

Data-driven projections suggest large opportunities to improve Europe's soybean self-sufficiency under climate change

Nicolas Guilpart ^{1*}, Toshichika Iizumi ², David Makowski ³

¹ Université Paris-Saclay, AgroParisTech, INRAE, UMR Agronomie, 78850, Thiverval-Grignon, France

² Institute for Agro-Environmental Sciences, National Agriculture and Food Research Organization (NARO), 3-1-3 Kannondai, Tsukuba, Ibaraki 305-8604, Japan

³ Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA-Paris, 75231, Paris, France

* Corresponding author: nicolas.guilpart@agroparistech.fr

Supplementary information

Content of the supplementary information

Supplementary figures for main text	2
Supplementary tables for main text.....	22
Supplementary Figures and Tables for methods	29
Supplementary details and discussion about the Random Forest model.....	35
Analysis of residuals	35
Variable importance	37
Additional references	38

Supplementary figures for main text

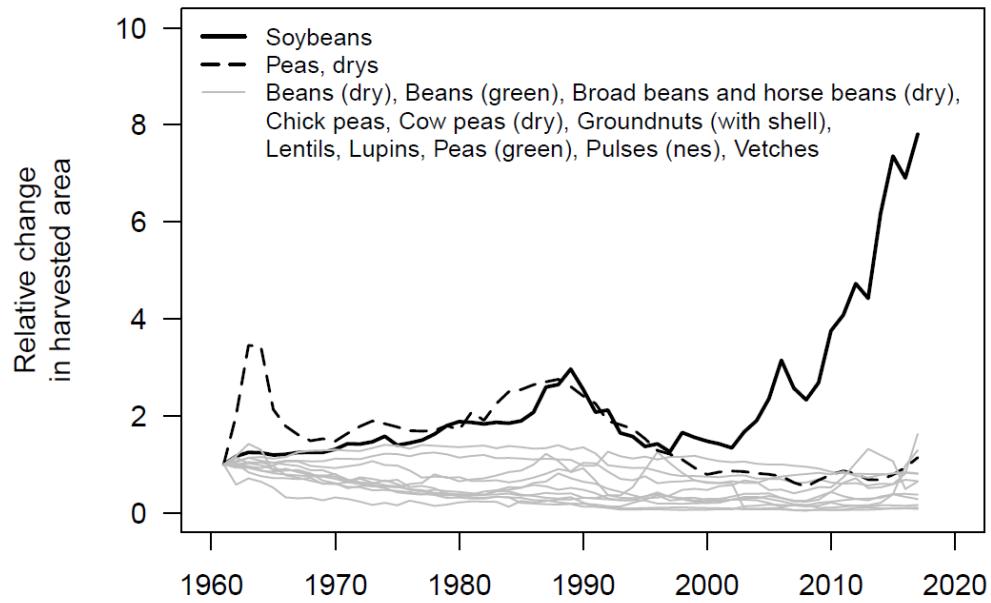


Figure S1. Historical trends in legumes harvested area in Europe from 1961 to 2017. Areas are expressed relatively to 1961 (set equal to 1). *Source: FAOSTAT¹.*

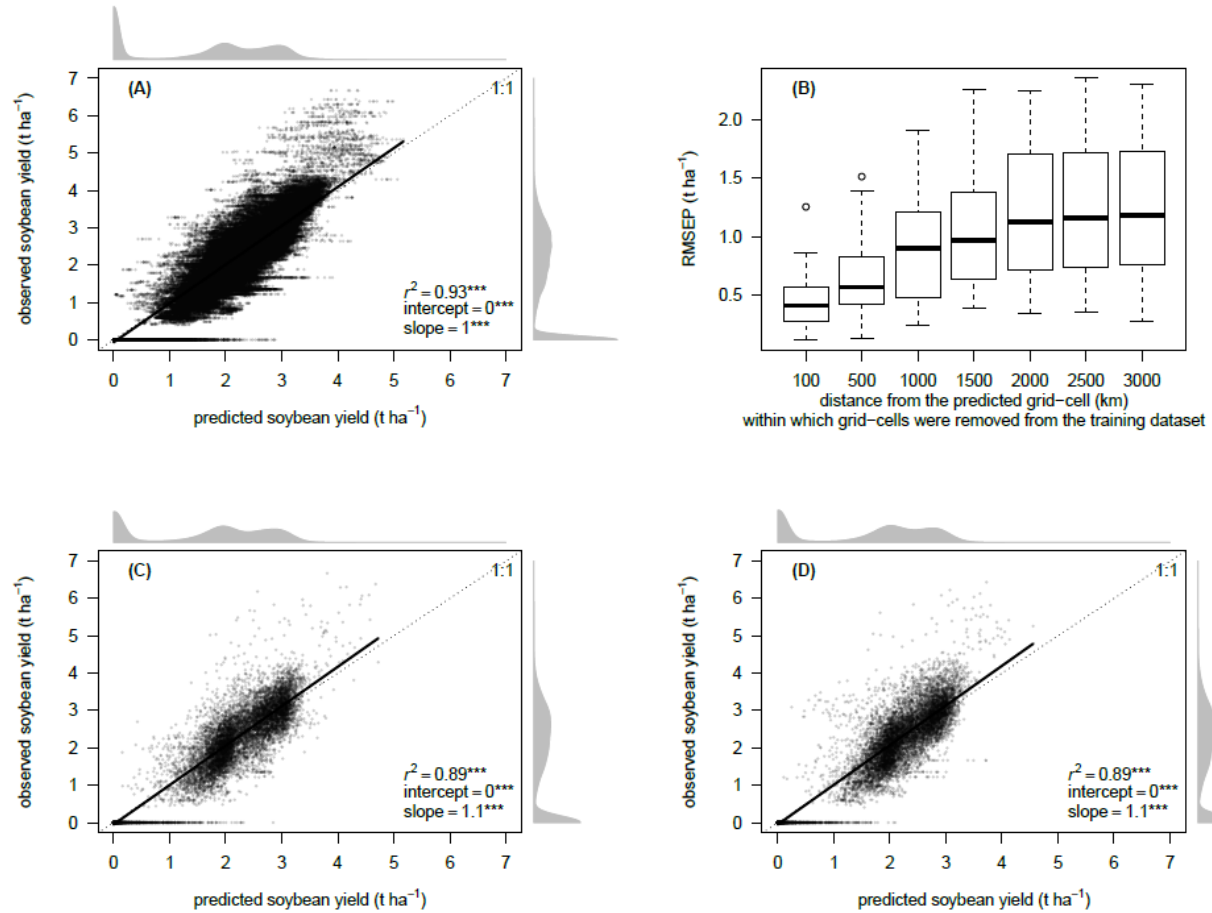


Figure S2. Assessment of the Random Forest algorithm. (A) The model is first evaluated using a classical bootstrap approach with 25 resamplings. (B) Model transferability in space is then evaluated by ensuring a minimum spatial distance between training and test datasets. Finally, model transferability 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.

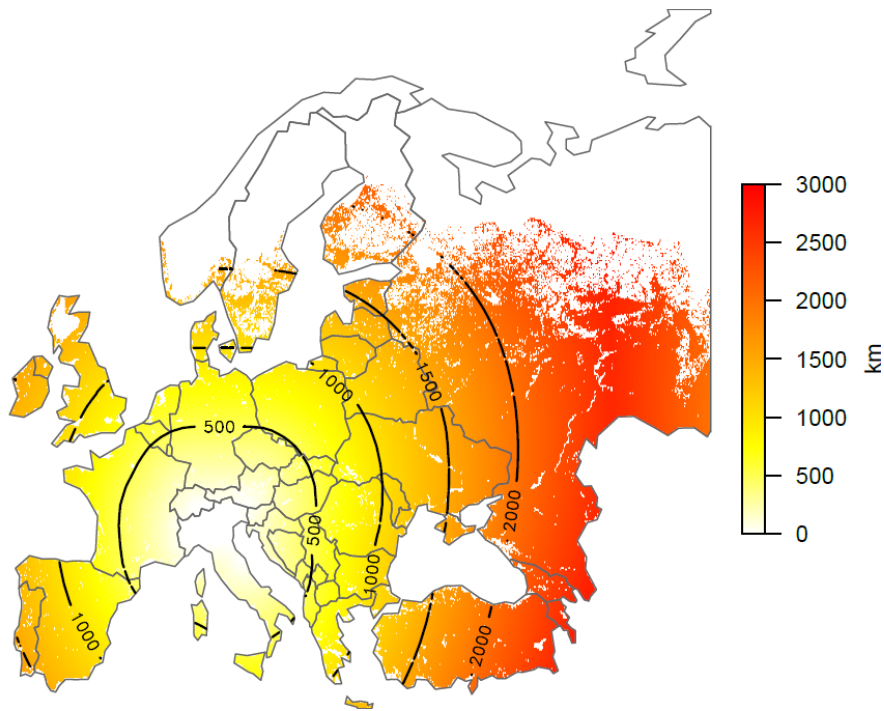


Figure S3. Distance to the nearest grid-cell with soybean used in the training dataset of the model. The Random Forest algorithm used to make soybean yield projections was trained using yield data from the Global Dataset of Historical Yields updated version^{2,3} in the following countries: Argentina, Brazil, Canada, China, India, Italy, and USA. The only yield data located in Europe in the training dataset are therefore from the north of Italy.

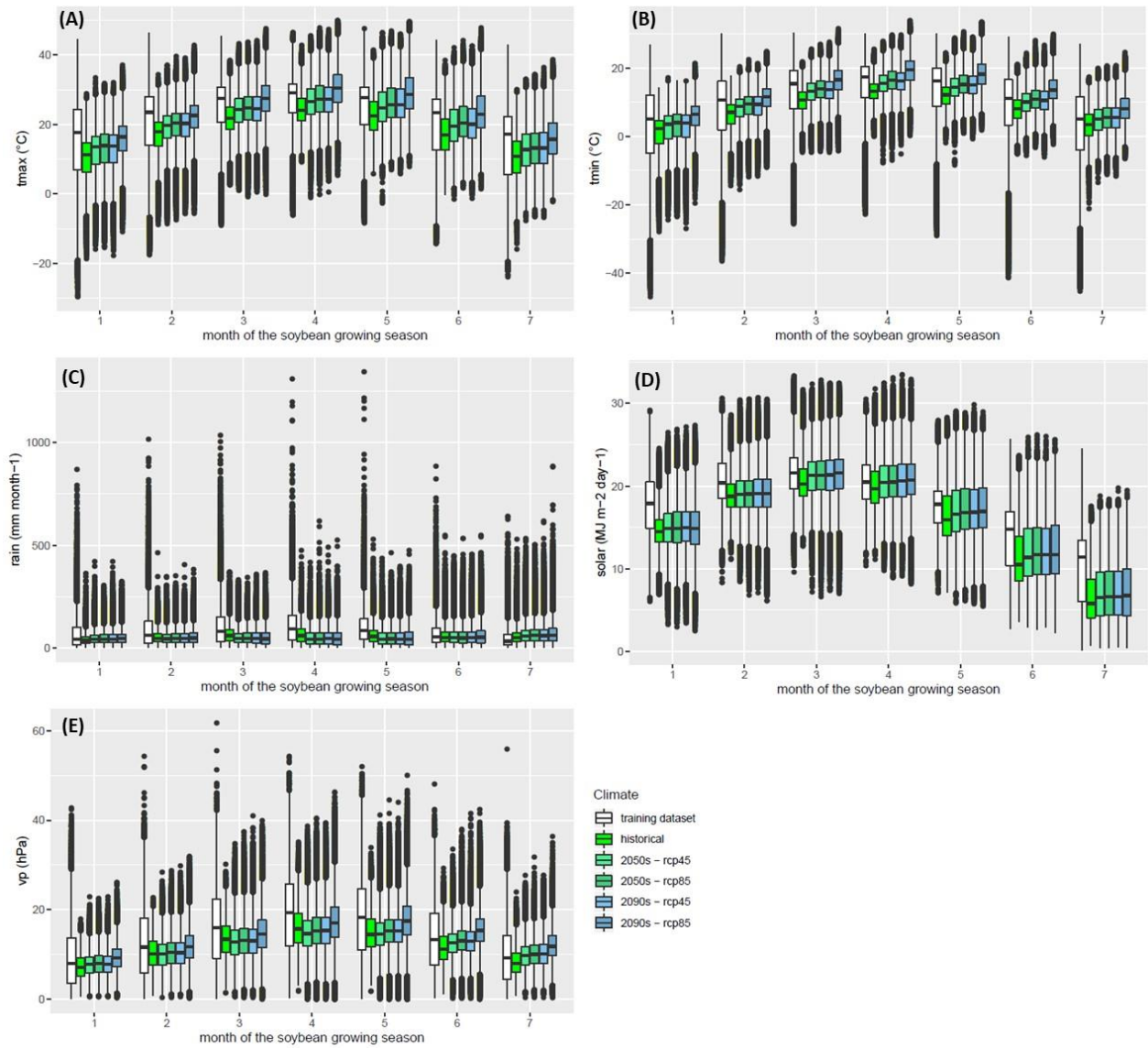


Figure S4. Comparison of observed ranges of climate variables in the dataset used for training the Random Forest algorithm and the datasets used to make soybean yield projections in Europe under historical and future climate scenarios. (A) Monthly average of daily maximum temperature. (B) Monthly average of daily minimum temperature. (C) Rainfall. (D) Solar radiation. (E) Vapour pressure. The GRASP dataset⁴ is used for model training at the global scale (“training dataset” in the color key) and to make projections of soybean yield in Europe under historical climate (1981-2010) (“historical” in the color key). Then eight Global Circulation Models⁵ (see methods) are considered for future climate scenarios, for two periods of time (2050s and 2090s) and two RCPs (RCP 4.5 and RCP 8.5).

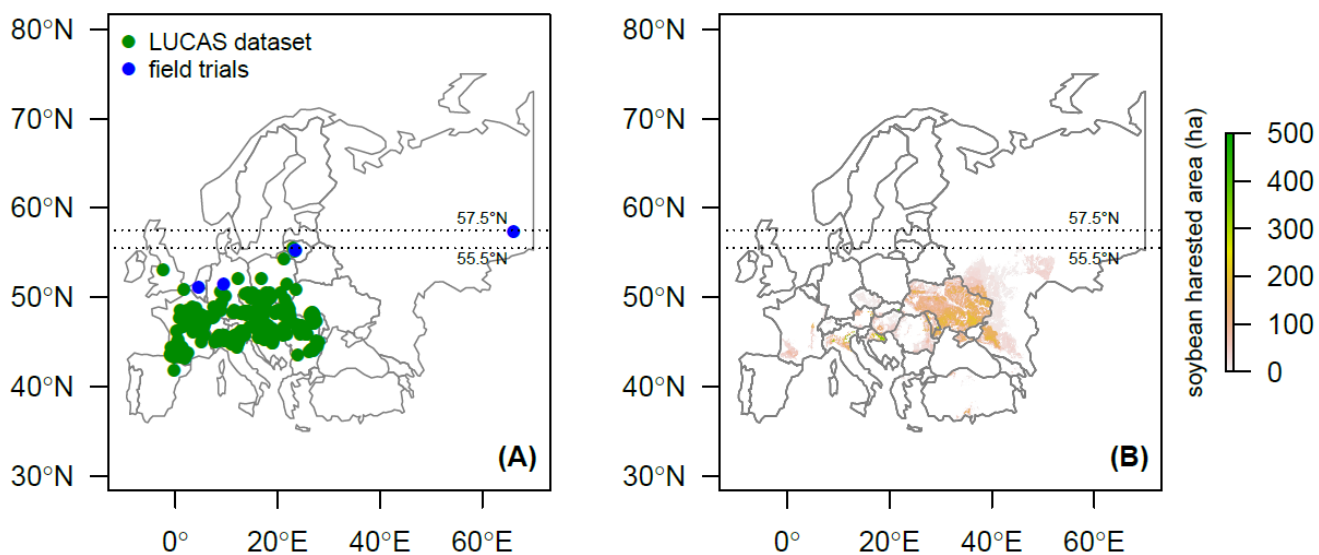


Figure S5. Evidence map of soybean cultivation at high latitude in Europe. (A) Locations of farmers soybean fields in 2018 reported by the Land Use and Coverage Area frame Survey (LUCAS) of the European Union (green dots), and locations of some selected published field experiments reporting satisfactory soybean yield (from 2.5 to 4 t ha⁻¹) at high latitudes in Europe (blue dots). See Table S4 for details about those field experiments. (B) Harvested soybean area in Europe according to the SPAM2010 model. Sources: the LUCAS dataset is available at <https://esdac.jrc.ec.europa.eu/projects/lucas> ; and the SPAM2010 dataset is available at: <https://www.mapspam.info/> .

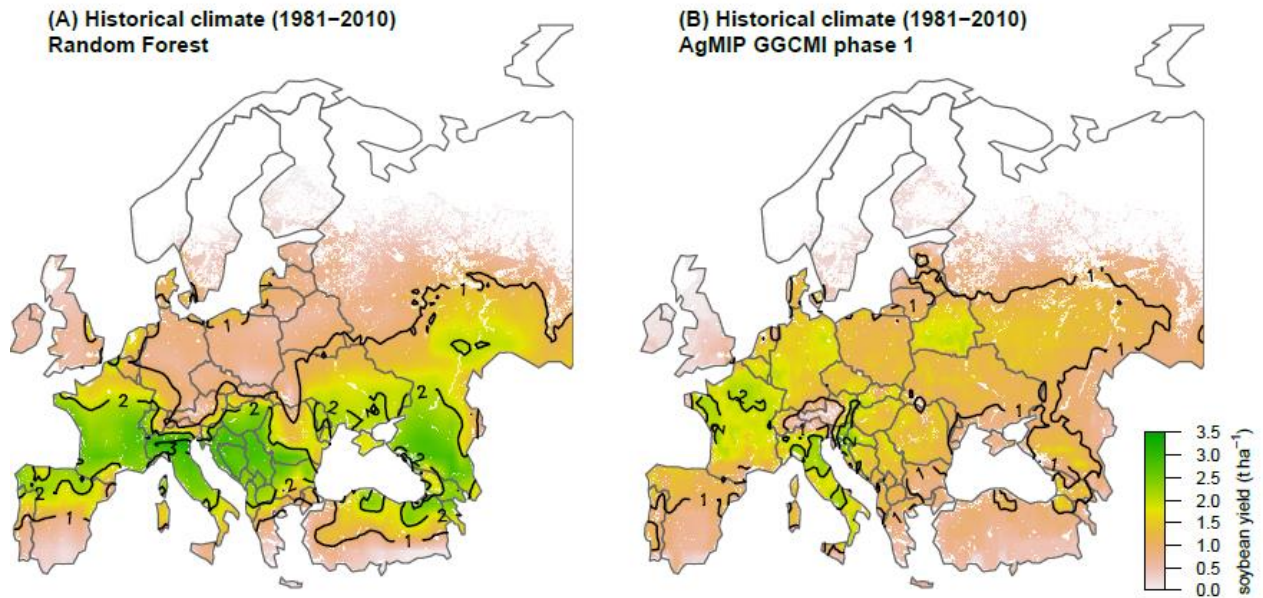


Figure S6. Historical simulated soybean yield (A) with Random Forest and (B) from AgMIP GGCM phase 1. Maps show median simulated yield over years for the Random Forest model, and over years and crop models for AgMIP GGCM phase 1. The following 8 crop models from AgMIP GGCM phase 1 were used: EPIC-BOKU, EPIC-IIASA, GEPIC, LPJmL, pAPSIM, pDSSAT, PEPIC, and PEGASUS (see Methods for details). Yield projections are shown only on agricultural area (cropland plus pasture), in the year 2000⁶.

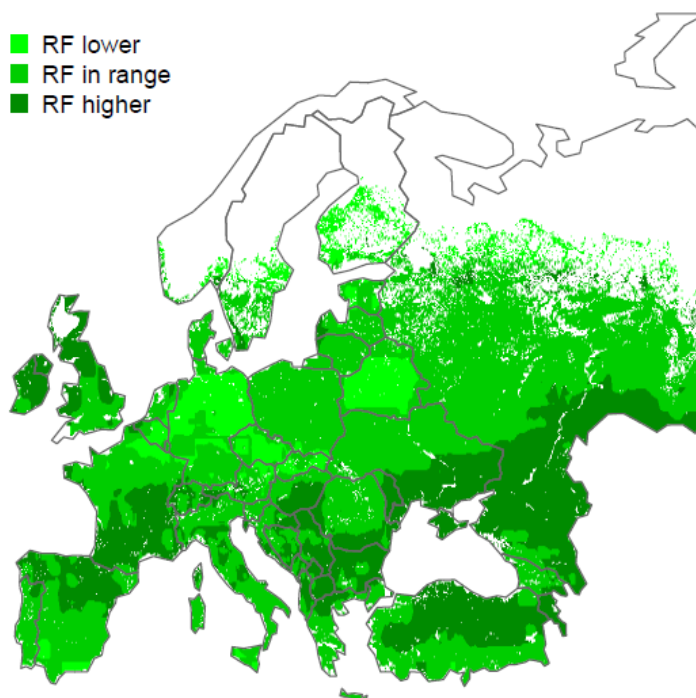


Figure S7. Map showing where soybean yield projection with RF are lower, higher, or within the range of AgMIP individual models simulations for the historical period (1981-2010). The following 8 crop models from AgMIP GGCM phase 1 were used: EPIC-BOKU, EPIC-IIASA, GEPIC, LPJmL, pAPSIM, pDSSAT, PEPIC, and PEGASUS (see Methods for details). Maps of AgMIP individual model simulations are shown on Figure S8. Yield projections are shown only on agricultural area (cropland plus pasture), in the year 2000⁶.

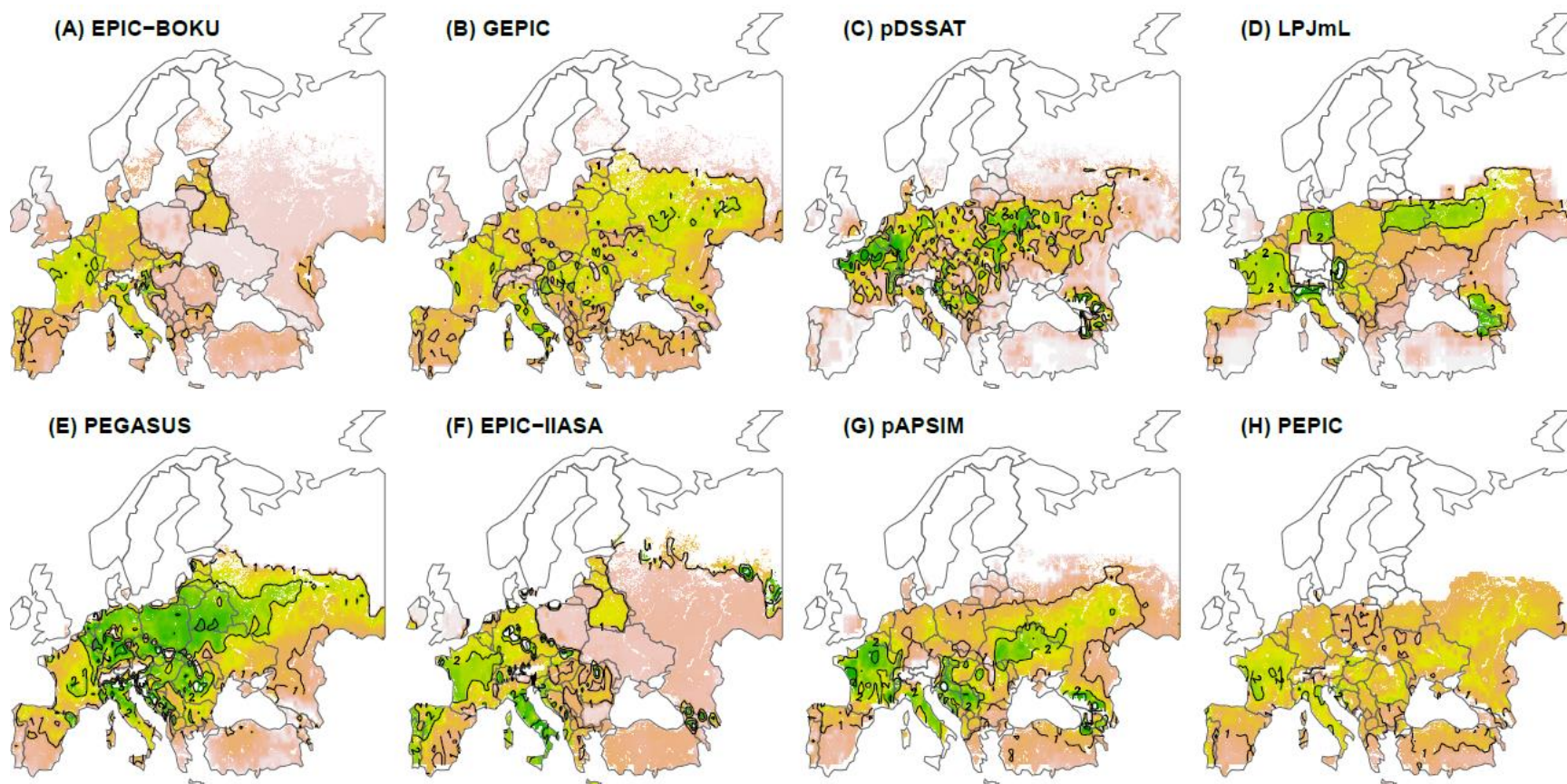


Figure S8. Maps of simulated historical soybean yield by the individual crop models from AgMIP GGCM1 phase 1 selected for the comparison with our Random Forest model projections. Maps show median simulated yield over years. Yield projections are shown only on agricultural area (cropland plus pasture), in the year 2000⁶.

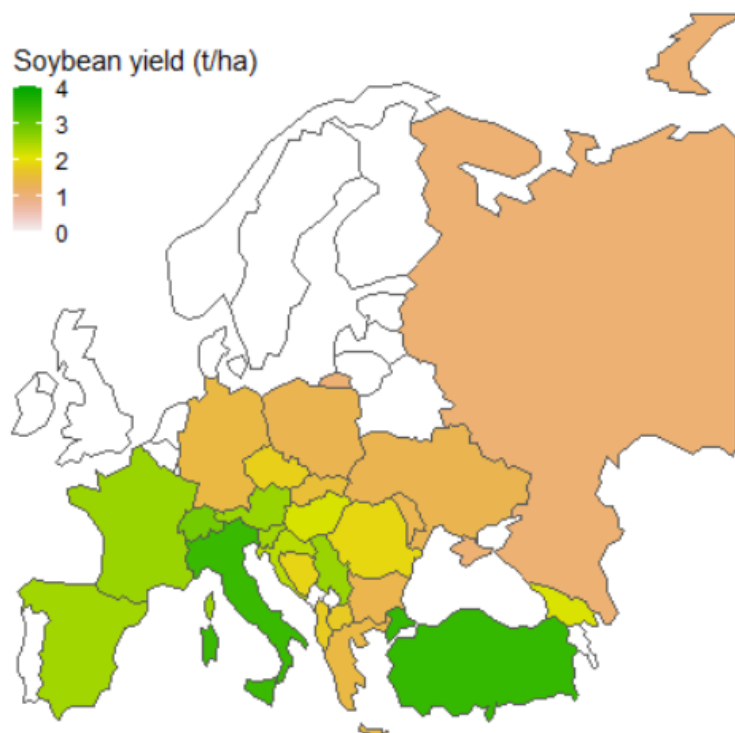


Figure S9. Actual soybean yield at the national level in Europe (average 2000-2010). *Source: FAOSTAT¹.*

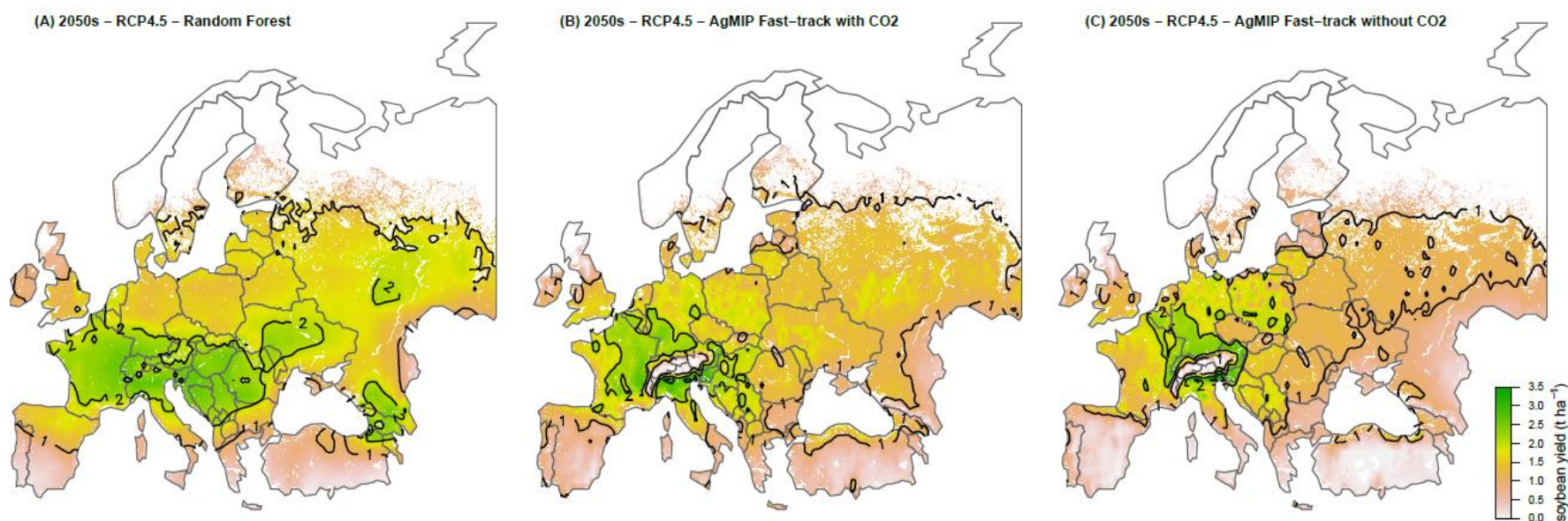


Figure S10. Simulated soybean yields by mid-century under RCP 4.5 (A) with Random Forest, (B) from AgMIP fast-track with CO₂, (C) from AgMIP fast-track without CO₂. Maps show median simulated yield over years (2050-2059) and global climate models for the Random Forest model, and over years, crop models and global climate model for AgMIP fast-track. Used crop and climate models from AgMIP fast-track are described in Table S11. Yield projections are shown only on agricultural area (cropland plus pasture), in the year 2000⁶.

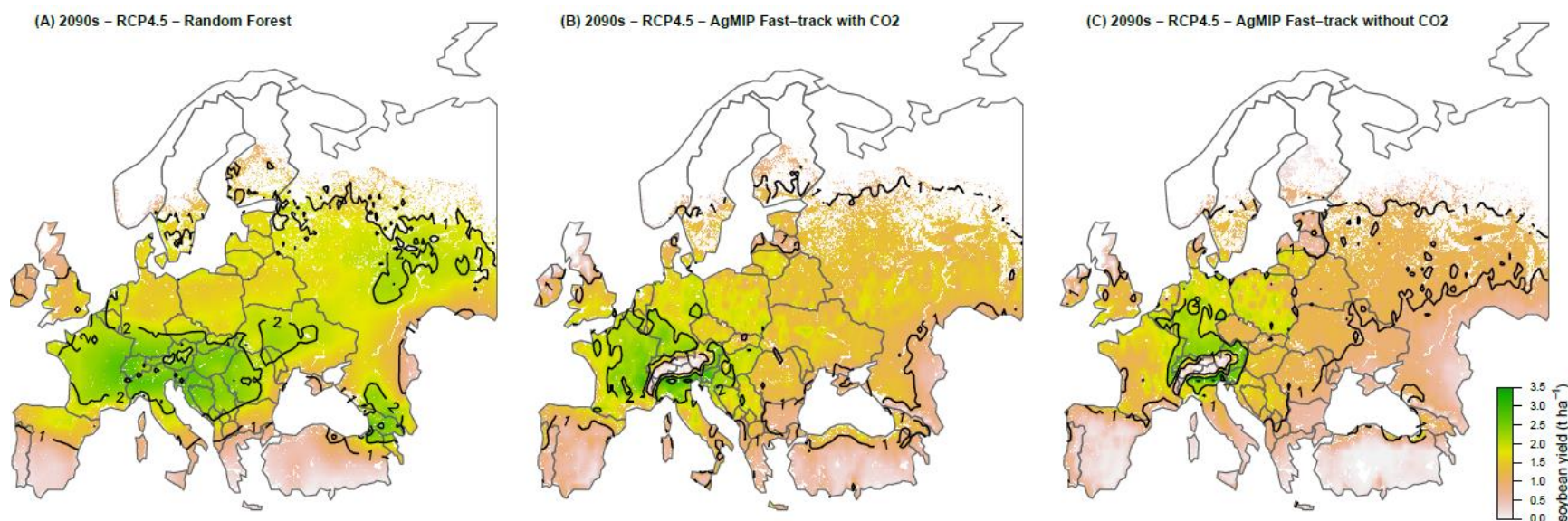


Figure S11. Simulated soybean yields by the end of the century under RCP 4.5 (A) with Random Forest, (B) from AgMIP fast-track with CO₂, (C) from AgMIP fast-track without CO₂. Maps show median simulated yield over years (2090-2099) and global climate models for the Random Forest model, and over years, crop models and global climate models for AgMIP fast-track. Used crop and climate models from AgMIP fast-track are described in Table S11. Yield projections are shown only on agricultural area (cropland plus pasture), in the year 2000⁶.

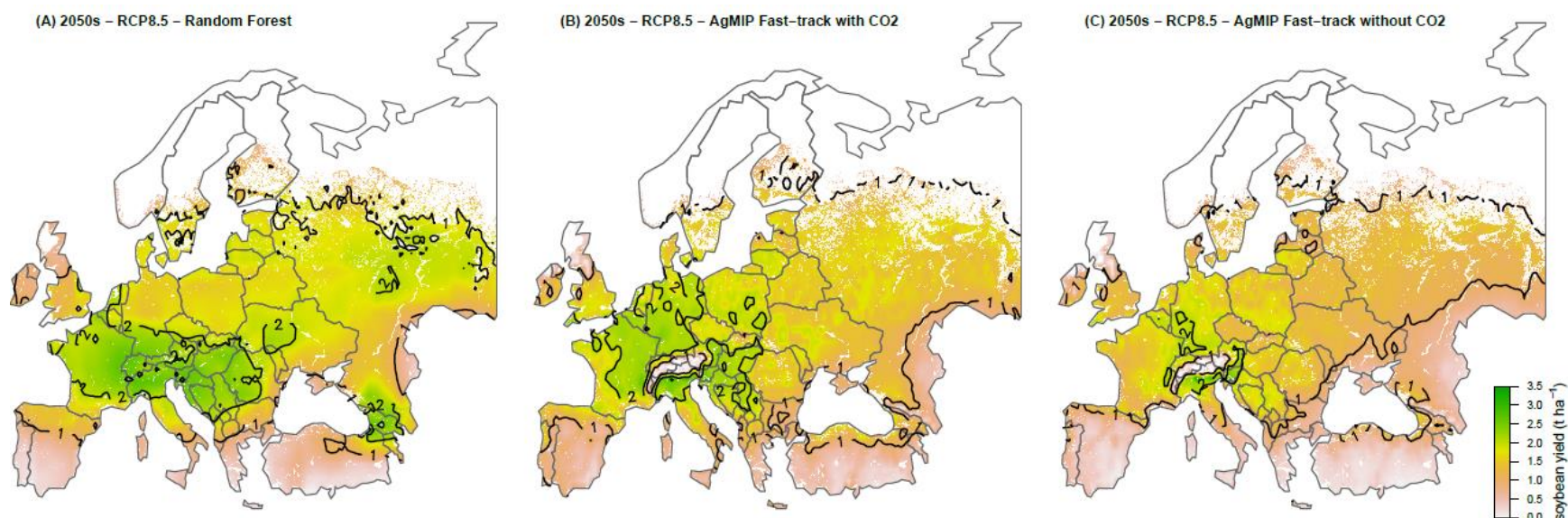


Figure S12. Simulated soybean yields by mid-century under RCP 8.5 (A) with Random Forest, (B) from AgMIP fast-track with CO₂, (C) from AgMIP fast-track without CO₂. Maps show median simulated yield over years (2050-2059) and global climate models for the Random Forest model, and over years, crop models and global climate models for AgMIP fast-track. Used crop and climate models from AgMIP fast-track are described in Table S11. Yield projections are shown only on agricultural area (cropland plus pasture), in the year 2000⁶.

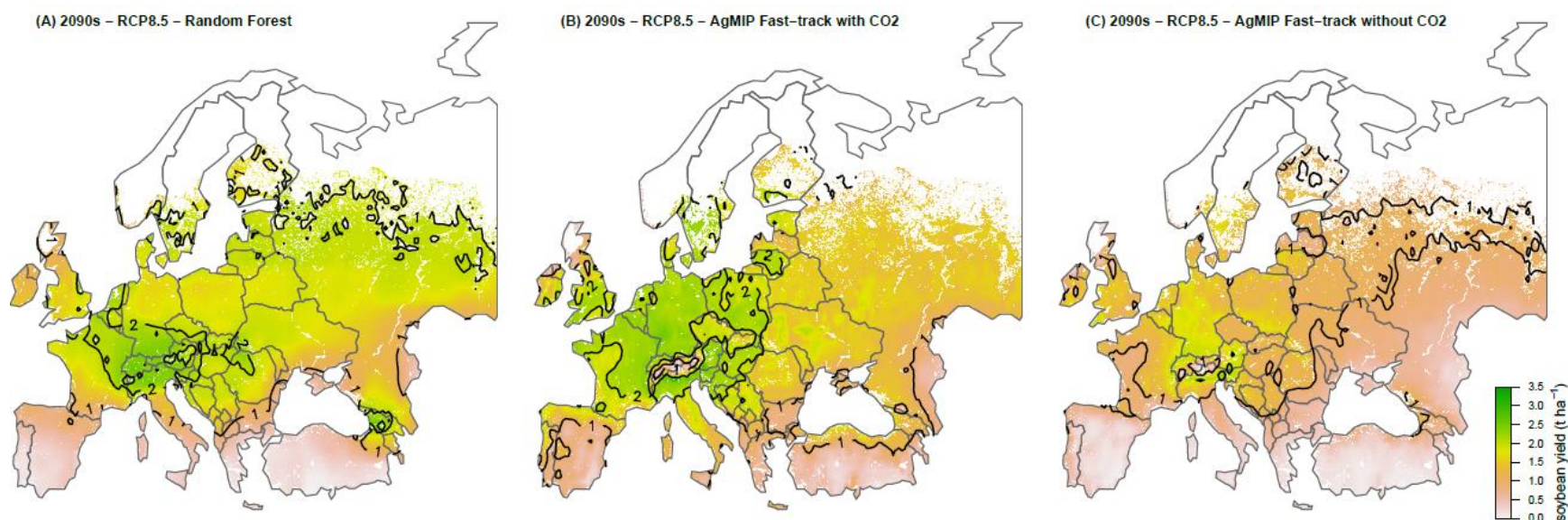


Figure S13. Simulated soybean yields by the end of the century under RCP 8.5 (A) with Random Forest, (B) from AgMIP fast-track with CO₂, (C) from AgMIP fast-track without CO₂. Maps show median simulated yield over years (2090-2099) and global climate models for the Random Forest model, and over years, crop models and global climate models for AgMIP fast-track. Used crop and climate models from AgMIP fast-track are described in Table S11. Yield projections are shown only on agricultural area (cropland plus pasture), in the year 2000⁶.

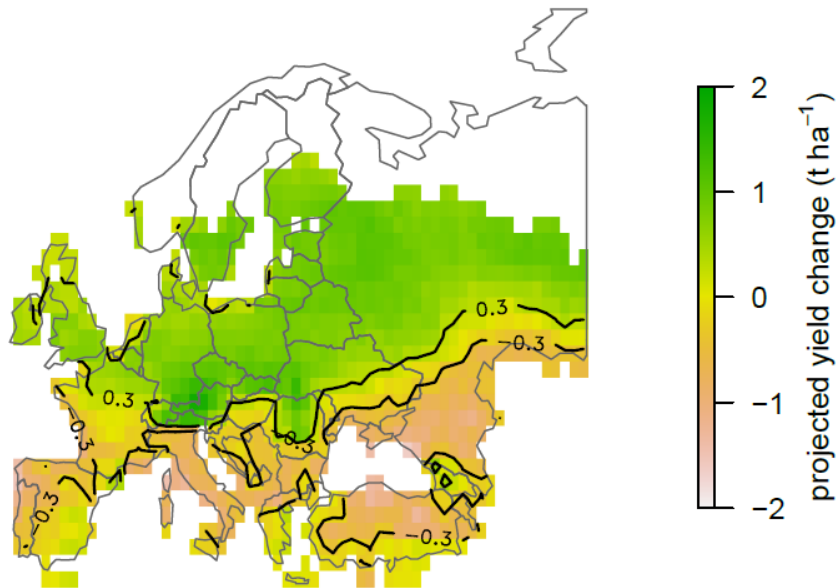


Figure S14. Projected soybean yield change between RCP 4.5 by mid-century and historical climate.

Projections are performed with the Random Forest algorithm and the GRASP dataset⁴ for historical climate (1981-2010), and 8 Global Circulation Models for future climate scenarios⁵. Grid-cells below -0.3 t ha^{-1} correspond to Group 1 (yield decrease) in the Linear Discriminant Analysis (see Figure 3 and Table 2), while grid-cells higher than $+0.3 \text{ t ha}^{-1}$ correspond to Group 2 (yield increase), and grid-cells between -0.3 and $+0.3 \text{ t ha}^{-1}$ correspond to Group 3 (marginal change).

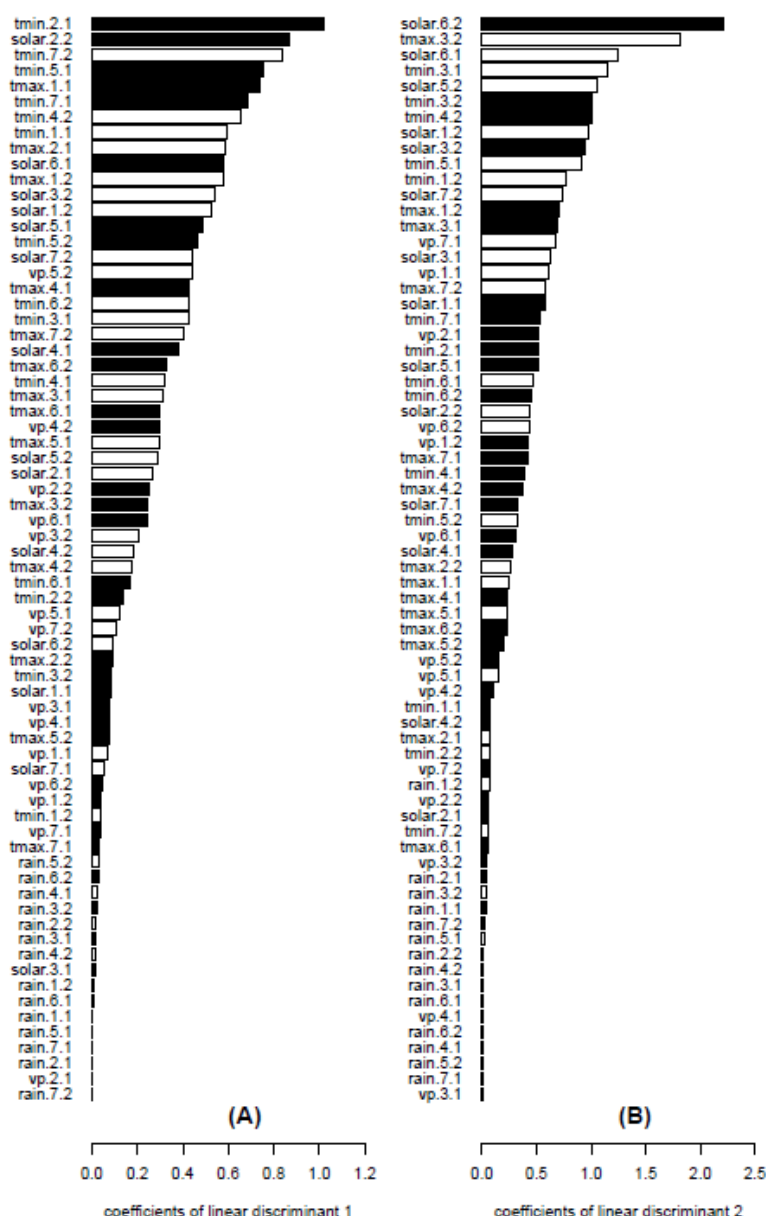


Figure S15. Climate variables contributions to linear discriminant 1 (panel A) and 2 (panel B). A linear discriminant analysis (LDA) was performed on climate variables for three groups of grid-cells defined by predicted soybean yield change between RCP 4.5 by 2050s and historical climate. Groups of grid-cells are defined as Group 1: yield decrease (projected yield change < - 0.3 t ha⁻¹), Group 2: yield increase (projected yield change > 0.3 t ha⁻¹), and Group 3: marginal change (yield change between -0.3 and +0.3 t ha⁻¹). White bars indicate a positive contribution, and black bars indicate a negative contribution. The higher the value, the higher the contribution of the corresponding climate input. Suffixes to climate variables names indicate, first, the month of the soybean growing season, and second, the time period ("1" standing for historical climate, and "2" standing for the 2050s under RCP 4.5). For example, "tmin.2.1" means "monthly average daily minimum temperature in the second month of the growing season under historical climate".

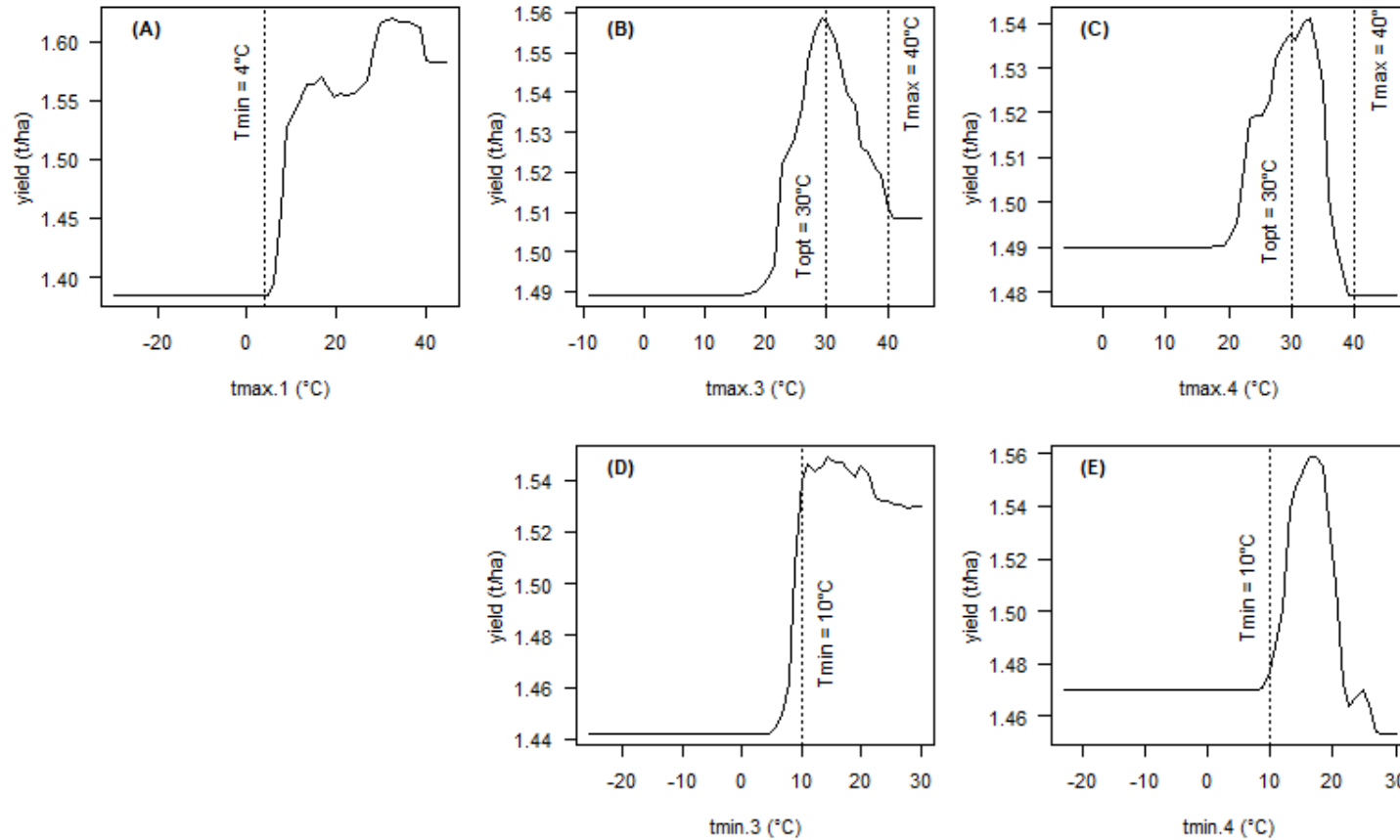


Figure S16. Partial dependence plots for selected variables of the Random Forest algorithm. (A) Daily maximum temperature in the first month of the growing season ($t_{max.1}$). (B) Daily maximum temperature in the third month of the growing season ($t_{max.3}$). (C) Daily maximum temperature in the fourth month of the growing season ($t_{max.4}$). (D) Daily minimum temperature in the third month of the growing season ($t_{min.3}$). (E) Daily minimum temperature in the third month of the growing season ($t_{min.3}$). Dotted lines indicate some cardinal temperatures extracted from the literature to show consistency between the Random Forest algorithm and the current knowledge of soybean physiology. Readers are referred to Table S7 for details about those cardinal temperatures.

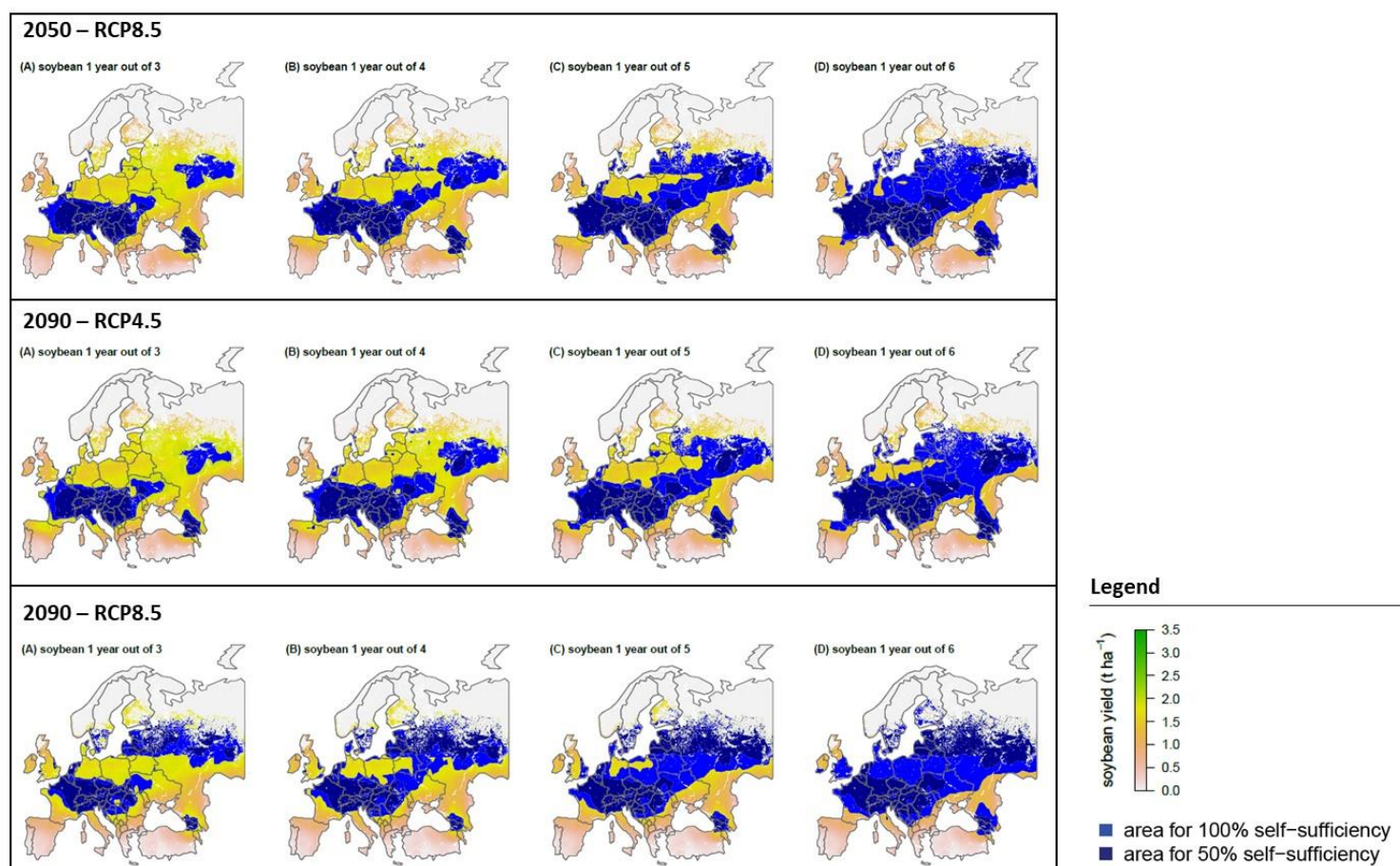


Figure S17. Area requirements for 50% and 100% soybean self-sufficiency in Europe by 2050s under RCP 4.5 and by 2090s under RCP 4.5 and 8.5. Based on soybean yield projections presented in Figure 1 and assuming various levels of soybean frequency in crop sequences (one year out for three, four, five 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 cropland⁶, which excludes permanent pastures in line with the Common Agricultural Policy of the European Union aiming at their protection⁷. Background colors indicate projected soybean yield in t ha⁻¹ as in Figure 1.

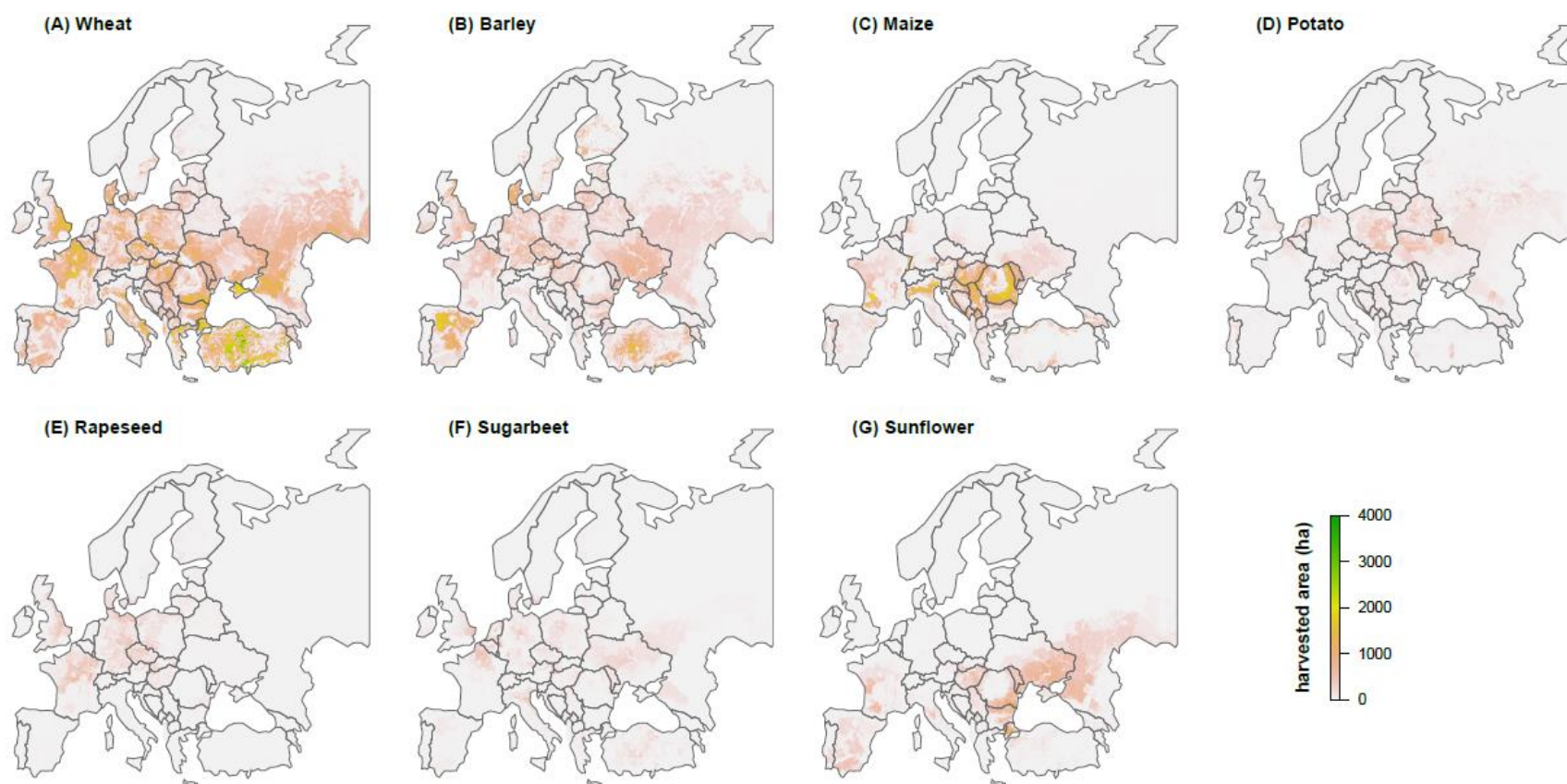


Figure S18. Harvested area maps of (A) wheat, (B) barley, (C) maize, (D) potato, (E) rapeseed, (F) sugarbeet, (G) sunflower in Europe around the year 2000⁸.

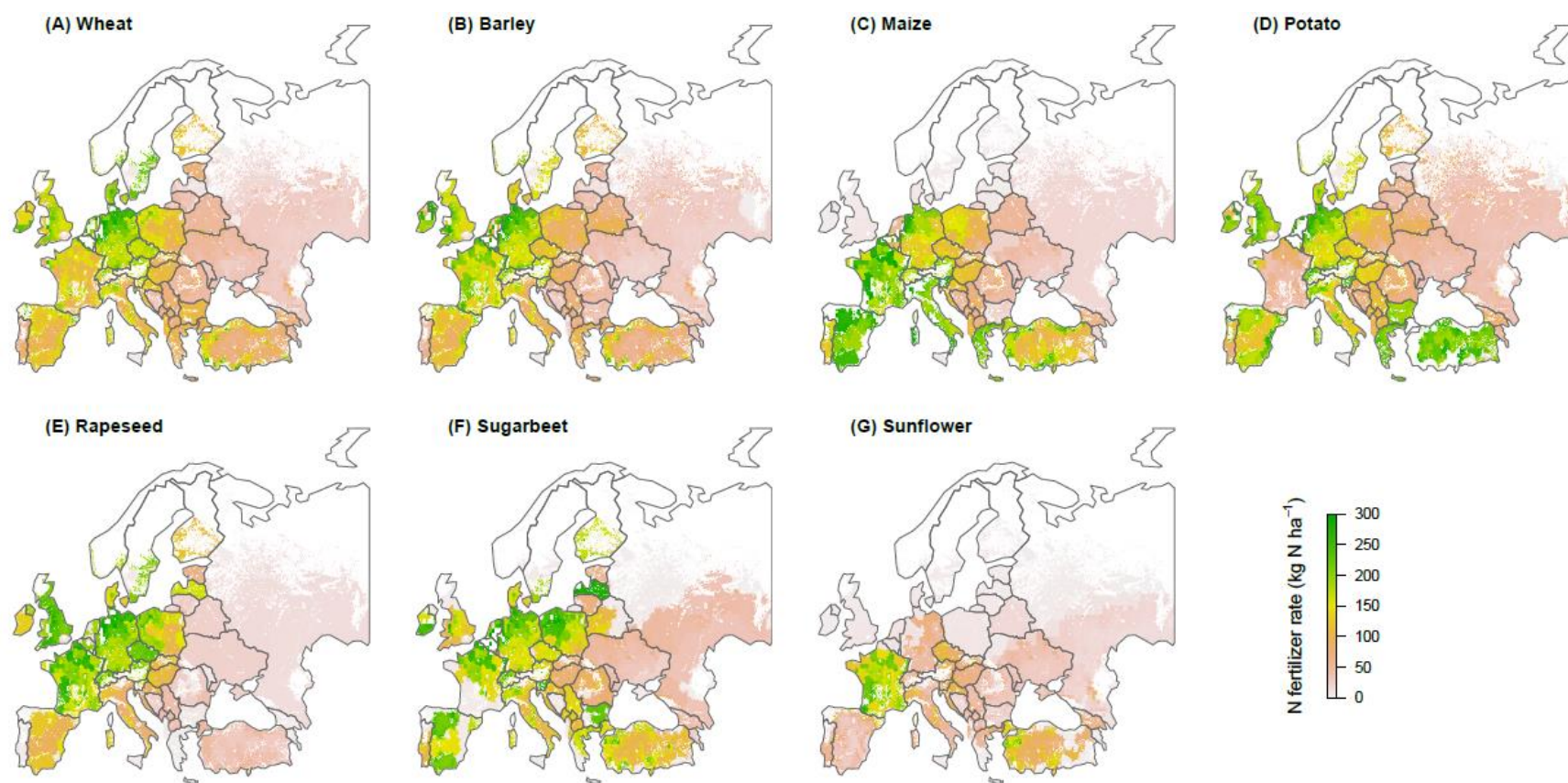


Figure S19. Crop specific nitrogen fertilizer rate for (A) wheat, (B) barley, (C) maize, (D) potato, (E) rapeseed, (F) sugarbeet, (G) sunflower in Europe around the year 2000⁹.

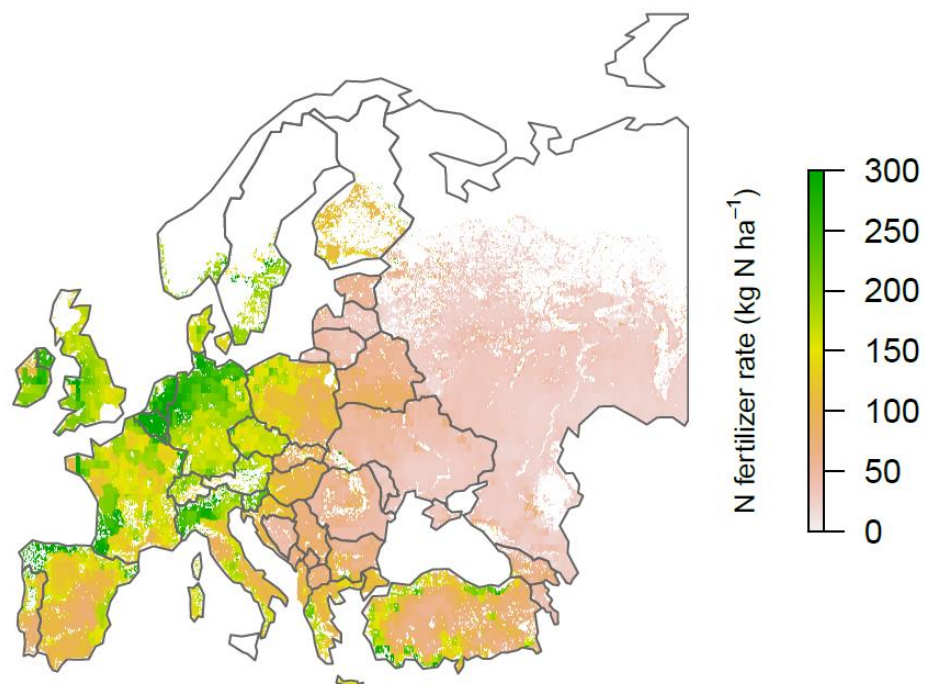


Figure S20. Average nitrogen fertilizer rate of major crops in Europe around the year 2000. Map shows the area-weighted nitrogen fertilizer rate based on crop specific harvested area and nitrogen fertilizer rate maps presented in Figure S18 and S19, respectively.

Supplementary tables for main text

Table S1. Soybean food balance in Europe (average 2007-2013). Soybean cake quantity is expressed in soybean grain equivalent assuming that 1 Mt of soybean grain gives 0.8 Mt of soybean cake. *Source: FAOSTAT¹.*

	Soybean cake	Soybean grain	Total (grain eq.)
	-----	Mt	-----
Domestic supply quantity	32	18	58
Domestic supply breakdown by source			
Production	12	5	20
Export Quantity	9	3	14
Import Quantity	29	16	52
Stock Variation	0	0	0
Domestic supply utilization			
Feed	32	2	42
Food	-	0	0
Processing	-	16	16
Losses	-	0	0
Seed	-	0	0
Other uses	0	0	0

Table S2. Changes in soybean suitable area due to climate change in Europe. All values are in Mha. Percentage change relative to historical climate is indicated in parenthesis.

Climate scenario		Area with predicted soybean yield		
		$\geq 1.5 \text{ t ha}^{-1}$	$\geq 2 \text{ t ha}^{-1}$	$\geq 2.5 \text{ t ha}^{-1}$
Historical climate (1981-2010)		207	106	37
2050s	RCP 4.5	279 (+35%)	95 (-10%)	10 (-73%)
	RCP 8.5	273 (+32%)	84 (-21%)	8 (-78%)
2090s	RCP 4.5	294 (+42%)	102 (-4%)	13 (-65%)
	RCP 8.5	262 (+27%)	45 (-58%)	0.1 (-100%)

Table S3. Percentage of climate data used for soybean yield projections that fall outside of the observed range of climate data used for model training. Rain: rainfall. Solar: solar radiation. Tmax: daily maximum temperature. Tmin: daily minimum temperature. VP: vapour pressure. For example, 0.13 % of the solar radiation data in month 4 of the soybean growing season under historical climate for the period 1981-2010 fall outside of the observed range for that variable in the dataset used for model training.

	Month*	Rain	Solar	Tmax	Tmin	VP
Historical climate						
1981-2010	1	0	0	0	0	0
	2	0	0	0	0	0
	3	0	0	0	0	0
	4	0	0.13	0	0	0
	5	0	0.03	0	0	0
	6	0	0	0	0	0
	7	0	0	0	0	0
RCP 4.5						
2050s	1	0	0.10	0	0	0
	2	0	0	0	0	0
	3	0	0.01	0	0	0
	4	0	0.35	0	0	0.01
	5	0	0.13	0	0	0
	6	0	0.00	0	0	0
	7	0	0	0	0	0
2090s	1	0	0.14	0	0	0
	2	0	0.01	0	0	0
	3	0	0.02	0	0	0
	4	0	0.41	0	0	0.01
	5	0	0.18	0	0	0
	6	0	0.00	0	0	0
	7	0	0	0	0	0
RCP 8.5						
2050s	1	0	0.15	0	0	0
	2	0	0.01	0	0	0
	3	0	0.04	0	0	0
	4	0	0.33	0.01	0	0.01
	5	0	0.14	0	0	0.01
	6	0	0.00	0	0	0
	7	0	0	0	0	0
2090s	1	0	0.34	0	0	0
	2	0	0.05	0	0	0
	3	0	0.06	0.03	0.01	0
	4	0	0.35	0.18	0.20	0.05
	5	0	0.13	0.06	0.27	0.02
	6	0	0	0.02	0.01	0.01
	7	0	0	0	0	0

*month of the soybean growing season (April to October in Europe)

Table S4. Details of the selected published soybean field experiments at high latitude in Europe reported in Figure S5.

Latitude	Country	Reported soybean yield	References
57.35 ° N	Russia	4 t ha ⁻¹	Kühling et al. (2018). Soybeans in high latitudes : effects of Bradyrhizobium inoculation in Northwest Germany and southern West Siberia. <i>Organic Agriculture</i> , 8(2), 159–171. ¹⁰
55.24 ° N	Lithuania	2.5 - 3 t ha ⁻¹	Kadžilienė et al. (2016). Legumes for sustainability of agroecosystems. In Z. Gaile & Š. Dace (Eds.), 20th Baltic Agronomy Forum (p. 56). Jelgava, Latvia. ¹¹
51.40 ° N	Germany	2.5 - 3 t ha ⁻¹	Zimmer et al. (2016). Effects of soybean variety and Bradyrhizobium strains on yield, protein content and biological nitrogen fixation under cool growing conditions in Germany. <i>European Journal of Agronomy</i> , 72, 38–46. ¹²
51.10 ° N	Belgium	2.5 - 3 t ha ⁻¹	Pannecouque et al. (2018). Screening for soybean varieties suited to Belgian growing conditions based on maturity, yield components and resistance to <i>Sclerotinia sclerotiorum</i> and <i>Rhizoctonia solani</i> anastomosis group 2-2IIIB. <i>Journal of Agricultural Science</i> , 1–8. ¹³

Table S5. Confusion matrix of the Linear Discriminant Analysis. The linear discriminant analysis was performed on climate on climate variables for three groups of grid-cells defined by predicted soybean yield change between RCP 4.5 by 2050s and historical climate. Groups are defined as Group 1: yield decrease (projected yield change < - 0.3 t ha⁻¹), Group 2: yield increase (projected yield change > 0.3 t ha⁻¹). Group 3: marginal change (yield change between -0.3 and +0.3 t ha⁻¹). Overall accuracy (fraction of correct predictions) is 89%.

Actual Group	Predicted group		
	Group 1	Group 2	Group 3
Group 1	207	1	43
Group 2	0	491	20
Group 3	26	25	188

Table S6. Analysis of climate variables associated with a decrease (Group 1), increase (Group 2), and marginal change (Group 3) in projected soybean yield under RCP 4.5 by mid-century relative to historical climate. Reported values of climate variables represent mean values for each group of grid-cells calculated with a Linear Discriminant Analysis (Figure 3). Groups of grid-cells are defined as Group 1: yield decrease (projected yield change < - 0.3 t ha⁻¹), Group 2: yield increase (projected yield change > 0.3 t ha⁻¹), and Group 3: marginal change (yield change between -0.3 and +0.3 t ha⁻¹). The GRASP dataset⁴ is used for historical climate, and the median over height Global Circulation Models⁵ is shown for RCP 4.5 by mid-century. Yield projections are performed with the Random Forest algorithm presented in Table 1 and Figure S2.

Climate variable	Month*	Group 1 (yield decrease)		Group 2 (yield increase)		Group 3 (marginal change)	
		Historical climate	2050s	Historical climate	2050s	Historical climate	2050s
Tmax (°C)	1	15,3	16,5	10,6	13,1	15,2	16,7
	2	21,1	22,6	17,6	19,6	20,9	22,4
	3	25,5	27,9	21,3	23,8	25,0	27,5
	4	28,6	31,3	23,4	25,9	27,8	30,4
	5	28,0	30,9	21,7	24,2	27,1	29,9
	6	23,4	26,1	16,5	18,9	22,8	25,3
	7	16,8	18,7	10,1	12,1	16,6	18,6
Tmin (°C)	1	4,9	5,8	1,7	3,2	5,0	6,1
	2	9,7	11,1	6,9	8,8	9,5	10,9
	3	13,3	15,8	10,5	13,2	13,1	15,5
	4	15,9	18,3	13,0	15,4	15,5	17,8
	5	15,3	18,0	11,7	14,0	15,0	17,5
	6	11,5	13,8	7,9	9,7	11,5	13,6
	7	7,0	8,3	3,3	4,9	7,1	8,6
Rain (mm month ⁻¹)	1	48,1	42,8	37,5	44,9	44,9	43,4
	2	49,2	41,4	53,7	48,0	44,5	41,9
	3	49,4	31,8	72,2	52,2	48,9	35,4
	4	31,6	19,3	75,1	52,1	39,3	26,9
	5	32,1	15,7	68,5	52,6	39,3	25,9
	6	42,6	27,6	56,9	56,9	45,7	36,4
	7	51,4	47,6	53,5	64,9	54,3	53,1
Solar (MJ m ⁻² day ⁻¹)	1	16,9	17,6	14,2	14,5	16,7	17,4
	2	21,0	21,4	18,4	18,6	20,7	21,1
	3	23,4	23,9	19,4	20,8	22,9	23,4
	4	23,7	24,2	18,9	19,8	22,9	23,4
	5	20,8	21,4	15,5	16,1	19,9	20,5
	6	15,8	16,6	10,3	11,0	15,1	15,8
	7	10,2	11,0	5,6	6,3	9,7	10,4
VP (hPa)	1	9,2	9,1	7,4	7,9	9,3	9,3
	2	12,4	12,1	10,7	10,3	12,3	12,0
	3	15,5	14,7	14,2	13,1	15,4	14,5
	4	17,8	16,5	16,3	15,1	17,8	16,6
	5	17,0	15,7	15,2	15,0	17,3	16,2
	6	13,9	14,3	11,7	12,8	14,1	14,7
	7	10,3	11,8	8,4	9,9	10,8	12,2

* month of the soybean growing season (April to October in Europe)

Table S7. Soybean cardinal temperatures. All values are expressed in degree Celsius (°C).

Process	Tmin	Topt	Tmax	References
Germination	4	30	40	¹⁴
Pollen germination	10-13	28.5 - 30	47	^{15,16}
Leaf photosynthesis	5	36	50	^{17,18}
Crop development (phenology)				
Pre-anthesis	5-7.6	30-31.5	40-45	¹⁹
Post-anthesis	3.6-6	23-26	39-40	¹⁹⁻²³

Table S8. Harvested area of wheat, barley, maize, potato, sunflower, rapeseed, and sugarbeet within the area needed to achieve 50% and 100% soybean self-sufficiency as identified by the random forest yield projections.

Scenario	Wheat (%)	Barley (%)	Maize (%)	Potato (%)	Sunflower (%)	Rapeseed (%)	Sugarbeet (%)	Total (Mha)
----- 100 % self-sufficiency -----								
Historical climate								
Soybean 1 year in 3	40	16	22	3	15	2	2	37
Soybean 1 year in 4	40	18	19	3	15	2	3	51
Soybean 1 year in 5	42	19	17	4	14	2	3	67
Soybean 1 year in 6	43	20	14	5	11	2	4	92
2050s – RCP 4.5								
Soybean 1 year in 3	37	16	24	5	9	4	4	40
Soybean 1 year in 4	38	17	20	6	10	4	4	52
Soybean 1 year in 5	39	19	17	7	10	3	4	64
Soybean 1 year in 6	39	20	15	8	10	3	4	80
2050s RCP 8.5								
Soybean 1 year in 3	40	16	22	6	7	5	4	39
Soybean 1 year in 4	40	18	19	8	7	4	5	49
Soybean 1 year in 5	40	20	17	8	7	4	5	61
Soybean 1 year in 6	40	21	14	9	8	4	4	79
----- 50 % self-sufficiency -----								
Historical climate								
Soybean 1 year in 3	39	10	31	3	12	2	3	16
Soybean 1 year in 4	41	11	27	3	13	2	2	23
Soybean 1 year in 5	41	14	24	3	13	2	2	30
Soybean 1 year in 6	40	16	22	3	15	2	2	38
2050s – RCP 4.5								
Soybean 1 year in 3	36	13	31	4	8	5	4	21
Soybean 1 year in 4	36	14	29	4	8	5	4	27
Soybean 1 year in 5	36	15	26	4	9	4	4	34
Soybean 1 year in 6	37	16	24	5	9	4	4	41
2050s – RCP 8.5								
Soybean 1 year in 3	37	14	29	4	7	6	4	21
Soybean 1 year in 4	38	15	26	4	7	5	4	28
Soybean 1 year in 5	39	16	24	5	7	5	4	33
Soybean 1 year in 6	40	17	22	7	7	4	4	40

Supplementary Figures and Tables for methods

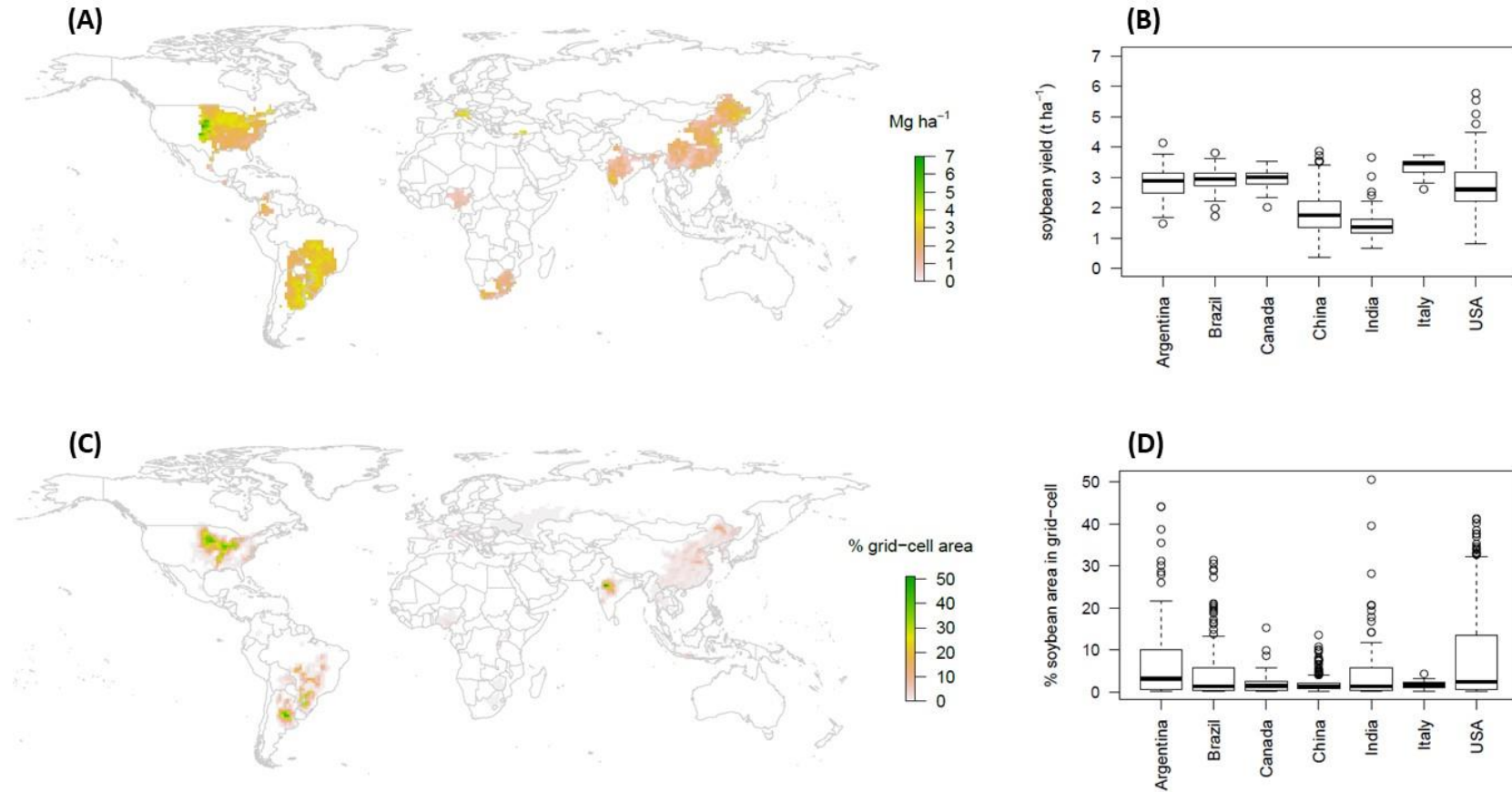


Figure S21. Soybean yield and harvested area from the global dataset of historical yields². (A) Global map of soybean yield in 2010. (B) Soybean yield in 2010 in selected countries. (C) Global map of soybean harvested area. (D) Soybean harvested area in selected countries.

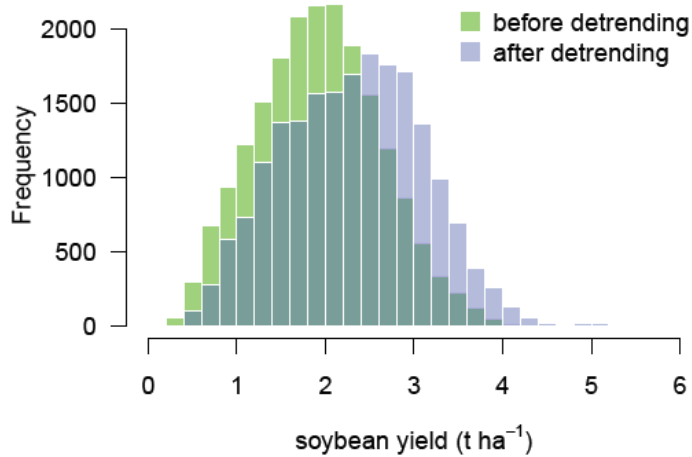


Figure S22. Histograms of soybean yield before and after detrending. Detrending is performed in order to remove the increasing trends of soybean yield time series due to improved cultivars and technological progress. Data shown in this Figure are for Argentina, Brazil, Canada, China, India, Italy, and USA.

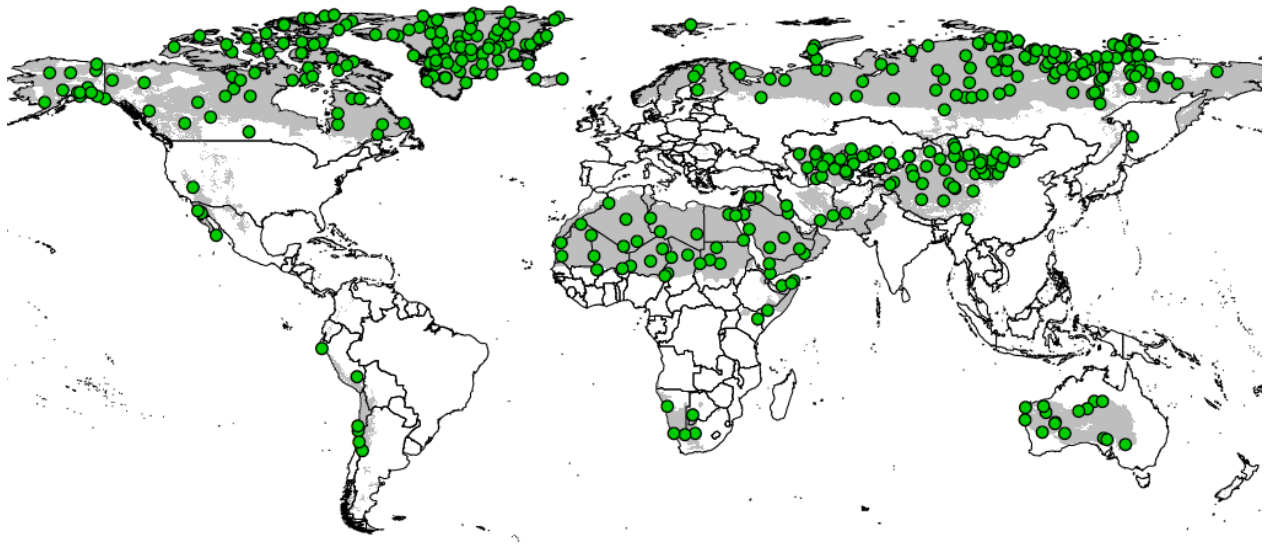


Figure S23. Map showing locations of true absences (soybean yield equals zero) added to the historical yield dataset. These data points (green dots) representing true absences of soybean were added to the dataset in order to extend the range of climate conditions captured by the model and to capture climate conditions leading to zero yield. These 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. Grey zones indicate climate zones from the Köppen-Geiger climate classification^{24,25} in which the true absences were randomly selected. See Table S9 for a short description of these climate zones.

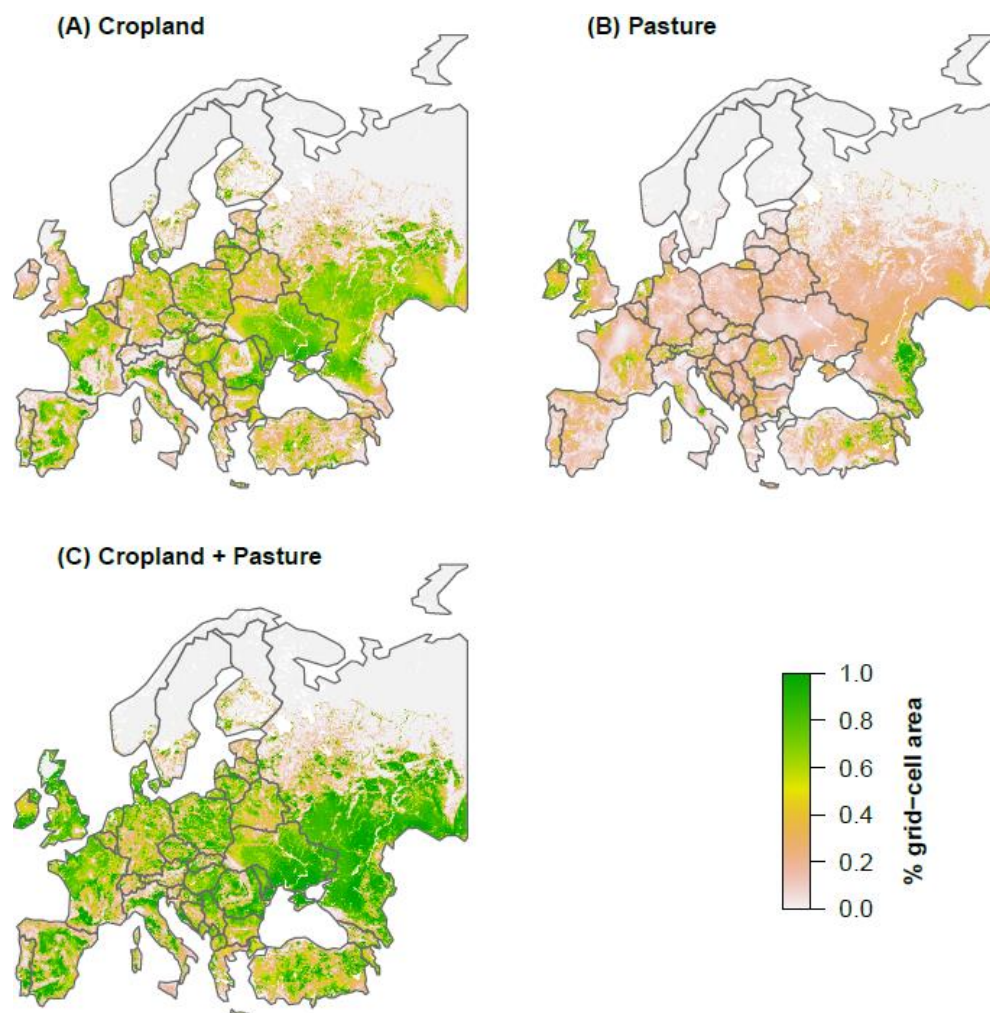


Figure S24. Cropland (A), pasture (B), and total agricultural area (C) maps in Europe around the year 2000⁶. Total agricultural area is calculated as the sum of cropland plus pasture areas.

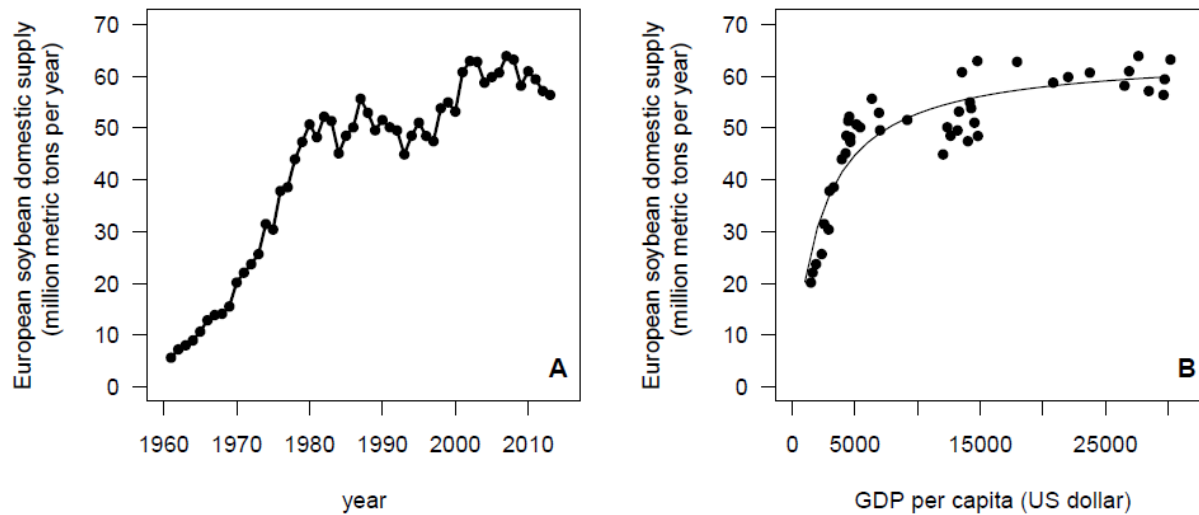


Figure S25. European soybean domestic supply as a function of year (A) and GDP per capita (B). The soybean domestic supply includes soybean grains and soybean cakes and is expressed in grain equivalent assuming that 1 kg of soybean grain gives 0.8 kg of soybean cakes. In panel B, the soybean domestic supply was modeled as a function of GDP per capita as follows: $y \sim (a * x) / (b + x)$ where y is the soybean domestic supply and x the GDP per capita. This model was fitted using the *nls()* function in R. Estimated values of parameters are $a = 64$ ($p < 0.001$) and $b = 2168$ ($p < 0.001$). All data are from FAOSTAT.

Table S9. Climate zones from the Köppen-Geiger climate classification in which true absences (soybean yield equals zero) were randomly selected. Locations of these climate zones can be seen in Figure S23.

Climate zone code	Description of the corresponding climate
BWh	arid, winter dry , hot arid
BWk	arid, winter dry , cold arid
Dfc	snow, fully humid, cool summer
Dfd	snow, fully humid, extremely continental
EF	polar frost
ET	polar tundra

Table S10 – Selected AgMIP data from GGCM Phase 1.

Crop model	Climate data	Irrigation	Management
EPIC-BOKU	GRASP	No	fullharm
EPIC-IIASA	GRASP	No	fullharm
GEPIC	GRASP	No	fullharm
LPJmL	GRASP	No	default
pAPSIM	GRASP	No	fullharm
pDSSAT	GRASP	No	fullharm
PEPIC	GRASP	No	fullharm
PEGASUS	AgMERRA	No	fullharm

- "default": the default setup of their model
- "fullharm": harmonized growing seasons (i.e. prescribed grid-cell- and crop-specific sowing and maturity dates) and fertilizer inputs
- "harmnon": for the same harmonized growing seasons but with unlimited nutrient supply (harmnon)

Table S11 – Selected data from the AgMIP fast-track dataset.

Climate model	RCP	CO2	Crop models				
			PEGASUS	pDSSAT	LPJmL	GEPIC	EPIC-BOKU
<i>GFDL-ESM2M</i>	<i>RCP4.5</i>	<i>yes</i>	X	X	X	X	
		<i>no</i>		X	X		
	<i>RCP8.5</i>	<i>yes</i>	X	X	X	X	X
		<i>no</i>	X	X	X		X
<i>HadGEM2-ES</i>	<i>RCP4.5</i>	<i>yes</i>	X	X	X	X	X
		<i>no</i>	X	X	X	X	X
	<i>RCP8.5</i>	<i>yes</i>	X	X	X	X	X
		<i>no</i>	X	X	X	X	X
<i>IPSL-CM5A-LR</i>	<i>RCP4.5</i>	<i>yes</i>	X	X	X	X	
		<i>no</i>		X	X		
	<i>RCP8.5</i>	<i>yes</i>	X	X	X	X	X
		<i>no</i>	X	X	X		X
<i>MIROC-ESM-CHEM</i>	<i>RCP4.5</i>	<i>yes</i>	X	X	X	X	
		<i>no</i>		X	X		
	<i>RCP8.5</i>	<i>yes</i>	X	X	X	X	X
		<i>no</i>	X	X	X		X
<i>NorESM1-M</i>	<i>RCP4.5</i>	<i>yes</i>	X	X	X	X	
		<i>no</i>		X	X		
	<i>RCP8.5</i>	<i>yes</i>	X	X	X	X	X
		<i>no</i>	X	X	X		X

Supplementary details and discussion about the Random Forest model

Analysis of residuals

As shown in Figure S2, the predictive ability of the Random Forest algorithm is good. An analysis of residuals (yield data - yield prediction) reveals a tendency of the model to overestimate low yields and underestimate high yield (Figure S26-A), highlighting a conservative behavior of the model. Nevertheless, histogram of residuals indicates that the distribution of residuals is symmetrical and centered, and that most residuals range between -1 to +1 t ha⁻¹ (Figure S26-B). Similar observations (good predictive ability and conservative behavior) about Random Forest when used to predict crop yields have been reported in the literature²⁶. Residuals show no association with latitude (Figure S27 A) and average in-season tmax (Figure S27 B), which indicates good performance of the model from low to high latitudes and from cold to hot years. However, residuals show an association with total in-season rainfall, with higher residuals at low rainfall values and lower residuals at high rainfall values (Figure S27 C-D). For example, when total in-season rainfall is lower than 600 mm (low rainfall) residuals ranged from -2.3 to 3.1 t ha⁻¹, and when total in-season rainfall is higher than 1500 mm (high rainfall) residuals ranged from -1.3 to 1.4 t ha⁻¹. A possible explanation for this is that soil properties are not included in model predictors. Indeed, at low rainfall values, the water-holding capacity of soils is important to buffer low and variable rainfall, while at high rainfall values the water-holding capacity of soils is not likely to have a significant effect on crop water availability. This would explain that RF predictive ability is lower at low rainfall value, although this requires further research. Interestingly, temperature does not seem to interact with the effect of total rainfall on model residuals as hot and cold years don't show any departure from the general residuals trend along the rainfall gradient (Figure S27 C-D).

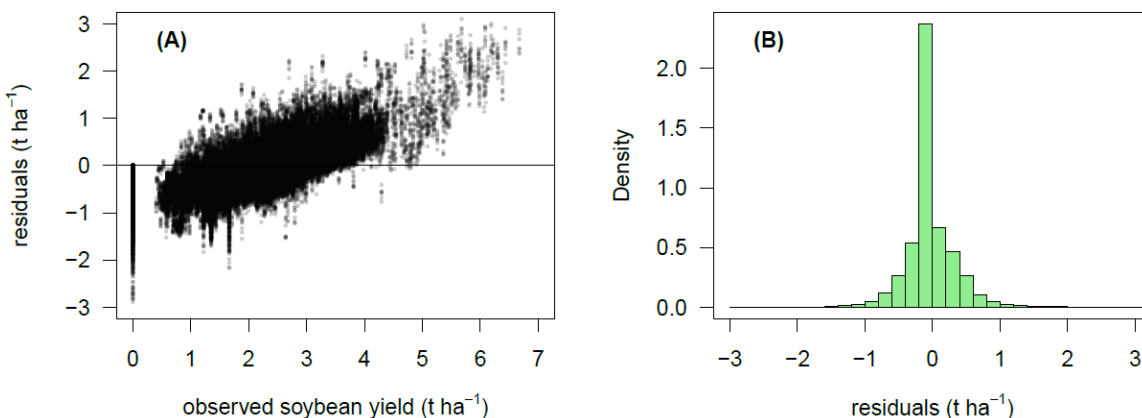


Figure S26. Analysis of the Random Forest residuals. (A) Residuals as a function of observed soybean yield. (B) Histogram of residuals. Residuals were calculated as data – prediction on 25 out-of-bag samples that were generated by bootstrap.

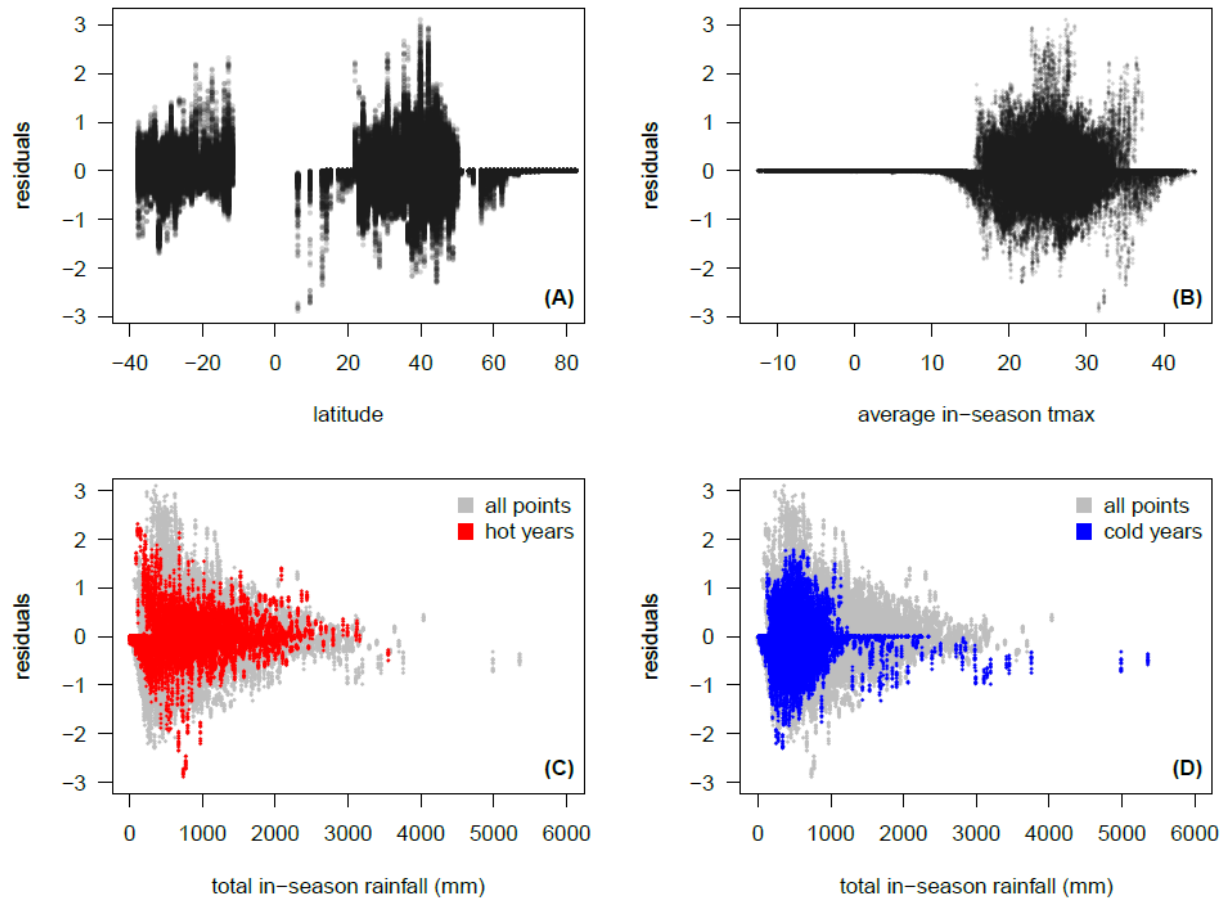


Figure S27. Random Forest residuals as a function of (A) latitude, (B) average in-season tmax, (C) total in-season rainfall (all points and hot years), (D) total in-season rainfall (all points and cold years). Residuals were calculated as data – prediction on 25 out-of-bag samples that were generated by bootstrap. Average in-season tmax was calculated as the average of monthly tmax over the 7 months of the growing season, and is expressed in degree Celsius. Total in-season rainfall was calculated as the sum of total monthly rainfall over the 7 months of the growing season. Hot years in panel (C) are defined as years for which average in-season tmax is higher than the 90th percentile of average in-season tmax (~30°C). Cold years in panel (D) are defined as years for which average in-season tmax is lower than the 40th percentile of average in-season tmax (~22°C).

Variable importance

Variable importance of the fitted Random Forest algorithm is presented in Figure S28. It highlights the importance of temperature in the model, which is consistent with the key role of temperature in soybean suitable area shifts due to climate change found with the LDA (Figure 3).

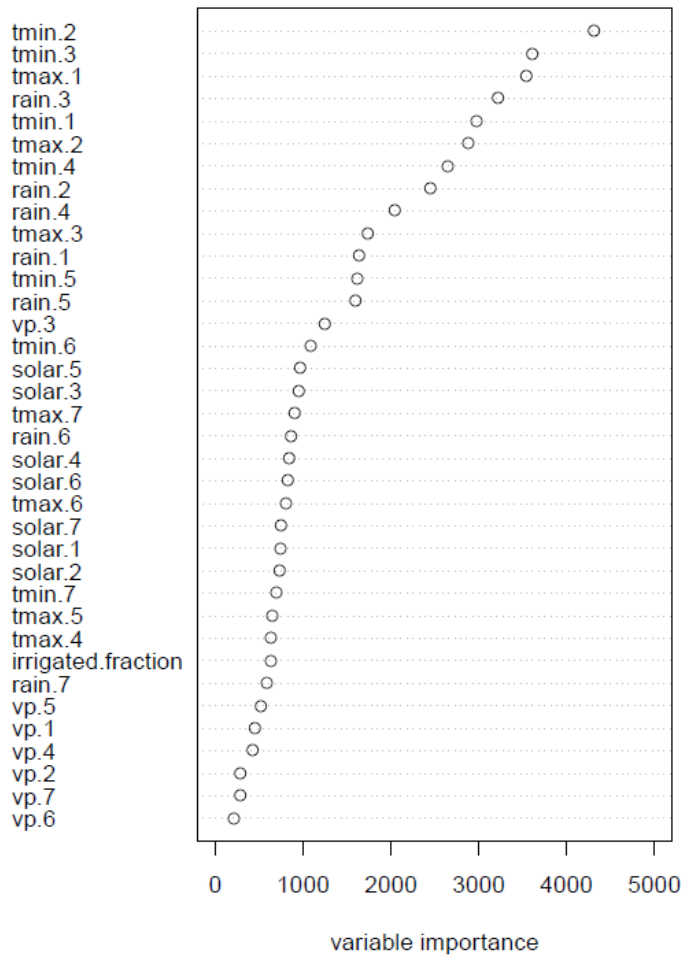


Figure S28. Measures of importance of the inputs used by Random Forest . The Random Forest algorithm (regression mode) was fitted with the function *ranger()* of the R package *Ranger* v0.10.1 with argument “importance” set to “impurity”. Variable importance is assessed by a measure of impurity which is the variance of the responses. Variables are sorted by decreasing order. *tmax*: monthly average daily maximum air temperature at 2m, *tmin*: monthly average daily minimum air temperature at 2m, *rain*: monthly total precipitation (mm month⁻¹), *solar*: monthly average daily solar radiation (MJ m⁻² day⁻¹), *vp*: monthly average daily vapor pressure (hPa). The numerical suffix refers to the month of the soybean growing season (from 1 to 7).

Additional references

1. Food and Agriculture Organization of the United Nations. FAOSTAT Statistics Database. (2019). Available at: <http://www.fao.org/faostat/en/#data>.
2. 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).
3. Iizumi, T. *et al.* Uncertainties of potentials and recent changes in global yields of major crops resulting from census- and satellite-based yield datasets at multiple resolutions. *PLoS One* **13**, e0203809 (2018).
4. Iizumi, T., Okada, M. & Yokozawa, M. A meteorological forcing data set for global crop modeling: Development, evaluation, and intercomparison. *J. Geophys. Res. Atmos. Res.* **119**, 363–384 (2014).
5. Taylor, K. e., Stouffer, R. J. & Meehl, G. A. An Overview of CMIP5 and experiment design. *Am. Meteorol. Soc.* **93**, 485–498 (2012).
6. Ramankutty, N., Evan, A. T., Monfreda, C. & Foley, J. A. Farming the planet : 1. Geographic distribution of global agricultural lands in the year 2000. *Global Biogeochem. Cycles* **22**, 1–19 (2008).
7. Commission, E. Sustainable land use (greening). Available at: https://ec.europa.eu/info/food-farming-fisheries/key-policies/common-agricultural-policy/income-support/greening_en. (Accessed: 27th February 2020)
8. Monfreda, C., Ramankutty, N. & Foley, J. A. Farming the planet : 2. Geographic distribution of crop areas , yields , physiological types , and net primary production in the year 2000. *Global Biogeochem. Cycles* **22**, 1–19 (2008).
9. Mueller, N. D. *et al.* Closing yield gaps through nutrient and water management. *Nature* **490**, 254–257 (2012).
10. Kühling, I., Hüsing, B., Bome, N. & Trautz, D. Soybeans in high latitudes : effects of Bradyrhizobium inoculation in Northwest Germany and southern West Siberia. *Org. Agric.* **8**, 159–171 (2018).
11. Kadžiulienė, Ž., Arlauskienė, A. & Šarūnaitė, L. Legumes for sustainability of agroecosystems. in *20th Baltic Agronomy Forum* (eds. Gaile, Z. & Dace, Š.) 56 (2016).
12. Zimmer, S. *et al.* Effects of soybean variety and Bradyrhizobium strains on yield, protein content and biological nitrogen fixation under cool growing conditions in Germany. *Eur. J. Agron.* **72**, 38–46 (2016).
13. Pannecoque, J. *et al.* Screening for soybean varieties suited to Belgian growing conditions based on maturity, yield components and resistance to Sclerotinia sclerotiorum and Rhizoctonia solani anastomosis group 2-2IIIB. *J. Agric. Sci.* 1–8 (2018). doi:10.1017/S0021859618000333
14. Lamichhane, J. R. *et al.* Analysis of soybean germination, emergence, and prediction of a possible northward establishment of the crop under climate change. *Eur. J. o* **113**, 125972 (2020).
15. Salem, M. A., Kakani, V. G., Koti, S. & Reddy, K. R. Pollen-Based Screening of Soybean Genotypes for High Temperatures. *Crop Sci.* **47**, 219–231 (2007).

16. Djanaguiraman, M., Schapaugh, W., Fritschi, F., Nguyen, H. & Prasad, P. V. Reproductive success of soybean (*Glycine max* L . Merrill) cultivars and exotic lines under high daytime temperature. *Plant Cell Environ.* **42**, 321–336 (2019).
17. Harley, P. C., Weber, J. a & Gates, D. M. Interactive effects of light, leaf temperature, CO₂ and O₂ on photosynthesis in soybean. *Planta* **165**, 249–263 (1985).
18. Setiyono, T. D. *et al.* Simulation of soybean growth and yield in near-optimal growth conditions. *F. Crop. Res.* **119**, 161–174 (2010).
19. Hesketh, J. D., Myhre, D. L. & Willey, C. R. Temperature Control of Time Intervals Between Vegetative and Reproductive Events in Soybeans 1 . *Crop Sci.* **13**, 250–254 (1973).
20. Egli, D. B. & Wardlaw, I. F. Temperature Response of Seed Growth Characteristics of Soybeans. *Agron. J.* **72**, 560–564 (1980).
21. Boote, K. J. *et al.* Elevated temperature and CO₂ impacts on pollination, reproductive growth, and yield of several globally important crops. *J. Agric. Meteorol.* **60**, 469–474 (2005).
22. Baker, J. T., Allen, L. H., Boote, K. J., Jones, P. & Jones, J. W. Response of Soybean to Air Temperature and Carbon Dioxide Concentration. *Crop Sci.* **29**, 98–105 (1989).
23. Brown, D. M. & Chapman, L. J. Soybean Ecology. II. Development-Temperature-Moisture Relationships from Field Studies. *Agron. J.* **52**, 496–499 (1960).
24. Kottek, M., Grieser, C., Beck, C., Rudolf, B. & Rubel, F. World Map of the Köppen-Geiger climate classification updated. *Meteorol. Zeitschrift* **15**, 259–263 (2006).
25. Rubel, F., Brugger, K., Haslinger, K. & Auer, I. The climate of the European Alps: Shift of very high resolution Köppen-Geiger climate zones 1800-2100. *Meteorol. Zeitschrift* **26**, 115–125 (2017).
26. Jeong, J. H. *et al.* Random Forests for Global and Regional Crop Yield Predictions. *PLoS One* **11**, e0156571 (2016).