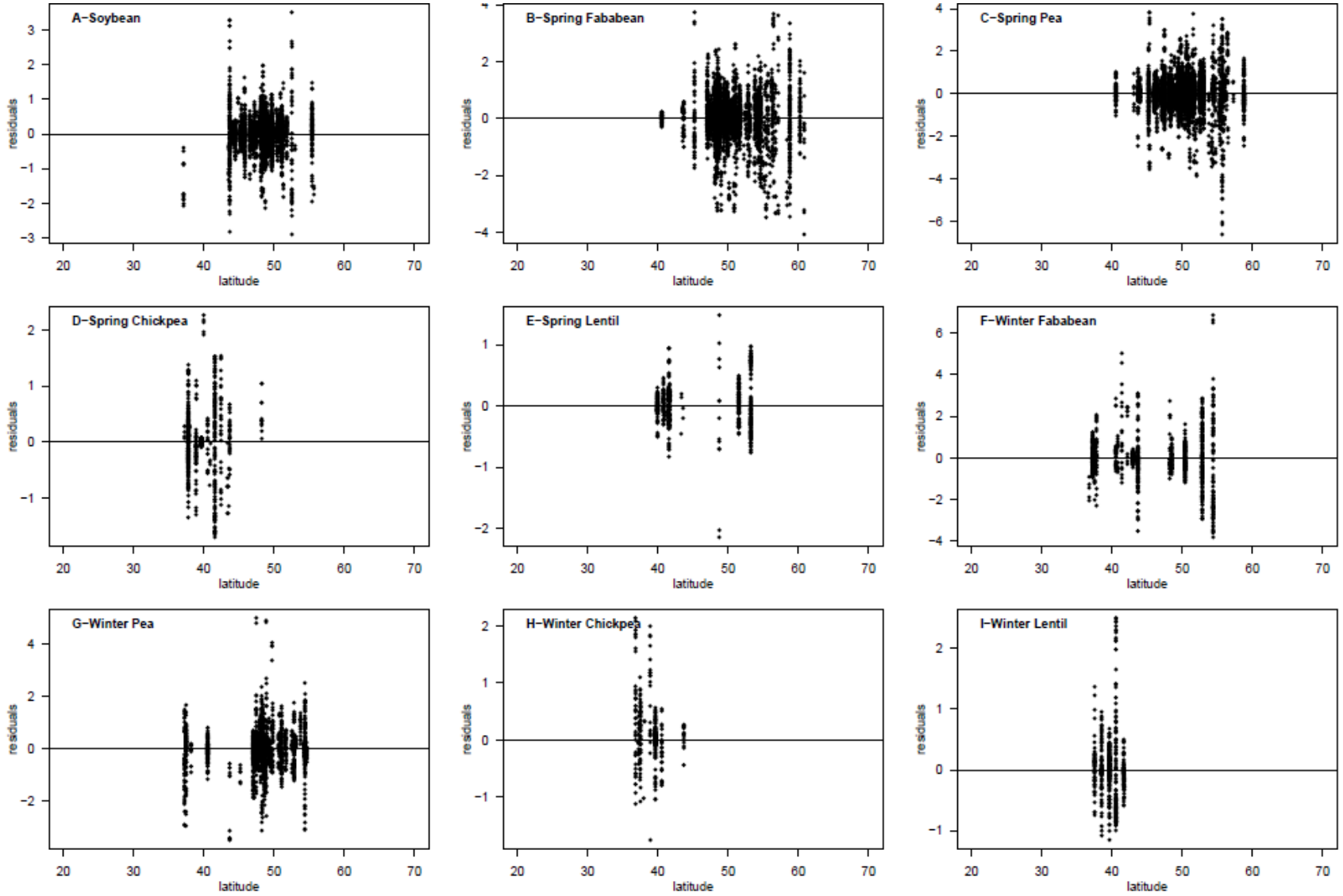

Figure 9. Analysis of model residuals: residuals as a function of latitude.

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

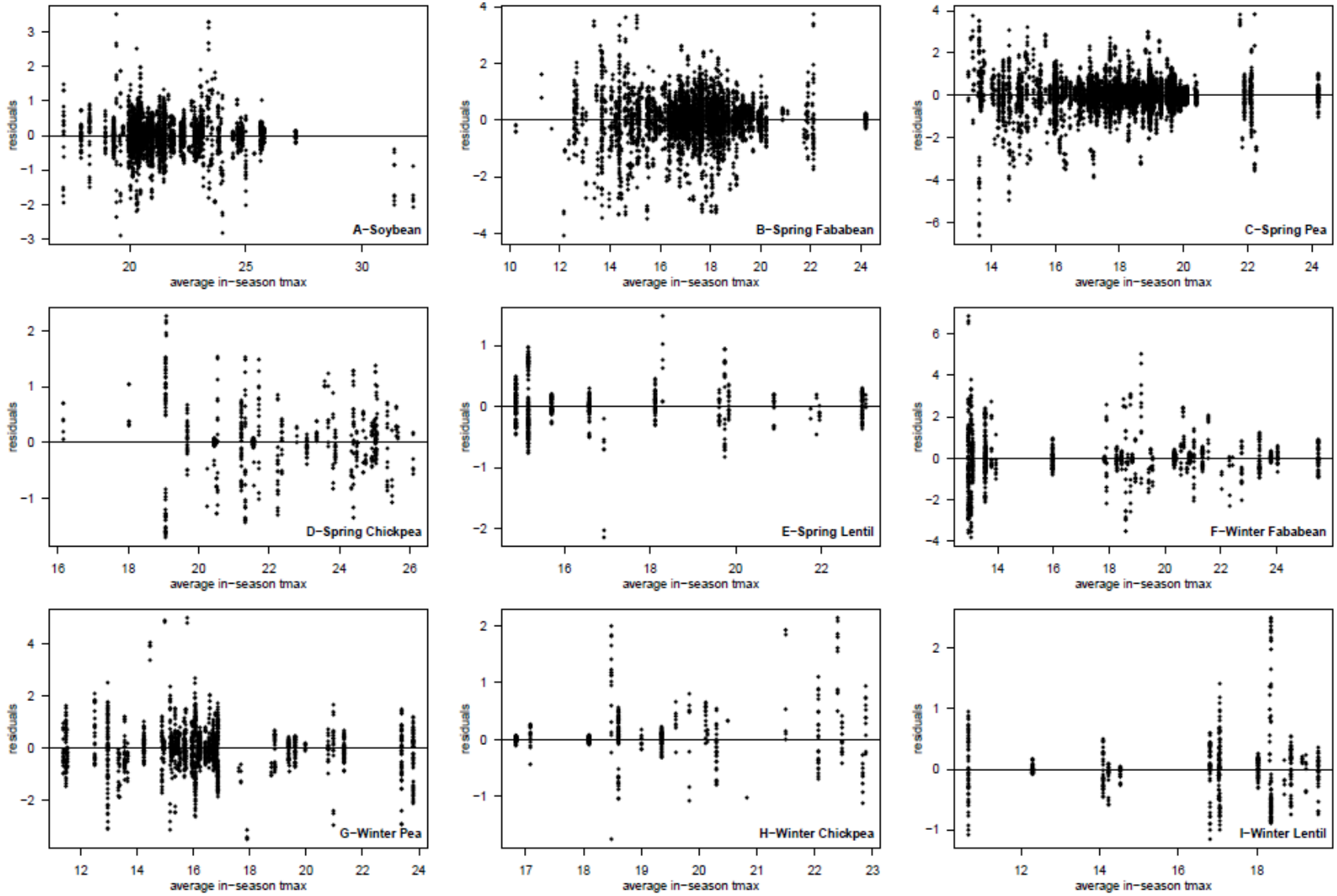

Figure 10. Analysis of model residuals: residuals as a function of average in-season tmax.

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

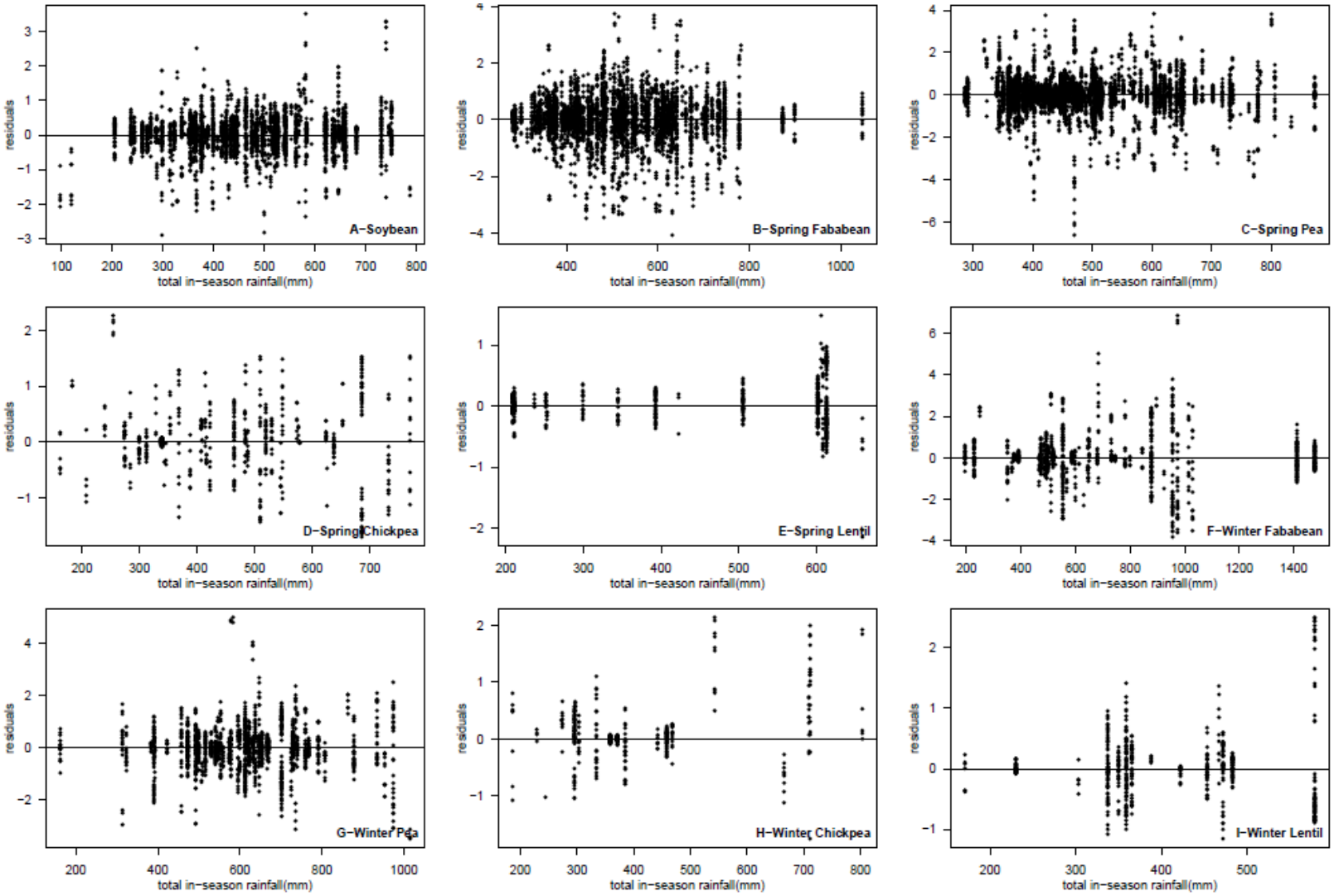

Figure 11. Analysis of model residuals: residuals as a function of total in-season rainfall.

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

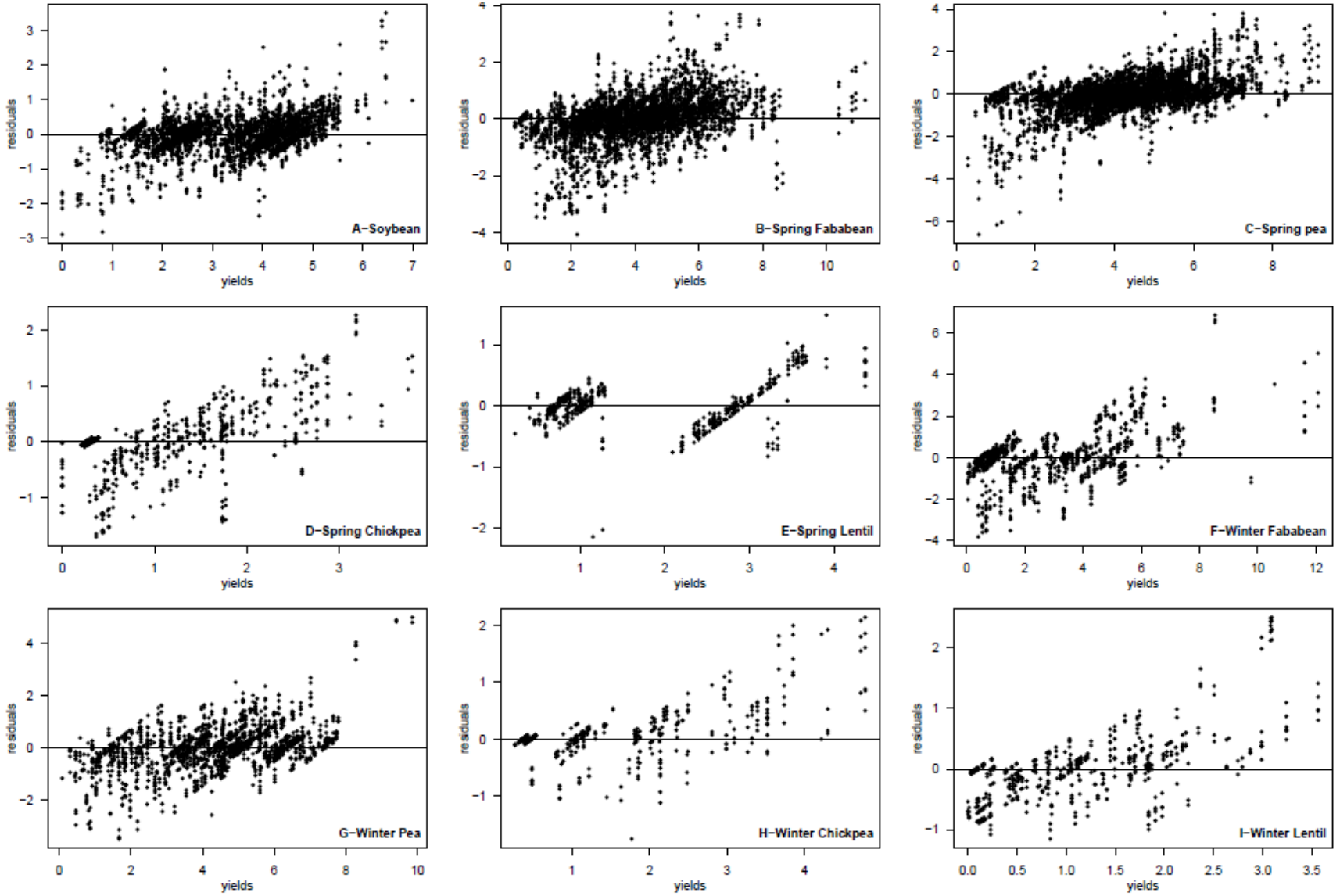

Figure 12. Analysis of model residuals: residuals as a function of observed yields.

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

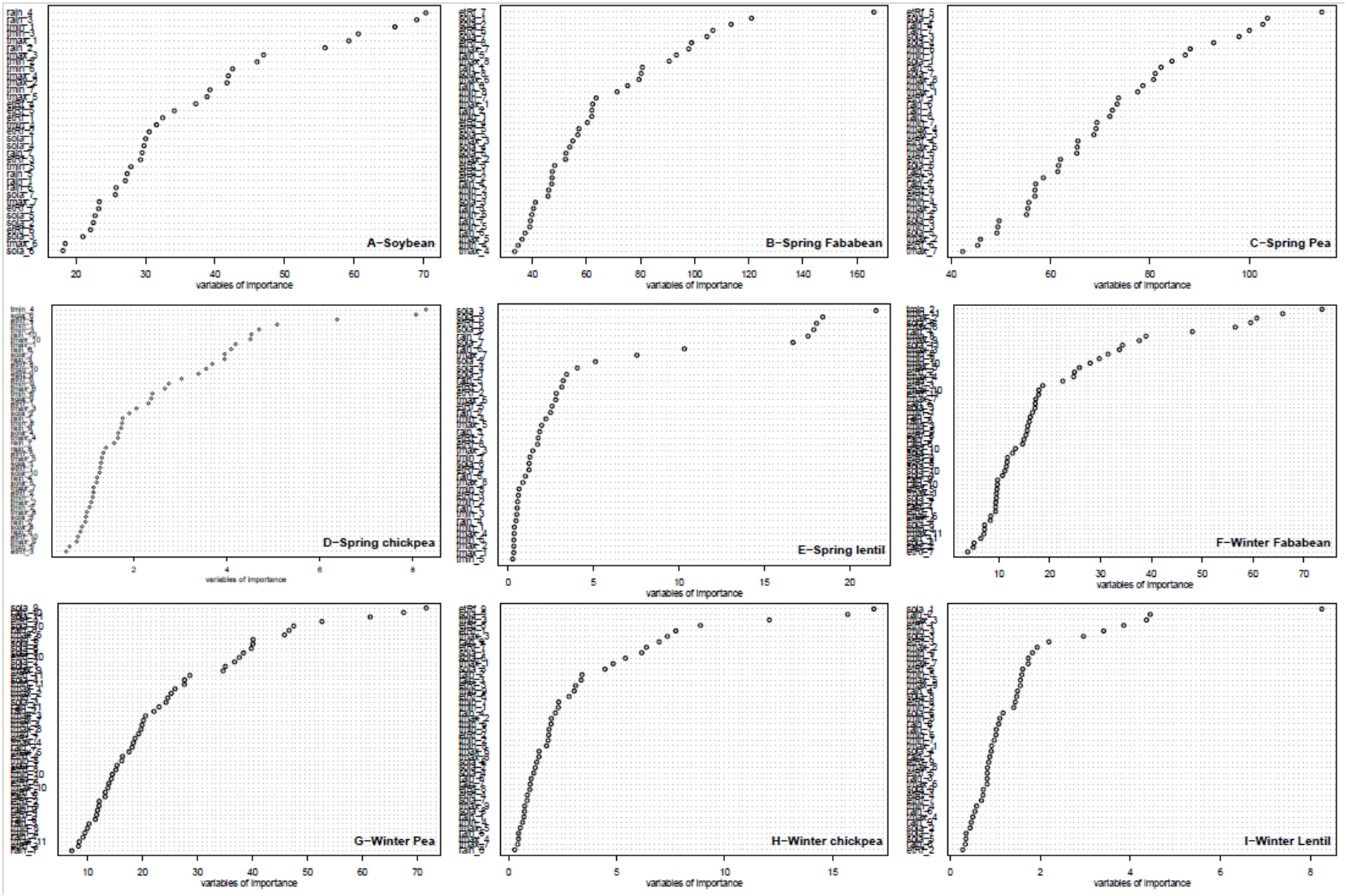

Figure 13. Variables importance plots derived from the Random Forest algorithm. *tmin* (°C) is the monthly average of daily minimum temperature, *tmax* (°C) is the monthly average of daily maximum temperature, *rain* (mm day^{-1}) is the monthly average of daily rainfall, *solar* (W m^{-2}) is the monthly average of daily downward shortwave radiation, *etRf* is the monthly average of daily reference evapotranspiration (mm day^{-1}). 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

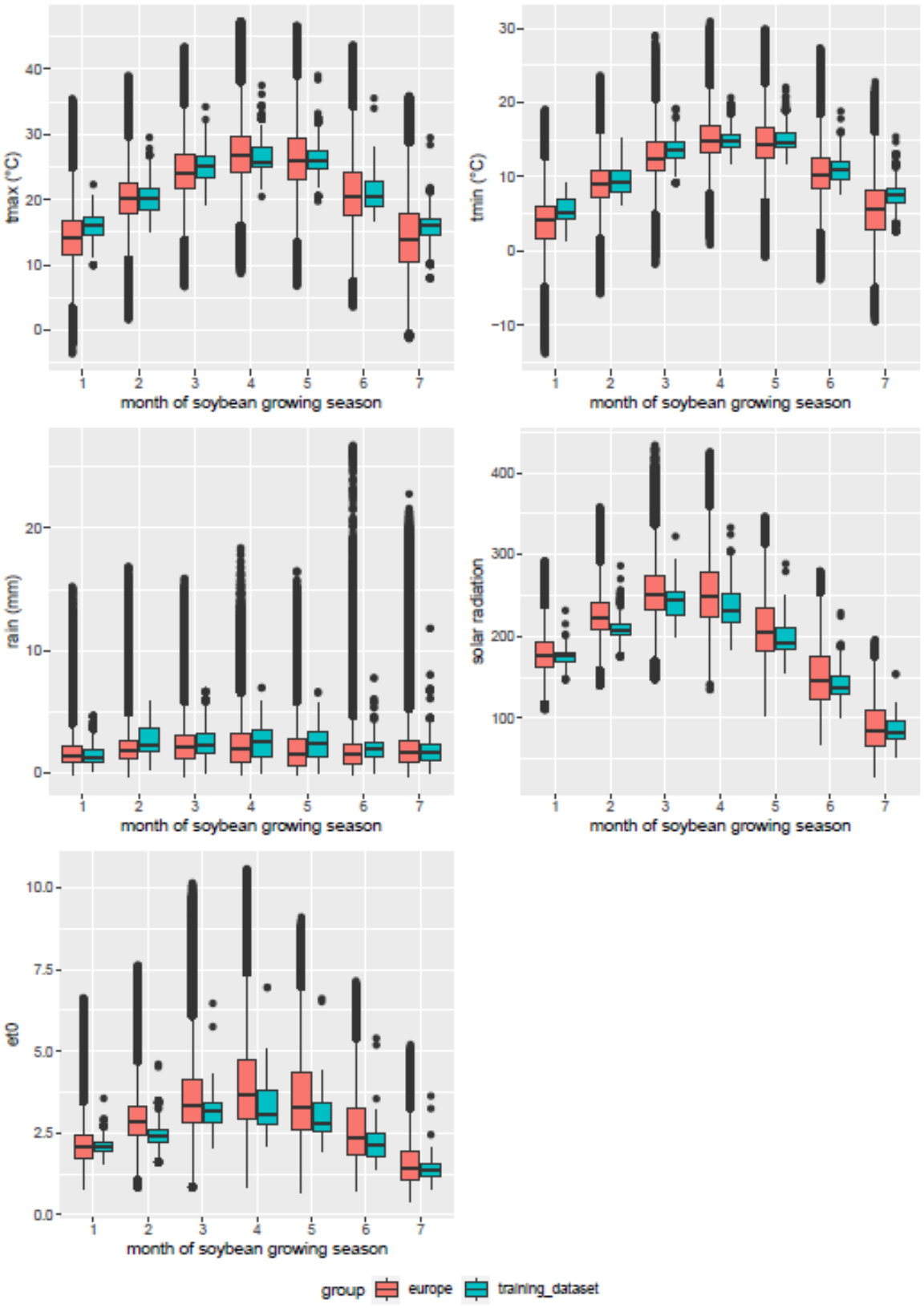

Figure 14. Comparison of climatic variables distribution as observed in the training dataset and over Europe for soybean. For Europe, the distribution concerns climate data from 2000-2020.

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

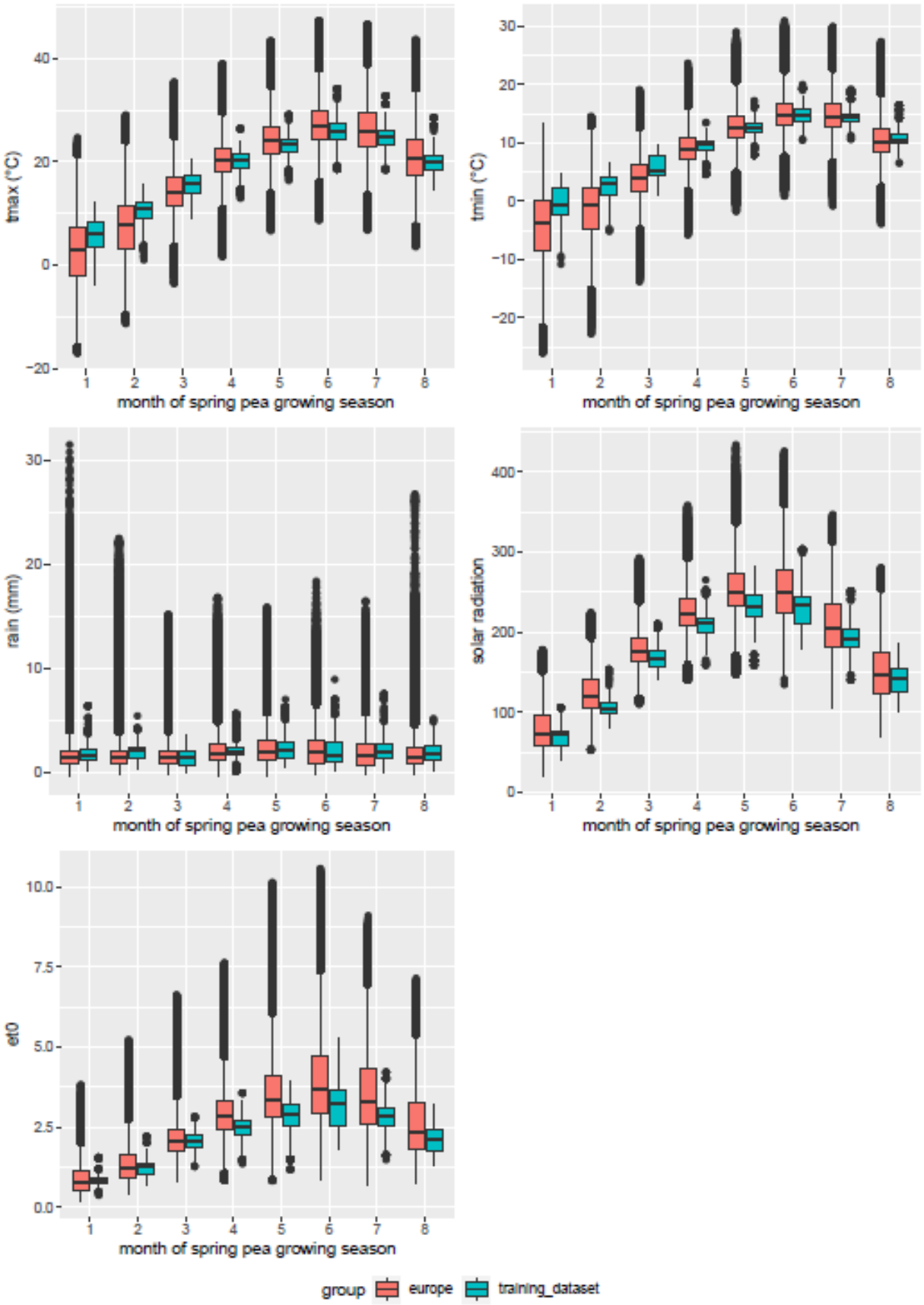

Figure 15. Comparison of climatic variables distribution as observed in the training dataset and over Europe for spring pea. For Europe, the distribution concerns climate data from 2000-2020.

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

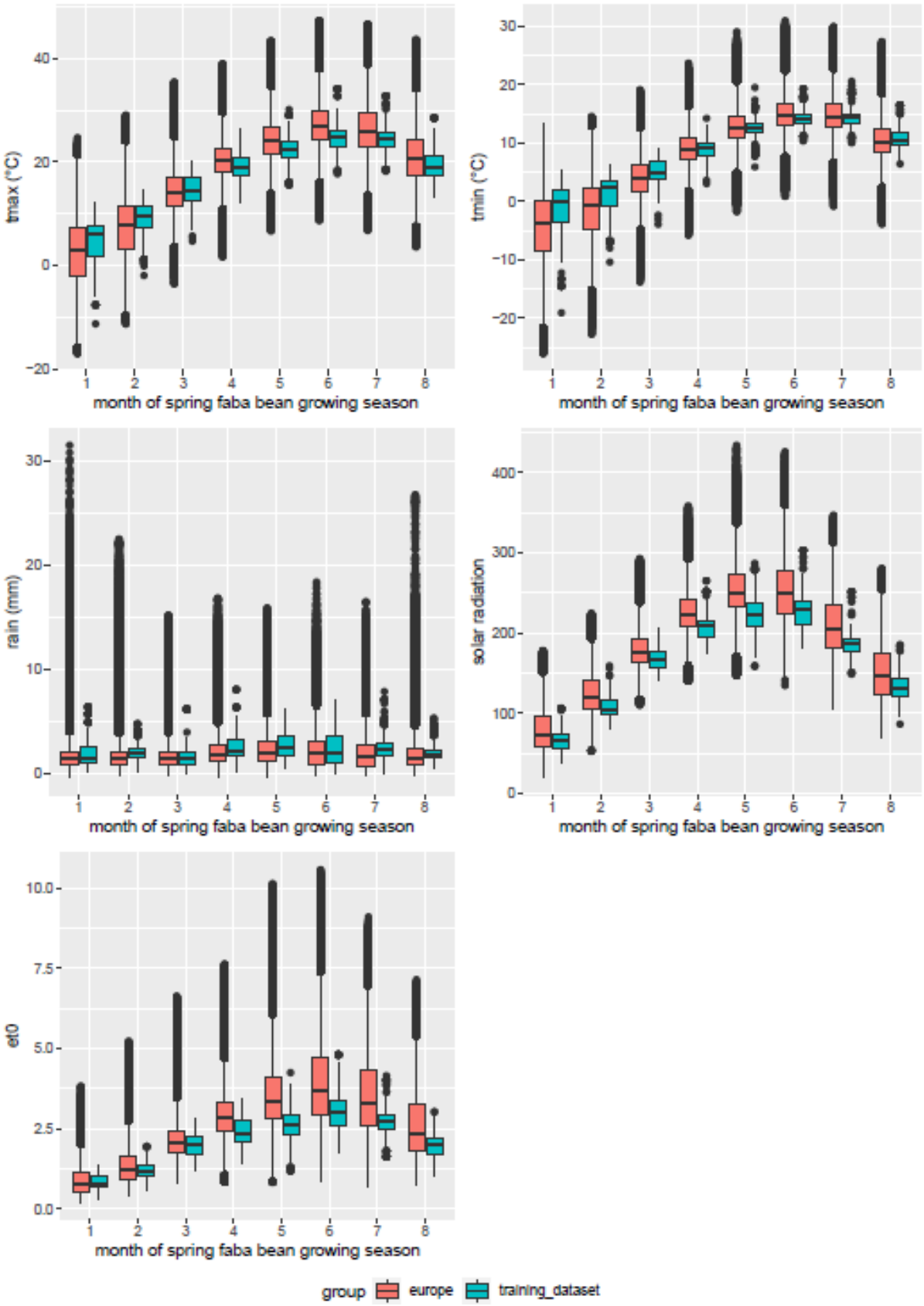

Figure 16. Comparison of climatic variables distribution as observed in the training dataset and over Europe for spring fababean. For Europe, the distribution concerns climate data from 2000-2020.

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

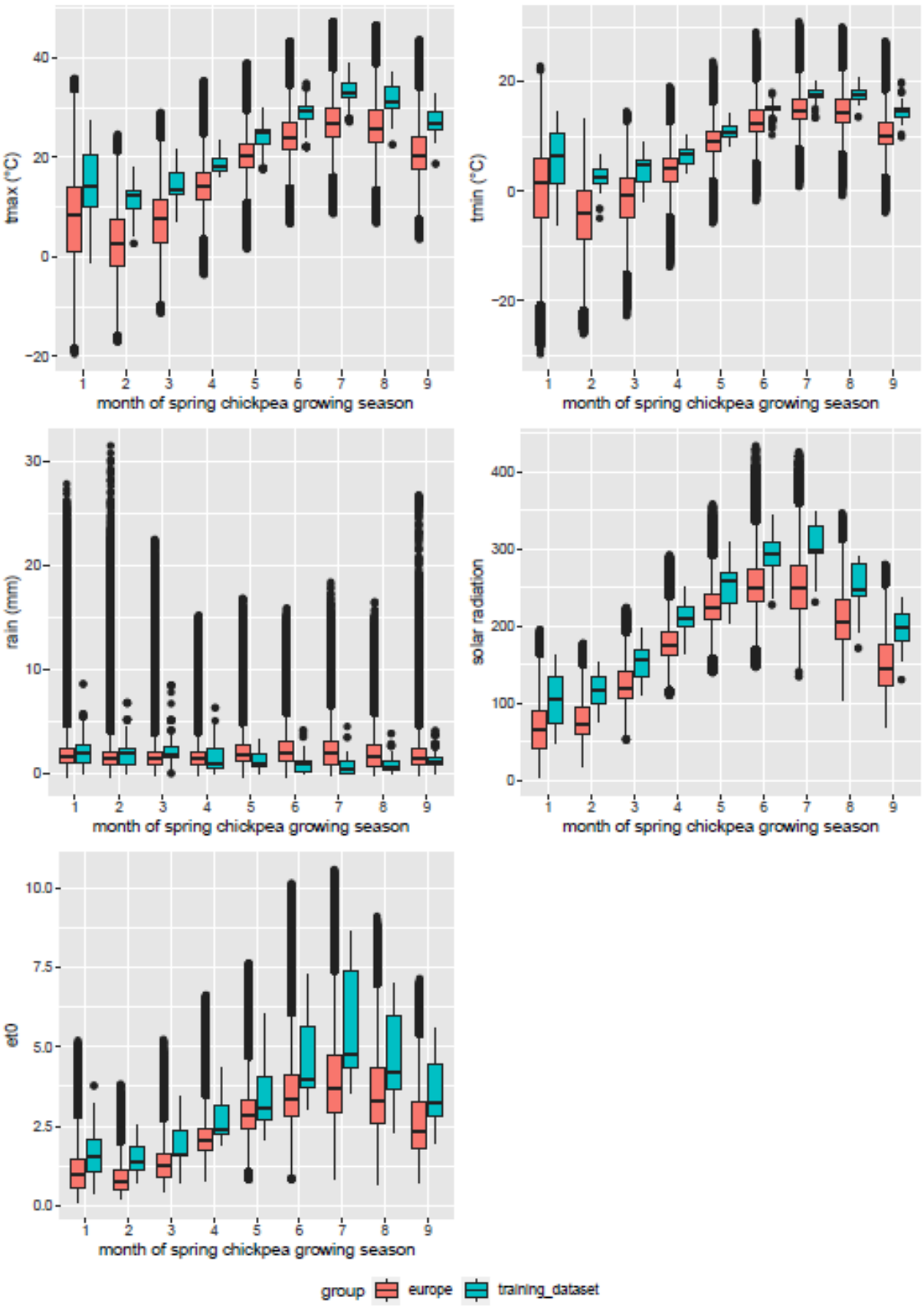

Figure 17. Comparison of climatic variables distribution as observed in the training dataset and over Europe for spring chickpea. For Europe, the distribution concerns climate data from 2000-2020.

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

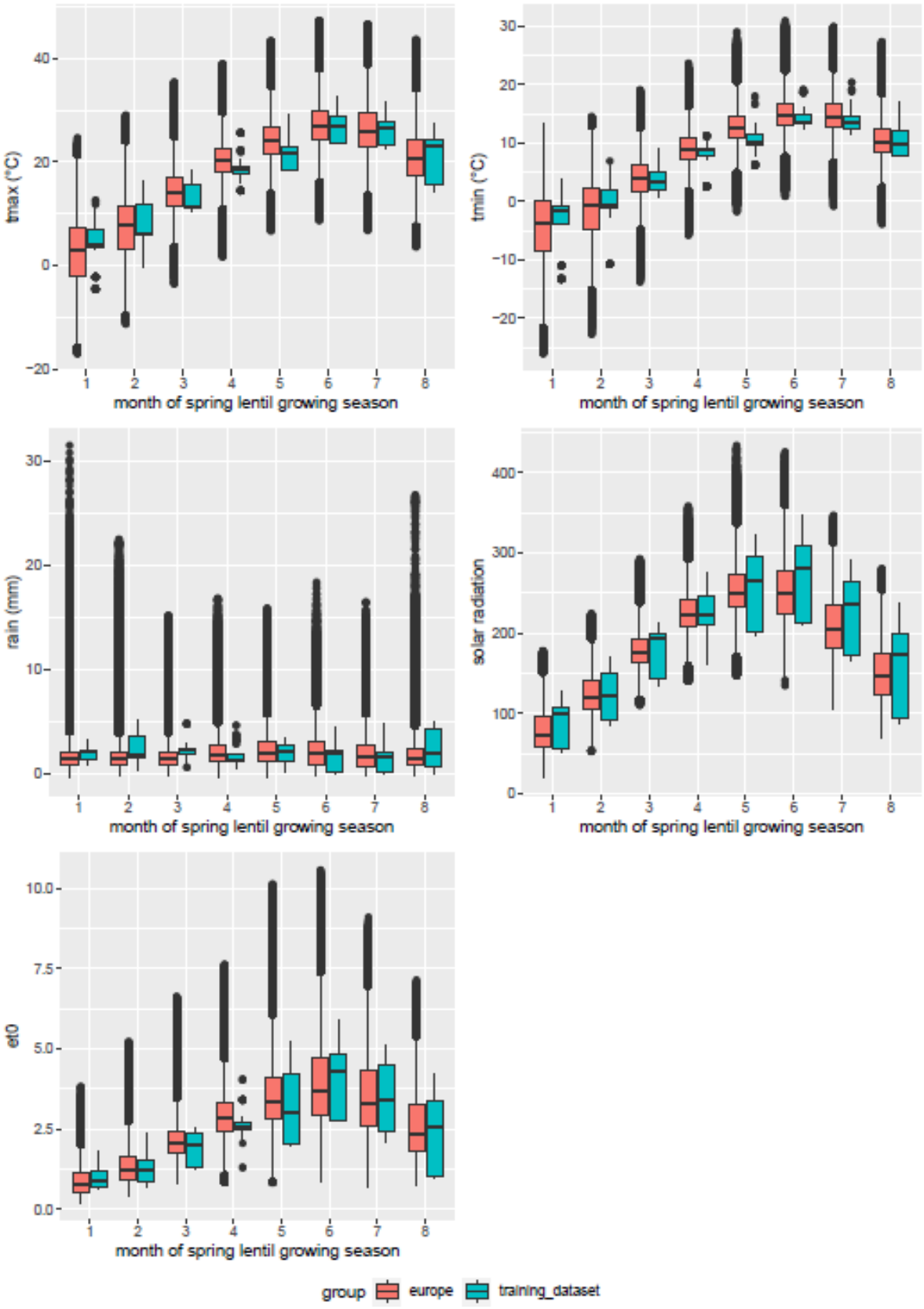

Figure 18.Comparison of climatic variables distribution as observed in the training dataset and over Europe for spring lentil. For Europe, the distribution concerns climate data from 2000-2020.

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

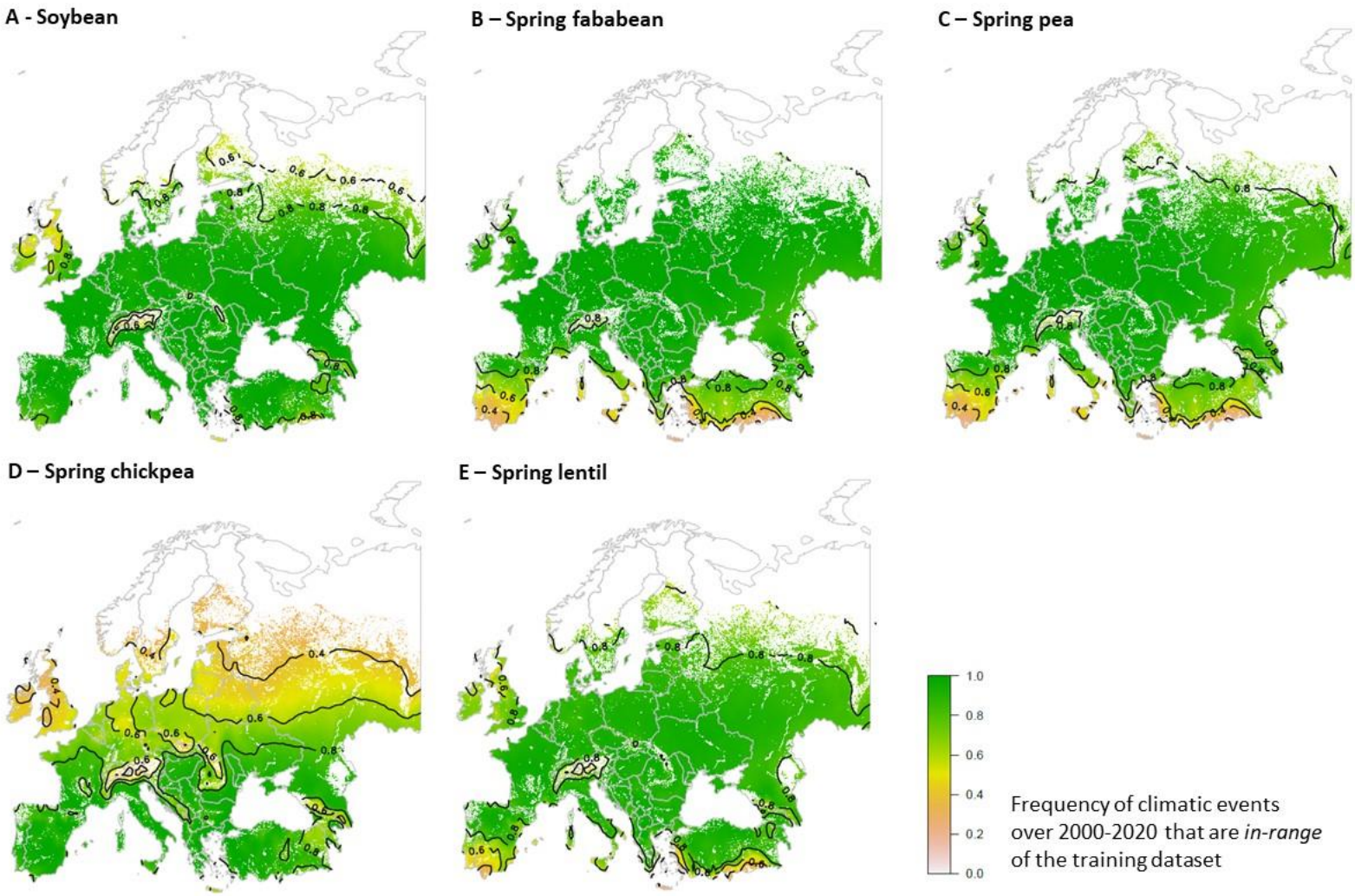

Figure 19. Frequency of climate events in-range of climate conditions observed in training dataset for spring crops. See Materials and methods section 4.2.3 for details. Pixels where the frequency is low denote areas where climatic conditions are different from climatic in the training dataset. Frequency was calculated on the 2000-2020 time period, taking into account crop-specific growing seasons.

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

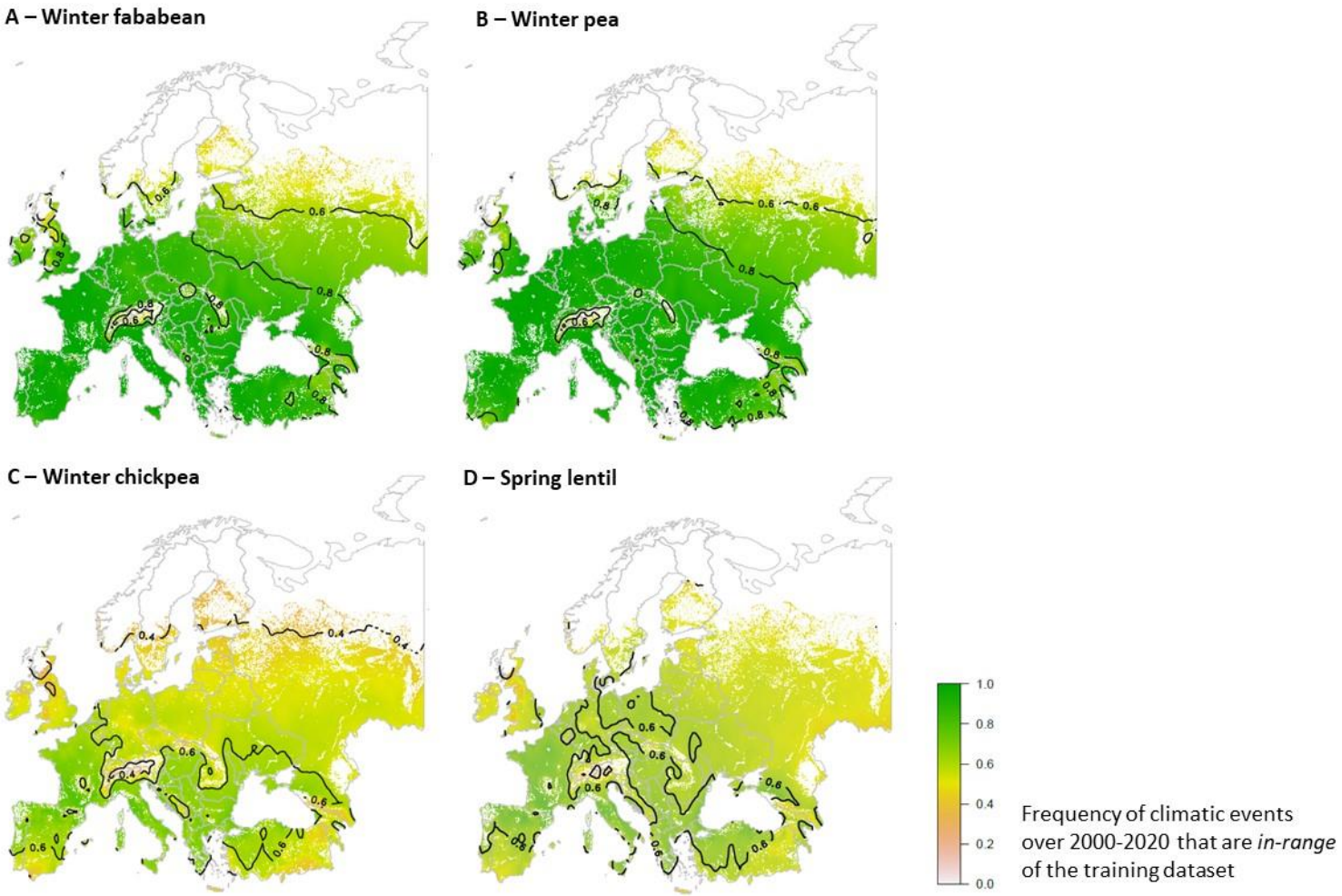

Figure 20. Frequency of climate events in-range of climate conditions observed in training dataset for winter crops. See Materials and methods section 4.2.3 for details. Pixels where the frequency is low denote areas where climatic conditions are different from climatic in the training dataset. Frequency was calculated on the 2000-2020 time period, taking into account crop-specific growing seasons

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

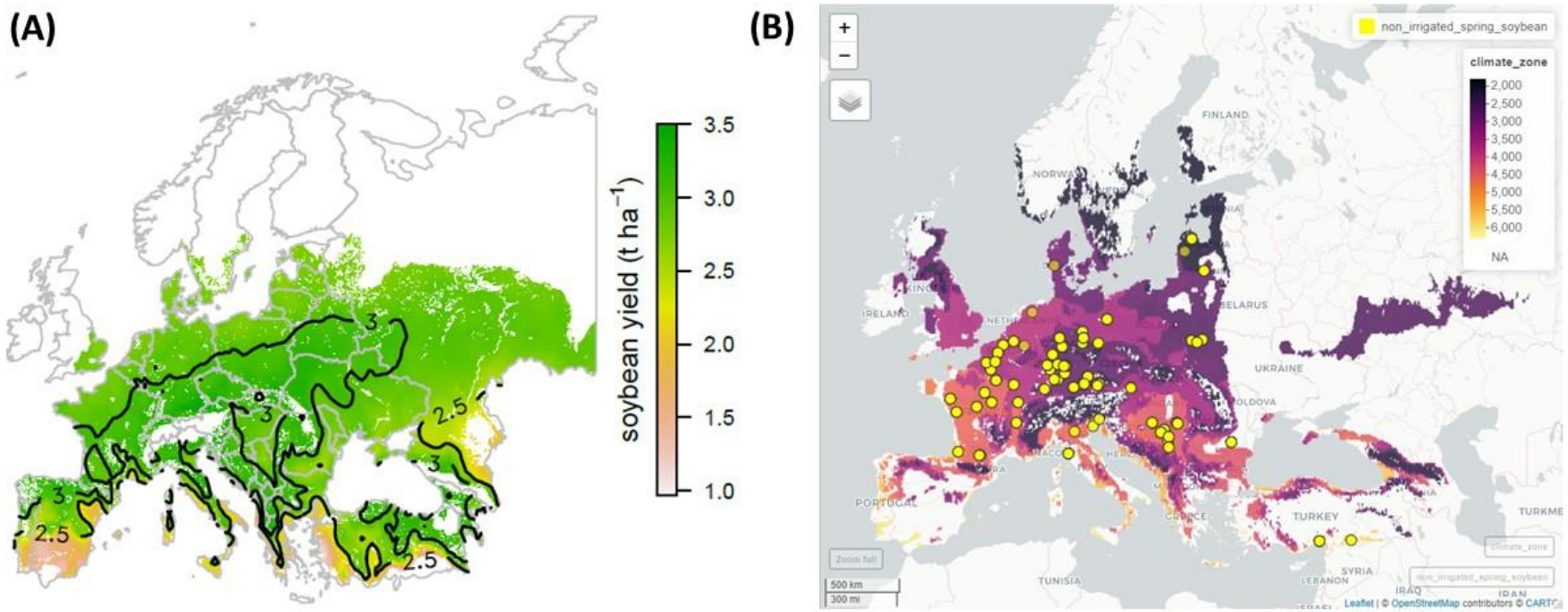

Figure 21. (A) Soybean projected yields (t ha^{-1} at 13% moisture) under historical climate (average 2000-2020) and (B) maps of locations of experiments used for model fitting (yellow dots) and climate zones containing at least one experiment (filled polygons). Projections are shown only on agricultural area (cropland plus pastures) in the year 2000 and where climate conditions are reasonably similar to climate conditions in training dataset (see text for details). 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

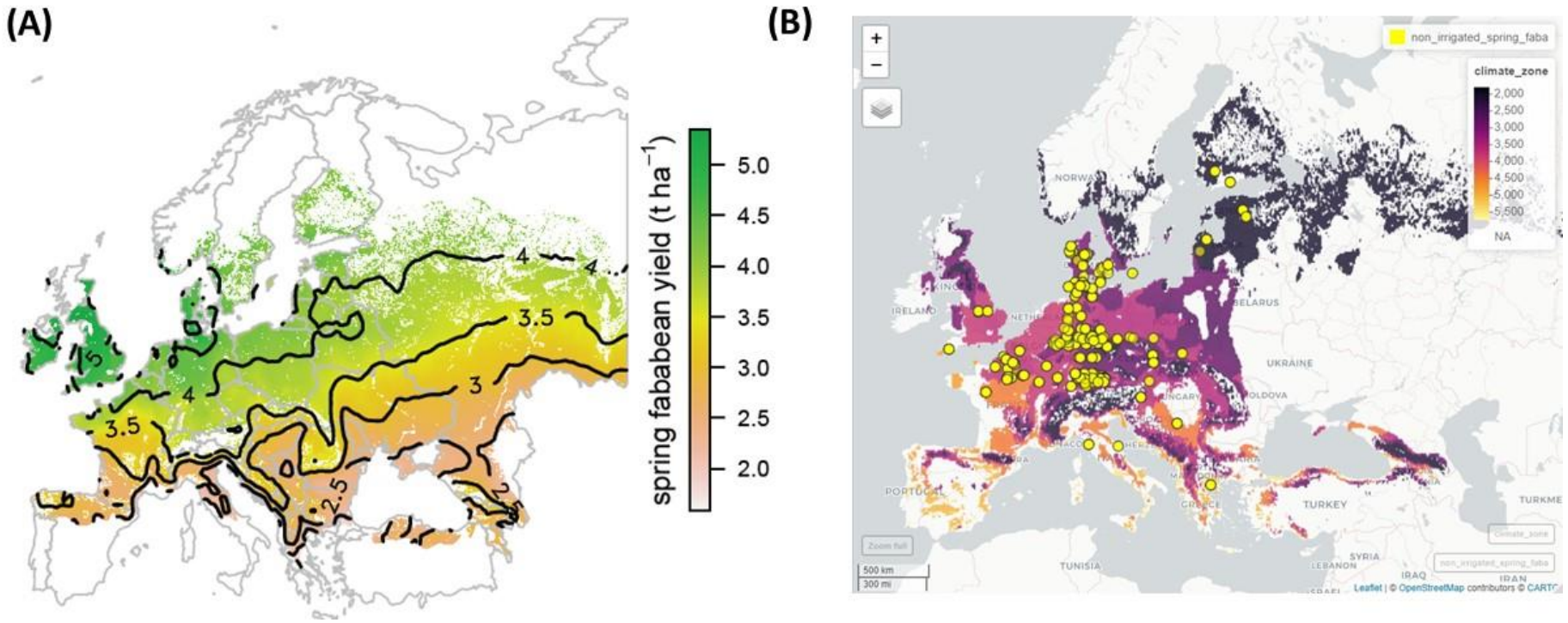

Figure 22. (A) Spring faba bean projected yields (t ha^{-1} at 13% moisture) under historical climate (average 2000-2020) and (B) maps of locations of experiments used for model fitting (yellow dots) and climate zones containing at least one experiment (filled polygons). Projections are shown only on agricultural area (cropland plus pastures) in the year 2000 and where climate conditions are reasonably similar to climate conditions in training dataset (see text for details). 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

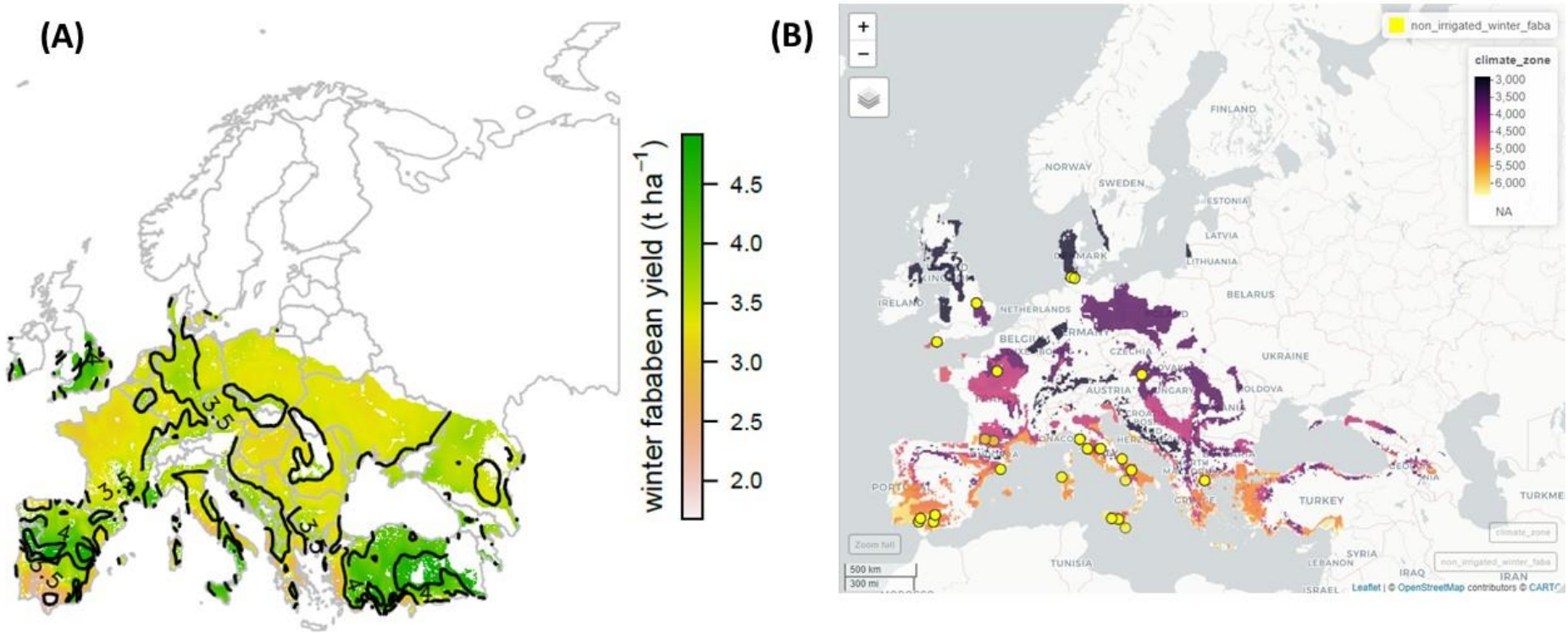

Figure 23. Winter faba bean projected yields (t ha^{-1} at 13% moisture) under historical climate (2000-2020) and maps of locations of experiments used for model fitting (yellow dots) and climate zones containing at least one experiment (filled polygons). The first map shows the median projected yield 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

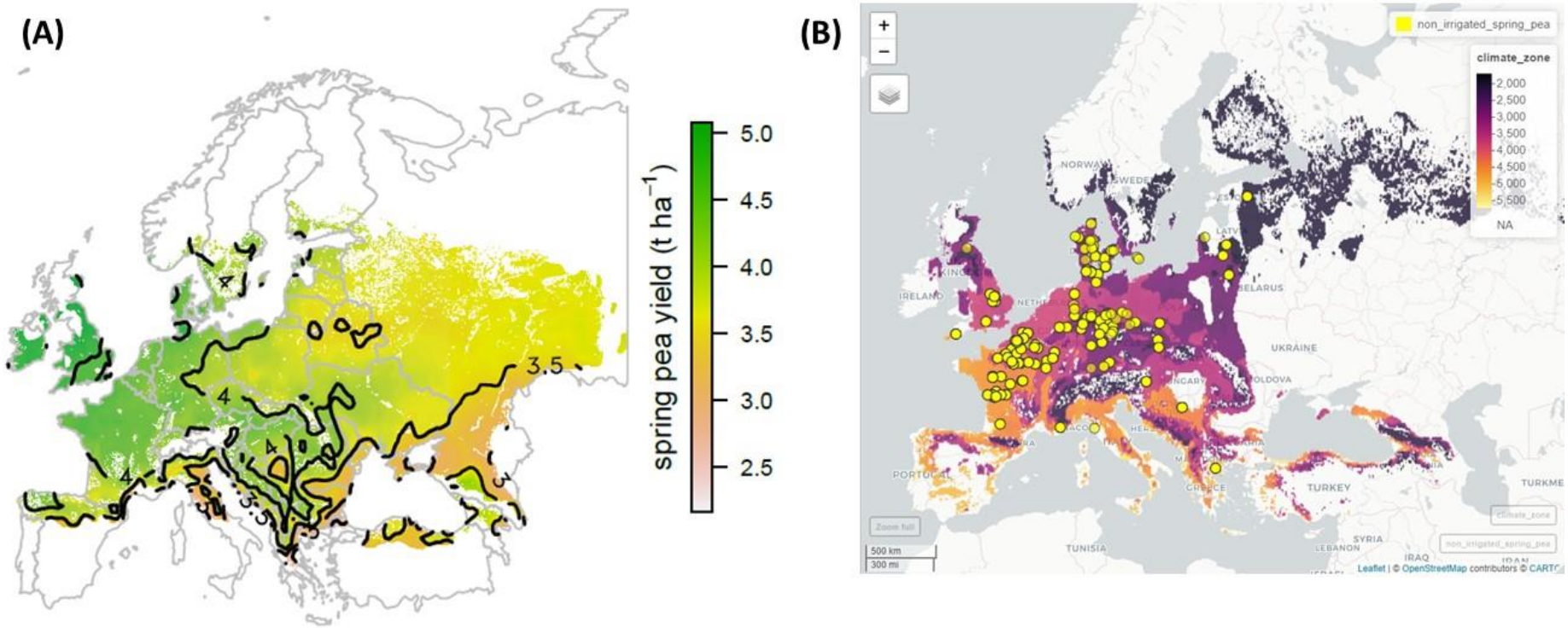

Figure 24. (A) Spring field pea projected yields (t ha^{-1} at 13% moisture) under historical climate (average 2000-2020) and (B) maps of locations of experiments used for model fitting (yellow dots) and climate zones containing at least one experiment (filled polygons). Projections are shown only on agricultural area (cropland plus pastures) in the year 2000 and where climate conditions are reasonably similar to climate conditions in training dataset (see text for details). 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

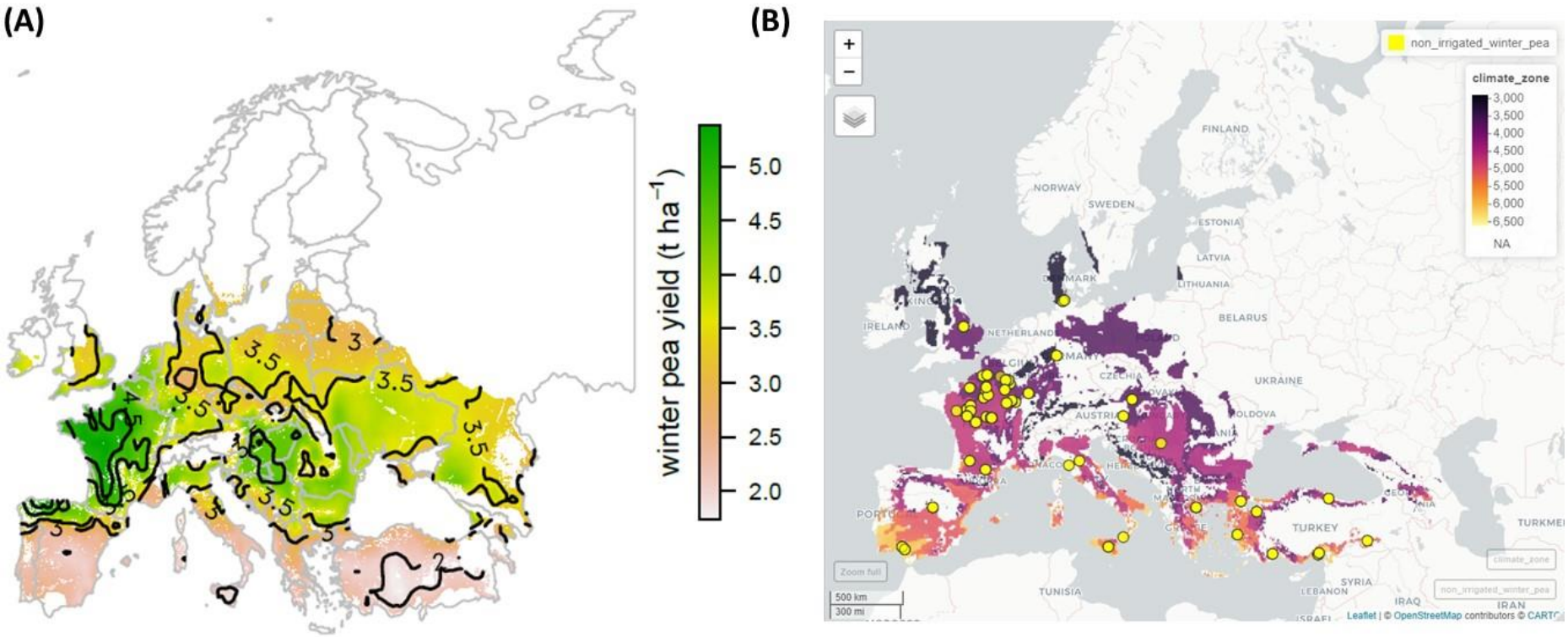

Figure 25. (A) Winter field pea projected yields (t ha^{-1} at 13% moisture) under historical climate (average 2000-2020) and (B) maps of locations of experiments used for model fitting (yellow dots) and climate zones containing at least one experiment (filled polygons). Projections are shown only on agricultural area (cropland plus pastures) in the year 2000 and where climate conditions are reasonably similar to climate conditions in training dataset (see text for details). 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

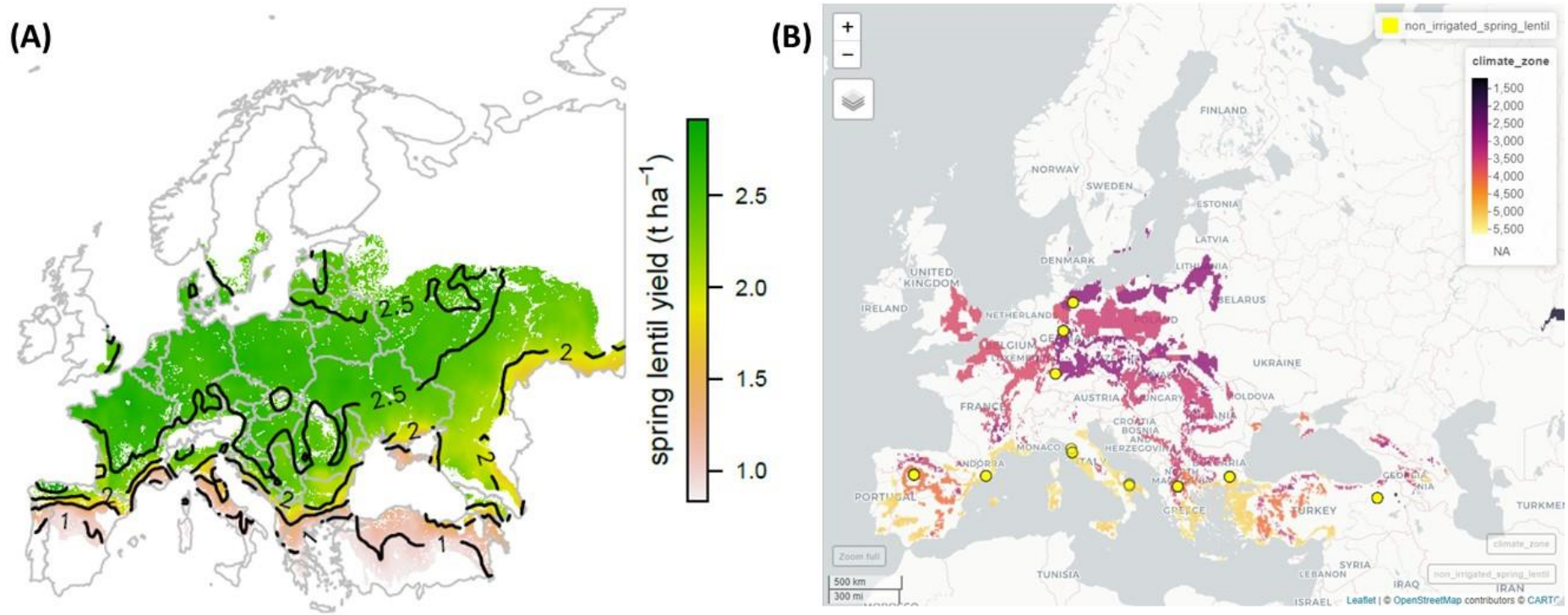

Figure 26. (A) Spring lentil projected yields (t ha^{-1} at 13% moisture) under historical climate (average 2000-2020) and (B) maps of locations of experiments used for model fitting (yellow dots) and climate zones containing at least one experiment (filled polygons). Projections are shown only on agricultural area (cropland plus pastures) in the year 2000 and where climate conditions are reasonably similar to climate conditions in training dataset (see text for details). 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

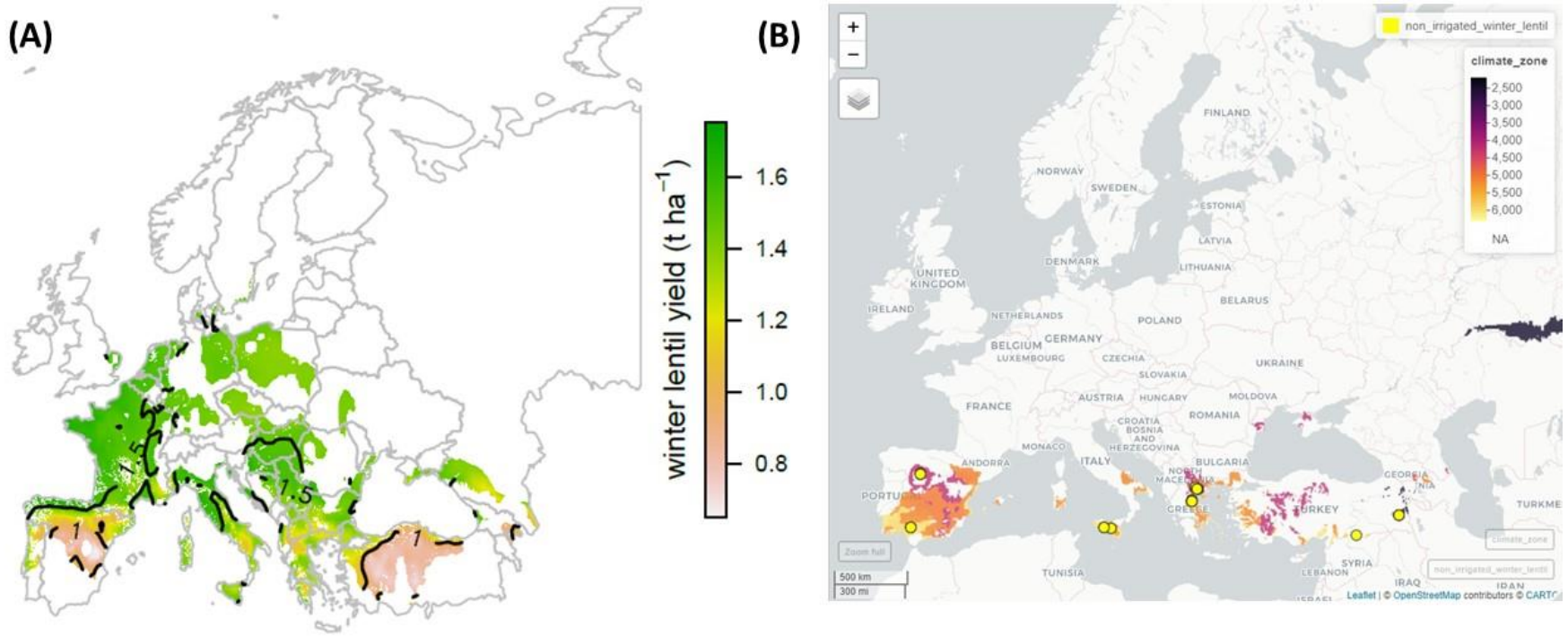

Figure 27. (A) Winter lentil projected yields (t ha^{-1} at 13% moisture) under historical climate (average 2000-2020) and (B) maps of locations of experiments used for model fitting (yellow dots) and climate zones containing at least one experiment (filled polygons). Projections are shown only on agricultural area (cropland plus pastures) in the year 2000 and where climate conditions are reasonably similar to climate conditions in training dataset (see text for details). 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

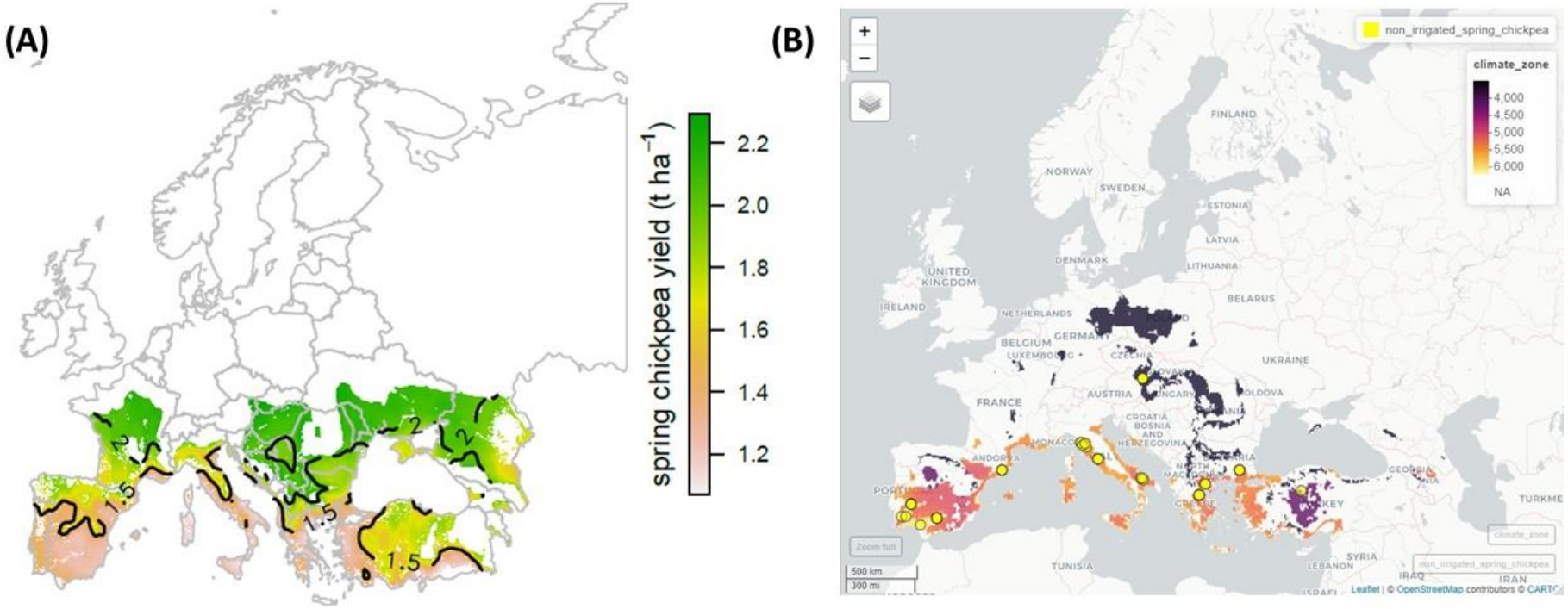

Figure 28. (A) Spring chickpea projected yields (t ha^{-1} at 13% moisture) under historical climate (average 2000-2020) and (B) maps of locations of experiments used for model fitting (yellow dots) and climate zones containing at least one experiment (filled polygons). Projections are shown only on agricultural area (cropland plus pastures) in the year 2000 and where climate conditions are reasonably similar to climate conditions in training dataset (see text for details). 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

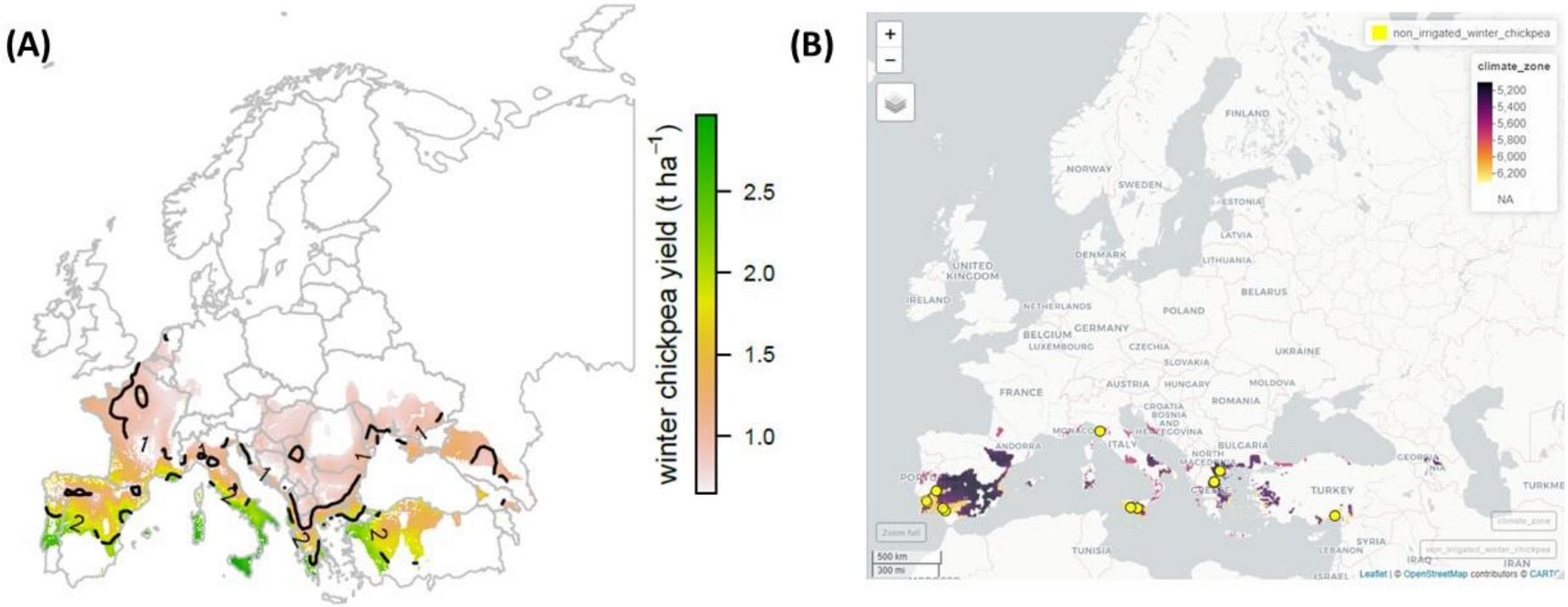

Figure 29. (A) Winter chickpea projected yields (t ha^{-1} at 13% moisture) under historical climate (average 2000-2020) and (B) maps of locations of experiments used for model fitting (yellow dots) and climate zones containing at least one experiment (filled polygons). Projections are shown only on agricultural area (cropland plus pastures) in the year 2000 and where climate conditions are reasonably similar to climate conditions in training dataset (see text for details). 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