

Seed yield and quality of soybeans across the US: an overview of management and
environmental characterization

by

Valentina Maria Pereyra Picabea

B.S., Universidad Nacional de Entre Rios, 2021

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Agronomy
College of Agriculture

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2024

Abstract

Soybean (*Glycine max* L.) is the main oilseed worldwide with the United States (US) producing 29 % of the global production (125 million Tn). The main targets of crop production in soybeans are yield (farmer profit), and seed quality in terms of protein (soybean meal for animal feed), and oil (for human consumption and biofuels). Seed yields are relatively stable due to the capacity of branches and nodes to adapt and maintain yield per area in diverse environmental conditions. However, soybean quality exhibits significant variation both within and between regions due to environmental, genetic, and management factors. Therefore, optimizing growth and understanding how management and environmental factors modulate seed yield and quality is critical for improving productivity and use of soybeans. This dissertation comprises six chapters (Chapter 1, Introduction, and Chapter 6, Conclusions). In Chapter 2, we explored metrics to characterize plant-to-plant spacing and its impact on yield across seeding rate and yield environments. We found that lack of plant-to-plant uniformity negatively impacted yield in both low- and medium-yield environments ($<3 \text{ Mg ha}^{-1}$) under low plant densities ($<20 \text{ plants m}^{-2}$), but with no effect at high yields (4.3 Mg ha^{-1}). In Chapter 3, we developed an assessment of yield and seed quality allocation to mainstems and branches of an old and modern soybean genotype at varying row spacings. The modern genotype presented 50% greater yields, with a three-fold greater contribution from branches compared to the old genotype at narrow rows ($<0.38 \text{ m}$). In addition, the modern genotype maintained protein levels despite high yields. In Chapter 4, we collected 235 on-farm seed and soil samples across 13 states of the United States (US). In addition, we collected management and weather data seeking to relate environment covariates to seed quality variation across regions. We predicted yield and seed oil concentration with better accuracy (R^2 0.56 and 0.39, respectively) than protein concentration

(R^2 0.09). Yield and oil in the south were mainly limited by delayed planting dates that exposed early stages to high temperatures and shortened the season length. In the north, limitations were related to low temperatures during seed formation due to long maturity groups or late planting dates. Finally, in Chapter 5, we built a robust model for predicting oil (R^2 0.54, MAE. 0.75%) and protein (R^2 0.45, MAE 1.13%) concentration in all soybean fields of the US. We gathered satellite data on weather and soil, calculated the green chlorophyll vegetation index (GCVI) and derived the growing season length at a 1 x 1 km resolution. The outcomes of this study will assist farmers to develop near-real time predictions of soybean quality changes not only for segregation at field-scale but providing essential information for soybean exports.

Seed yield and quality of soybeans across the US: an overview of management and
environmental characterization

by

Valentina Maria Pereyra Picabea

B.S., Universidad Nacional de Entre Rios, 2021

A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Agronomy
College of Agriculture

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2024

Approved by:

Major Professor
Dr. Ignacio A. Ciampitti

Copyright

© Valentina M Pereyra Picabea 2024.

Abstract

Soybean (*Glycine max* L.) is the main oilseed worldwide with the United States (US) producing 29 % of the global production (125 million Tn). The main targets of crop production in soybeans are yield (farmer profit), and seed quality in terms of protein (soybean meal for animal feed), and oil (for human consumption and biofuels). Seed yields are relatively stable due to the capacity of branches and nodes to adapt and maintain yield per area in diverse environmental conditions. However, soybean quality exhibits significant variation both within and between regions due to environmental, genetic, and management factors. Therefore, optimizing growth and understanding how management and environmental factors modulate seed yield and quality is critical for improving productivity and use of soybeans. This dissertation comprises six chapters (Chapter 1, Introduction, and Chapter 6, Conclusions). In Chapter 2, we explored metrics to characterize plant-to-plant spacing and its impact on yield across seeding rate and yield environments. We found that lack of plant-to-plant uniformity negatively impacted yield in both low- and medium-yield environments ($<3 \text{ Mg ha}^{-1}$) under low plant densities ($<20 \text{ plants m}^{-2}$), but with no effect at high yields (4.3 Mg ha^{-1}). In Chapter 3, we developed an assessment of yield and seed quality allocation to mainstems and branches of an old and modern soybean genotype at varying row spacings. The modern genotype presented 50% greater yields, with a three-fold greater contribution from branches compared to the old genotype at narrow rows ($<0.38 \text{ m}$). In addition, the modern genotype maintained protein levels despite high yields. In Chapter 4, we collected 235 on-farm seed and soil samples across 13 states of the United States (US). In addition, we collected management and weather data seeking to relate environment covariates to seed quality variation across regions. We predicted yield and seed oil concentration with better accuracy (R^2 0.56 and 0.39, respectively) than protein concentration

(R^2 0.09). Yield and oil in the south were mainly limited by delayed planting dates that exposed early stages to high temperatures and shortened the season length. In the north, limitations were related to low temperatures during seed formation due to long maturity groups or late planting dates. Finally, in Chapter 5, we built a robust model for predicting oil (R^2 0.54, MAE. 0.75%) and protein (R^2 0.45, MAE 1.13%) concentration in all soybean fields of the US. We gathered satellite data on weather and soil, calculated the green chlorophyll vegetation index (GCVI) and derived the growing season length at a 1 x 1 km resolution. The outcomes of this study will assist farmers to develop near-real time predictions of soybean quality changes not only for segregation at field-scale but providing essential information for soybean exports.

Table of Contents

List of Figures	xi
List of Tables	xvii
Acknowledgements	xix
Dedication	xxi
Chapter 1 - General Introduction	1
Chapter 2 - Early-season plant-to-plant spatial uniformity can affect soybean yields	6
Abstract	6
2.1 Introduction	7
2.2 Material and Methods	10
2.2.1 Sites description and field operations	10
2.2.2 Study design	11
2.2.3 Data collection and spacing uniformity variables	12
2.2.4 Statistical analysis and software	15
2.3 Results	17
2.3.1. Treatment effects on yield and spatial stand uniformity metrics	17
2.3.2 Spatial stand uniformity effects on soybean yield	19
2.4 Discussion	23
2.5 Conclusion	26
Chapter 3 - Soybean seed yield, protein, and oil concentration for a modern and old genotype under varying row spacings	27
Abstract	27
3.1 Introduction	28
3.2 Material and methods	30
3.2.1 Field conditions and management	30
3.2.2 Field determinations	31
3.2.3 Statistical analysis	31
3.3 Results	33
3.4 Discussion	39
3.5 Conclusion	41

Chapter 4 - Assessing on-farm drivers of soybean seed protein, and oil concentration	43
Abstract	43
4.1 Introduction	44
4.2 Material and Methods	46
4.2.1 <i>On-farm database on soybean yield, management practices, soil, and seed quality</i>	46
4.2.2 Weather and crop modelling	47
4.2.3 Data analysis	48
4.2.3.1 Variable importance assessment	48
4.2.3.2 Regional influence and relationships	50
4.3 Results	51
4.4 Discussion	59
4.5 Conclusion	62
Chapter 5 - Soybean seed protein and oil concentration prediction in the US during 2022 and 2023	63
Abstract	63
5.1 Introduction	64
5.2 Material and methods	66
5.2.1 On-farm database on soybean yield, management practices, soil, and seed quality	66
5.2.2 Satellite data	67
5.2.3 Model Evaluation and prediction	69
5.3 Results	70
5.4 Discussion	74
5.5 Conclusion	77
Chapter 6 - Conclusions	78
Limitations and Future Research	80
References	82
Appendix A - Chapter 2 Supplementary Material	95
Appendix B - Chapter 3 Supplementary Material	100
Appendix C - Chapter 4 Supplementary Material	105
Appendix D - Chapter 5 Supplementary Material	107

List of Figures

- Figure 2.1. Plant position classification based on the observed spacing to the previous plant within a row compared to the theoretical plant spacing (TPS). The TPS is the targeted (perfect) plant spacing of a plant population given a seeding rate and row width. The difference between TPS and observed spacing to the previous plant was rated as double ($< 17.5\%$ of TPS); misplaced by 66% (between $17.5-50.5\%$ of TPS), misplaced by 33% (between $50.5-83.5\%$ of TPS), perfectly spaced plants (between $83.5-116.5\%$ of TPS), short-skip plant (between $116.5-149.5\%$ of TPS), long-skip plants (between $149.5-182.5\%$ of TPS), double-skip plants (between $182.5-215.5\%$ of TPS), and greater than double-skip plants ($>215.5\%$ of TPS)..... 14
- Figure 2.2. Relationship between stand uniformity variables and soybean yield for US studies. ATEI mean and sd: achieved versus targeted evenness index mean and standard deviation, EI mean and sd: evenness index mean and standard deviation, Perfect: percentage of perfectly spaced plants, R^2 : coefficient of determination, RMSE: root mean square error. All stand uniformity variables presented a significant slope at $\alpha = 0.05$ 20
- Figure 2.3. Relationship of spacing standard deviation (Spacing sd, cm) and achieved versus targeted evenness index standard deviation (ATEI sd) to soybean yield. Different colors and line types denote different countries (Argentina, Arg - full line, red points; United States, US - dashed line, blue points). R^2 = coefficient of determination, RMSE= root mean square error..... 21
- Figure 2.4. Relationship between soybean yield (Mg ha^{-1}) and ATEI sd (achieved versus targeted evenness index standard deviation) in low- (b, mean of 2.7 Mg ha^{-1}), medium- (d, mean of 3.0 Mg ha^{-1}), and high- (f, mean of 4.3 Mg ha^{-1}) yield environments. The size of the bubbles indicates the relative density (plant density to seeding rate ratio), and colors indicate lower (red), middle (blue), and upper (green) thirds of achieved plant density distribution within each yield environment. Panels a, c, and e show the density distribution of the obtained plant density (plants ha^{-1}) by yield environment, full line and dashed lines denotes the median, and lower-upper terciles, respectively. Yield environments are based on ANOVA of site-year versus yield (Appendix Fig. A.1). Interaction between ATEI sd and yield

environment was significant ($\alpha = 0.05$). R^2 = coefficient of determination, RMSE = root mean square error.	23
Figure 3.1 Soybean seed yield (kg ha^{-1}) of old (1980) and modern (2013) soybean genotypes at row spacings of 0.19, 0.38, 0.76, and 1.52 m in 2020 (panel A) and 2021 (panel B) growing seasons. The error bars represent the 95% confidence intervals around the estimated expected value.	34
Figure 3.2 Soybean seed yield from branches (%; A and B), and from branches and main stems (kg ha^{-1} , C and D; dashed and full line, respectively) of old (1980) and modern (2013) genotypes at row spacings of 0.19, 0.38, 0.76, and 1.52 m, in 2020 (panels A and C) and 2021 (panels B and D) growing seasons. The error bars represent the 95% confidence intervals around the estimated expected value. Insets A1 and A2 present the median number of pods in branches across nodes of the main stem averaged over all tested row spacings in 2020 and 2021 growing seasons, respectively.	37
Figure 3.3 Soybean seed oil (panels A and B) and protein (panels C and D) concentration (%) in the whole plant of the old (1980) and modern (2013) genotype at row spacings of 0.19, 0.38, 0.76, and 1.52 m, in 2020 (panels A and C) and 2021 growing seasons (panels B and D). The error bars represents the 95% confidence intervals around the estimated expected value.	38
Figure 3.4 Soybean seed oil (panels A and B) and protein (panels C and D) concentrations (%) in branches (dashed lines) and main stems (full line) (C and D) of the old (1980) and modern (2013) genotype at row spacings of 0.19, 0.38, 0.76, and 1.52 m, in 2020 (panels A and C) and 2021 growing seasons (panels B and D). The error bars represents the 95% confidence intervals around the estimated expected value.	39
Figure 4.1. Location of the surveyed farmer fields (circles) in the US during the 2022 and 2023 growing seasons, showing the total per state (A). Histogram displaying yield (Mg ha^{-1}), seed oil and protein concentrations (%) in both 2022 and 2023 growing seasons (B).	53
Figure 4.2. Variable importance (feature gain) of XGBoost model to predict yield (Mg ha^{-1} , A), oil (%; B), and protein (%; C) concentration based on each of the thirteen-training sets (each point). Shaded areas indicate the .25, .5, .75, and .95 quantiles. Maximum and minimum temperature ($T_{\max}$ and $T_{\min}$, respectively), maturity group (MG), precipitation (Prec), vapor pressure deficit (VPD), number of days during	

seed fill with maximum temperatures $> 30^{\circ}\text{C}$ ($\text{sDaysT}_{\text{max}} > 30$), soil organic matter (OM). Weather was summarized for the vegetative (v), reproductive pre-seed filling (r), and reproductive seed filling (s) periods. Crop growth stages were simulated with the Decision the Support System for Agrotechnology Transfer (DSSAT) model.....	55
Figure 4.3 Location of the surveyed farmer fields (circles) colored according to the cluster classification, showing the total per cluster (A). Histogram displaying yield (Mg ha^{-1}), protein, and oil concentration (%) for each cluster (B).....	57
Figure 4.4. Soybean yield in relation to maximum temperature in VE-R1 (A), maturity group (B), and vapor pressure deficit (VPD, kPa) between R5-R7 (C). Additionally, soybean oil concentration in relation to minimum temperature between R5-R7 (D), number of days between R1-R5 (E), and minimum temperature between R1-R5 (F). Lastly, protein concentration in relation to precipitation (mm) between R1-R5 (G), VE-R1 (H), and R5-R7 (I). The colors represent three clustered regions: North (green), Center (Corn Belt, purple), and South (yellow). Growth stages were simulated with DSSAT. The relationships were fitted using generalized additive models. The coefficient of determination (R^2) and mean absolute error (MAE) are shown.....	58
Figure 5.1. Workflow overview. Green Chlorophyll Vegetation Index (GCVI), soil organic matter (OM), minimum and maximum temperature (T_{min} & T_{max}), beginning of pod (R3) and full seed (R6) phenological stages obtained from CROPGRO simulation model, coefficient of determination (R^2), concordance correlation coefficient (CCC), mean absolute error (MAE), crop data layer (CDL).	67
Figure 5.2. Location of the farmer fields (circles) in the US during the 2022 and 2023 growing seasons, showing the total per state (A). Histogram displaying seed oil and protein concentrations (%) in both 2022 and 2023 growing seasons (B).	72
Figure 5.3. Comparison of observed versus predicted oil (%), A) and protein concentration (%), B) of soybean on the testing dataset (30%) of the final model. The dots represent observed values from a farmer field, the dashed line shows the 1:1 relationship, and the solid line denotes the overall relationship between the observed and predicted data. Coefficient of determination (R^2), root mean square error (RMSE), concordance correlation coefficient (CCC), mean absolute error (MAE), percentage lack of accuracy (PLA), and percentage	

lack of precision (PLP) are shown. Variable importance (feature gain) of XGBoost model for oil (C) and protein (D) concentration for each of the 10 training datasets from the cross-validation with the best combination of hyperparameters. Shaded areas indicate .25, .50, .75, and .95 quantiles.	73
Figure 5.4. Example map showing the prediction of seed oil and protein in soybean fields near Manhattan, KS during 2023.	74
Appendix Figure A.1. Effect of site-year on yield, plant density, and all community-scale metrics of spatial uniformity (ATEI mean, ATEI sd, EI mean, EI sd, Doubles, Mis 33, Mis 66, Perfect, Short-skip, Long-skip, Double-skip and >Double-skip). Different colors denote site-year used as training (red) and testing (blue) data sets. Fisher's LSD pair-wise comparisons was performed on significative variables. Means followed by the same letter are not significantly different at $P = 0.05$. References: ATEI sd = achieved versus targeted evenness index standard deviation, EI sd = evenness index standard deviation, Double skip = percentage of double skips plants, Mis 66 = percentage of plants misplaced by 66%, U.S.: United States, Arg: Argentina, Dry: dryland, Irr: Irrigated, 19 and 20: 2019 and 2020 calendar years.....	98
Appendix Figure B.1. Density plot of obtained plant density (plants ha ⁻¹) for the evaluated seeding rates (148 and 445 thousand seeds ha ⁻¹) in 2020 and 2021. White lines denote the median of plant density.	100
Appendix Figure B.2. Canopy coverage (%) of soybean between 50 and 100 days after sowing (R2 and R6 phenological stages, respectively) of the old (1980) and modern (2013) genotype at row spacings of 0.19, 0.38, 0.76, and 1.52 m, in 2020 and 2021. Dashed lines represent 95% of light interception.	101
Appendix Figure B.3. Density distribution of the difference between the modern and old genotypes for yield (kg ha ⁻¹) in 2020 and 2021. Lines denote the 95 % quantiles, and the mean.	101
Appendix Figure B.4. Density distribution of the difference between the modern and old genotype for yield in branches (% , top panel) and main stems (kg ha ⁻¹ , bottom panel) in 2020 and 2021. Lines denote the 95 % quantiles, and the mean.	102

Appendix Figure B.5. Density distribution of the difference between the modern and old genotype for the seed protein (top panel) and oil (bottom panel) concentration (%) in 2020 and 2021. Lines denote the 95 % quantile, and the mean.	103
Appendix Figure B.6. Density distribution of the difference between the modern and old genotype for the seed protein (top panel) and oil (bottom panel) concentration (%) in branches and main stems in 2020 and 2021. Lines denote the 95 % quantile, and the mean.	104
Appendix Figure C.1. Correlation matrix of environmental descriptors. Pearson's correlation is displayed with the pie charts, indicating the direction and magnitude of each correlation. Only variables significantly correlated are displayed ($p < 0.05$). Seed protein and oil (%), yield (kg ha^{-1}), row spacing (RowSp, m), seeding rate (SeedingR, seeds m^{-2}), maturity group (MG), Season length from planting (SeasonL), Latitude (Lat), Longitude (Lon), radiation (Radn, MJ m^{-2}), maximum and minimum temperature (T_{max} and T_{min} , respectively, $^{\circ}\text{C}$), vapor pressure deficit (VPD, kPa), precipitation (Prec, mm), beginning of flowering (R1), beginning of seed (R5), physiological maturity (R7), sowing date (Sowing), number of days during seed fill with maximum temperature $> 30^{\circ}\text{C}$ ($\text{DaysT}_{\text{max}} > 30$), number of days during seed fill with minimum temperature $< 15^{\circ}\text{C}$ ($\text{DaysT}_{\text{min}} < 15$), clay (%), silt (%), sand (%), soil organic matter (OM, %), pH, soil No^{-3} (ppm), soil P (ppm), and cation exchange capacity (CEC, $\text{meq } 100 \text{ g}^{-1} \text{ soil}$).	105
Appendix Figure C.2. Comparison of observed versus predicted yield (Mg ha^{-1} , A), oil concentration (% , B), and protein concentration (% , C) of soybean. The observed dots represent the median value for protein and oil, and the reported yield from a farmer field of the U.S. that was not used during model training, the dashed line shows the 1:1 relationship, and the solid line denotes the overall relationship between the observed and predicted data. Coefficient of determination (R^2), concordance correlation coefficient (CCC), mean absolute error (MAE), percentage lack of accuracy (PLA), and percentage lack of precision (PLP) are shown.	106
Appendix Figure D.1. Example harmonic fits between day of the year and the Green Chlorophyll Vegetation Index (GCVI). Dots represent cloud-free observations from Sentinel-2.	107

Appendix Figure D.2. Agreement between the start (A), peak (B), and end (C) of the season obtained with GCVI versus the emergence, beginning of seed formation (R5), and harvest maturity (R8), respectively, obtained from DSSAT. The coefficient of determination (R^2) and mean absolute error (MAE) are shown.	108
Appendix Figure D.3. Correlation matrix of environmental descriptors. Pearson's correlation is displayed with the pie charts, indicating the direction and magnitude of each correlation. Only variables significantly correlated are displayed ($p < 0.05$). Soybean seed protein and oil (%), Latitude (Lat), Longitude (Lon), precipitation (Prec, mm), radiation (Rad, MJ m^{-2}), maximum and minimum temperature (T_{max} and T_{min} , respectively, $^{\circ}\text{C}$), green chlorophyll vegetation index (GCVI), clay (%), soil organic matter (OM, %), pH.....	109
Appendix Figure D.4. Comparison of observed versus predicted oil (% , A) and protein concentration (% , B) of soybean on all held-out datasets from the cross validation with best combination of hyperparameters. The dots represent observed values from a farmer field, the dashed line shows the 1:1 relationship, and the solid line denotes the overall relationship between the observed and predicted data. Coefficient of determination (R^2), root mean square error (RMSE), concordance correlation coefficient (CCC), mean absolute error (MAE), percentage lack of accuracy (PLA), and percentage lack of precision (PLP) are shown.	110

List of Tables

Table 2.1. Effect of planter type, seeding rate, and their interaction on variables from plant position classification for all US studies. References: percentage of perfectly spaced plants (Perfect), percentage of plants misplaced by 66% (Mis 66), percentage of plants misplaced by 33% (Mis 33), percentage of double plants (Double), percentage of short skips plants (Short-skip), percentage of long skip plants (Long-skip), percentage of double skips plants (Double-skip), and percentage of greater than double skip plants (> Double-skip).	18
Table 2.2. Effect of planter type, seeding rate, and their interaction on yield and stand uniformity variables for all site-years. References: Spacing between plants standard deviation (Spacing sd), achieved versus targeted evenness index mean and standard deviation (ATEI and ATEI sd, respectively), and evenness index mean and standard deviation (EI and EI sd, respectively).	19
Table 3.1 Soybean biomass (kg ha ⁻¹) in R2 (full flowering) and R5 (beginning seed filling) phenological stages of old (1980) and modern (2013) genotypes at row spacings of 0.19, 0.37, 0.76, and 1.52 m, in 2020 and 2021. The numbers between parenthesis indicate the 95% confidence interval around the estimated expected value.	34
Table 4.1 Summary of season characteristics, soil, and management variables for each state showing the median and range of observations for each variable.	52
Table 4.2 Summary of season characteristics, weather, and soil variables for each cluster showing the median and the range (between brackets) of observations for each variable. ..	56
Table 5.1. Summary of season characteristics, soil, and management variables for each state showing the median and range (between brackets) of observations for each variable.	71
Appendix Table A.1. Effect of planter type, seeding rate, and their interaction for each of the four US site years. References: Spacing between plants standard deviation (Spacing sd), percentage of perfectly spaced plants (Perfect), percentage of plants misplaced by 66% (Mis 66), percentage of plants misplaced by 33% (Mis 33), percentage of double plants (Double), percentage of short skips plants (Short-skip), percentage of long skip plants (Long-skip), percentage of double skips plants (Double-skip), percentage of greater than double skip	

plants (> Double-skip), achieved versus targeted evenness index mean and standard deviation (ATEI and ATEI sd, respectively), and evenness index standard deviation (EI sd).

..... 95

Appendix Table A.2. ANOVA table of ATEI sd (achieved versus targeted evenness index standard deviation) and seeding rate (SR) level effects on soybean yield (Mg ha^{-1}). Low, medium, and high SR levels based on SR lower than 200 seeds m^{-2} , between 200 and 300 seeds m^{-2} , and greater than 300 seeds m^{-2} , respectively..... 97

Acknowledgements

I would like to thank the Facultad de Ciencias Agropecuarias from Universidad Nacional de Entre Rios and all professors that teach and inspired me during my undergraduate degree in Agriculture Engineer. Thanks to INTA Paraná, especially to Ricardo Melchiori and Nicolas Maltese, for giving me the space to conduct my undergraduate thesis research and pave the path that led me to K-State. Thanks for mentoring me and introducing me to science. Thanks to all my friends from Oro Verde, Crespo, and beyond. Thanks for my family for always looking up for me and protecting me.

Special gratitude to my mentor and advisor, Dr. Ignacio Ciampitti, for extending and expanding his expertise continuously with so many, for sharing your passion and inspiring me with it. Thanks for your mentorship that goes well beyond the science, for encouraging me to think different, to question perspectives and boundaries, and rethink what's possible. Thank you for helping me to align my career goals and connecting me with the right people. Also, special mention to all members of Ciampitti Lab throughout the years, thanks for your work and help on collecting data. I saw many of you sacrificing a lot to come here pursuing professional and personal growth and facing numerous challenges along the way, you also inspired me.

I thank Dr. Vara Prasad for involving and entrusting me in international projects and giving me opportunities to expand my career goals, thanks for supporting me. I also thank the rest of my committee member, Dr. Trevor Hefley and Dr. Ajay Sharda, for your guidance and support throughout my Ph.D. I am also grateful to the Food Security Scholars from the Staley School of Leadership for the discussions around global and local food security, you gave me a space to express myself and my aspirations. Jenna Wallace, Adelaide Easter, Natalie Nusz, and Mary Kay Siefers, thank you for your work on the Student Advisory Leadership Team, thank

you for caring so much about your impact, for showing passion, conviction, and for seeking change. I also want to thank the Officers of the Graduate Student Council of K-State for all your work and shared experiences throughout the years we worked together, thanks to Dean Claudia Petrescu and Dr. Megan Miller for your leadership.

Gratitude for John Deere, the United Soybean Board, and the North Central Soybean Research Program for the financial support and allowing me to conduct the research during my Ph.D. program.

Dedication

To my family and to my friends, thanks for your support and love.

To the little me, my inner child, we've come a long long way.

To all Ciampitti Lab, you inspired me forever.

This was possible because of you.

Chapter 1 - General Introduction

The global production of oilseeds is estimated at 690 million tons, with soybeans representing 62% of this total. Therefore, soybeans are the leading source of protein meal globally (71%, 271 million tons) and the second leading source of vegetable oil (15%, 65 million tons) after palm oil (USDA-FAS, 2024). The United States is the second-largest producer of soybeans (125 million tons, 29%), as well as soybean meal (52 million tons, 19%) and soybean oil (13 million tons, 20%) (USDA-FAS, 2024). A substantial proportion of U.S. soybean seeds (40%) are exported and subsequently processed in the importing countries. While soybean meal is primarily utilized for animal feed, soybean oil is predominantly employed in human food products (55%) and as a biodiesel feedstock (18%) (OECD, FAO-UN, 2024). The future of soybean demand is influenced by several factors, including the use of soybeans as animal feed, regulations pertaining to deforestation-free and GMO-free products, and biofuel policies that promote the use of renewable diesel (Egli, 2024; OECD, FAO-UN, 2024). However, from an agronomic perspective, the primary objective remains to be the improvement of soybean seed yield and quality, particularly in terms of protein and oil, in order to meet market demands at the farm level.

Soybean development is characterized by two overlapping phases: a vegetative and a reproductive growth phase. The vegetative phase begins at emergence (VE) (Fehr and Caviness, 1977) and continues until the onset of flowering (R1), with vegetative growth occurring until late reproductive stages. Vegetative growth is partitioned into main stems and branches and controlled by resource availability (i.e., radiation, water, temperature, and soil nutrients) and plant density (plants ha⁻¹), with lower plant densities favoring branching (Carpenter and Board, 1997a). The reproductive phase can be further divided into the seed set period (R1-R5) where

pod and seed number are determined, and the seed filling period (R5-R7) where seed weight and quality are determined (Vogel et al., 2021). The overlap of vegetative and reproductive phases, combined with the ability to adjust the number of branches provide soybeans with great phenotypic plasticity in growth and final seed yield.

The key management practices that influence soybean phenotypic plasticity include planting date, maturity group, seeding rate, and row spacing. For a given location, the planting date, maturity group, and environmental conditions collectively inform the selection of an appropriate seeding rate. The agronomic optimum seeding rate or plant density (AOSD and AOPD, respectively) is defined as the minimum necessary to maximize yield. High-yield environments can achieve higher yields per plant and adapt to lower AOSR than low-yield environments (Carciochi et al., 2019), supported mainly by branch growth (Carpenter and Board, 1997a; Frederick et al., 2001a). However, a non-uniform plant spacing can affect yield per plant and penalize yield at the community-scale (Masino et al., 2018a). Therefore, in Chapter 2 we aimed to assess the effect of plant-to-plant-spatial uniformity on soybean yield at different seeding rates (100 to 550 seeds m⁻²) and yield environments.

The impact of row spacing on soybean phenotypic plasticity is also significant, as it affects final yield due to a strong correlation with the time of canopy closure required to achieve 95% light interception. Yield improvements with narrow rows are more likely to occur with late planting dates, short maturity groups, or high temperatures, which shortens the crop season length.

Modern varieties, however, show higher yield per plant supported by greater yield on main stem but mainly on branches (Suhre et al., 2014a). Hence, modern varieties may have higher yield stability across different row spacings. Nevertheless, the allocation of yield to the main stem and branches at varying row spacings remains less understood when comparing a modern versus an

older soybean genotype. Therefore, in Chapter 3, we explored the yield and seed quality in the main stem and branches at different row spacings for an old and a modern genotype.

Soybean seed quality, mainly protein and oil concentration, is the most important aspect for the industry and upstream actors of the value chain that gear and shape market requirements and demand. Regional and farm-level variations in seed quality are modulated by factors related to the genotype, environment, and management (GxExM) (Rotundo et al., 2016; Rotundo and Westgate, 2009a). Thus, complex GxExM interactions result in substantial seed quality variations within and across regions (Naeve and Huerd, 2008; Rotundo et al., 2016), complicating efforts to define patterns and cause-effects. Assessing changes in seed quality at regional scales is a challenging and ongoing area of research. Integrating weather, soil, and crop management data to further understand and more accurately predict the composition of soybean seeds at a regional level is crucial for decision-making and fair pricing. In Chapter 4, we explored the integration of seed quality (using seed samples derived from farmer fields), together with yield, crop management, weather, and soil data to explain variations in seed quality across the main producing regions of the US.

Regional predictions of soybean seed protein and oil concentration can facilitate the segregation of production, improve the efficiency of marketing and trade, and provide insights into the potential for product differentiation. Research efforts have successfully explored regional variations in seed quality (Assefa et al., 2019; Cerrudo et al., 2023; Rotundo et al., 2016).

However, variation at lower than regional scales is high and inference on management and weather remains limited. The use of satellite data can prove successful for inference if the data set captures the ecophysiological generating process of the variation in question. Satellite data offers high spatial resolution on weather, soil, and applications related to the leaf area index

(LAI) (Gasó et al., 2023; Hosseini et al., 2015), crop phenology (Gao and Zhang, 2021; Lou et al., 2023), and plant biomass (Joshi et al., 2023). For instance, the green chlorophyll vegetation index (GCVI) has been demonstrated to predict protein concentration around a week after the peak, which is closely associated with maximum LAI values around the critical period for yield determination (R3-R5) (Hernandez et al., 2023). Therefore, satellite data presents a promising opportunity to capture within region variability and infer correlation and causation. In chapter 4, we used satellite data on soil, weather, and calculated GCVI to predict soybean protein and oil concentration. We utilized GCVI to calculate the length of the growing season and summarized weather and GCVI during vegetative, early, and late reproductive stages. Lastly, we used a robust model to create prediction maps of quality across all soybean fields in the US.

Chapter 2 - Early-season plant-to-plant spatial uniformity can affect soybean yields

**Published in Scientific Reports.*

Valentina M. Pereyra, Leonardo M. Bastos, André Froes de Borja Reis, Ricardo J. M. Melchiori, Nicolas E. Maltese, Stefania C. Appelhans, P. V. Vara Prasad, Yancy Wright, Edwin Brokesh, Ajay Sharda, Ignacio A. Ciampitti. Early-season plant-to-plant spatial uniformity can affect soybean yields. *Scientific Reports*. 2022, 12:17128. <https://doi.org/10.1038/s41598-022-21385-z>

Abstract

Increased soybean (*Glycine max* L. Merrill) seed costs have motivated interest in reduced seeding rates to improve profitability while maintaining or increasing yield. However, little is known about the effect of early-season plant-to-plant spatial uniformity on the yield of modern soybean varieties planted at reduced seeding rates. The objectives of this study were to (i) investigate traditional and devise new metrics for characterizing early-season plant-to-plant spatial uniformity, (ii) identify the best metrics correlating plant-to-plant spatial uniformity and soybean yield, and (iii) evaluate those metrics at different seeding rate (and achieved plant density) levels and yield environments. Soybean trials planted in 2019 and 2020 compared seeding rates of 160, 215, 270, and 321 thousand seeds ha⁻¹ planted with two different planters, Max Emerge and Exact Emerge, in rainfed and irrigated conditions in the United States (US). In addition, trials comparing seeding rates of 100, 230, 360, and 550 thousand seeds ha⁻¹ were conducted in Argentina (Arg) in 2019 and 2020. Achieved plant density, grain yield, and early-season plant-to-plant spacing (and calculated metrics) were measured in all trials. All site-years were separated into low- (2.7 Mg ha⁻¹), medium- (3 Mg ha⁻¹), and high- (4.3 Mg ha⁻¹) yielding environments, and the tested seeding rates were separated into low (< 200 seeds m⁻²),

medium (200–300 seeds m⁻²), and high (> 300 seeds m⁻²) levels. Out of the 13 metrics of spatial uniformity, standard deviation (sd) of spacing and of achieved versus targeted evenness index (herein termed as ATEI, observed to theoretical ratio of plant spacing) showed the greatest correlation with soybean yield in US trials ($R^2 = 0.26$ and 0.32 , respectively). However, only the ATEI sd, with increases denoting less uniform spacing, exhibited a consistent relationship with yield in both US and Arg trials. The effect of spatial uniformity (ATEI sd) on soybean yield differed by yield environment. Increases in ATEI sd (values > 1) negatively impacted soybean yields in both low- and medium-yield environments, and in achieved plant densities below 200 thousand plants ha⁻¹. High-yielding environments were unaffected by variations in spatial uniformity and plant density levels. Our study provides new insights into the effect of early-season plant-to-plant spatial uniformity on soybean yields, as influenced by yield environments and reduced plant densities.

2.1 Introduction

Soybean (*Glycine max* L. Merrill) seed costs have doubled since 1997 (USDA - ERS, 2021a) mainly due to increased genetic yield potential (Rincker et al., 2014a), the introduction of herbicide traits (Fernandez-Cornejo et al., 2014; Shi and Chavas, 2009), and more recently the increase in use of industrial seed treatments (Gaspar et al., 2017; USDA - ERS, 2021b). Today, farmers have adopted reduced seeding rates (David A. Hennessy, Yuyuan Che and Hongli Feng, 2019) supported by the improved planting equipment (Mark Jeschke and Nanticha Lutt, 2016), enhanced seed quality (related to seed vigor) and seed treatments (Cox and Cherney, 2011; Gaspar et al., 2017, 2015), and increased yield contribution from branches in modern soybean varieties (Suhre et al., 2014b). Maintaining yield while reducing seeding rates is largely

associated with the growth plasticity of soybean plants (Carpenter and Board, 1997b) (i.e., the ability of individual plants to produce compensatory yield through formation of branches). However, a plant density threshold exists, below which soybean plants cannot produce compensatory growth, and thus yield is penalized. This point is defined as the agronomical optimum seeding rate and plant density (AOSR, and AOPD, respectively) (Carciochi et al., 2019). Although AOSR and AOPD are commonly assessed at the community-scale, a non-uniform plant-to-plant spacing may penalize yield at the community-scale. Therefore, early-season plant-to-plant spatial uniformity may be crucial to modern soybean production, characterized by high yield potential and reduced seeding rates (Carciochi et al., 2019).

Plant-to-plant spatial uniformity seems to affect yield in soybeans mainly when the genotype by environment (GxE) interaction hinders the ability of the plants to produce compensatory yield due to limitations of resources (i.e., drought/heat stress, nutrient deficiencies, reduced growth cycle)(Masino et al., 2018b; Moore, 1991; Stivers and Swearingin, 1980). For instance, soybean yield in rainfed conditions has been negatively impacted (by 6-9%) when planted in non-equidistant patterns (Masino et al., 2018b; Moore, 1991; Stivers and Swearingin, 1980; Xu et al., 2021); however, longer maturity groups (MG) showed capacity to overcome more effectively lack of spatial plant uniformity and yield more relative to shorter MG (Masino et al., 2018b; Stivers and Swearingin, 1980). In the other hand, yield of irrigated soybeans may be maintained in both equidistant and non-equidistant plant spacings, even though plant size variability may increase (Andrade and Abbate, 2005). These results suggest that the ability of soybean plants to produce compensatory yield is impaired by availability of resources, reducing yield in a larger proportion under non-uniform relative to uniform fields. Research into the association between spatial uniformity and soybean yield remains scarce. Improving our

understanding of this relationship would substantiate the potential of precision planting techniques via their benefits on seed singulation and seeding rate recommendations, enhancing the efficiency of soybean production (Lee et al., 2008).

The variable commonly used to assess spatial uniformity on a plant-to-plant basis is spacing between plants (average distance to neighboring plants) and is commonly summarized at the community-scale by calculating the standard deviation (sd) (Andrade and Abbate, 2005) or coefficient of variation (cv) (Masino et al., 2018b) of distance between plants. The use of these methods to characterize plant spatial uniformity can be problematic for three reasons: i) the spacing between plants does not account for the different distances (or uneven spacing) to each neighboring plant, or the closeness to the theoretical plant spacing (of a given seeding rate and row width); ii) opposing plant distance outcomes (i.e., skipped and double plants) can generate similar statistical values while their effect on yield is contrasting; and iii) standard deviation (sd)(Andrade and Abbate, 2005) and coefficient of variation (cv)(Masino et al., 2018b) of distance between plants are strongly influenced by a few large spacings and by seeding rate (i.e. lower seeding rates result in greater spacing sd) (S. D. Kachman and J. A. Smith, 1995). Thus, there is a need for community-scale variables that i) are more sensitive at detecting different spacing outcomes; ii) provide insight into both spacing and overall distribution of plants, and iii) are based on data collected from farmer fields that are more representative of realistic production conditions.

Following this rationale, our working hypothesis is that early-season plant-to-plant spatial uniformity is more relevant for soybean yield under limited resources (low-yielding environments) and at low seeding rates (low levels of achieved plant density). Therefore, the objectives of this study were to i) explore both traditionally used and newly devised metrics to

characterize early-season plant-to-plant spatial uniformity, ii) identify the best metrics correlating the plant-to-plant spatial uniformity with soybean yield, and iii) evaluate those metrics at different seeding rate (and achieved plant density) levels and yield environments.

2.2 Material and Methods

2.2.1 Sites description and field operations

A total of six field studies were conducted in two different regions over two seasons. Four studies (two dryland and two irrigated) were in Kansas, United States (dryland: 39° 4' 30'' N, -96° 44' 43'' W, irrigated: 39° 4' 25'' N, -96° 43' 12'' W) during the 2019 and 2020 growing seasons (hereafter referred to as USDry19, USIrr19, USDry20, and USIrr20 studies). The remaining two studies (dryland) were in Entre Rios, Argentina (31° 50' 49'' S; 60° 32' 16'' W) during the 2018/2019 and 2019/2020 growing seasons (hereafter referred to as Arg19 and Arg20 studies). The soils were Fluventic Hapludolls [silt loam, 40% sand, 13% clay, 47% silt, organic matter (OM) 1.7%, 7.7 pH, 31.1 ppm P(Bray⁻¹)] at the US dryland studies, and Pachic Argiudolls [silty clay loam, 10.1% sand, 30.6% clay and 59.3% silt, OM 3.2%, 6.8 pH, 34.7 ppm P (Bray⁻¹)] at the US irrigated studies. At the Argentinian studies soil was a Vertic Argiudoll in 2019 [silty clay loam to clay loam, 3.9% sand, 27.6% clay, 67.9% silt, OM 2.65%, 7.2 pH, 12.5 ppm P (Bray⁻¹)] and an Acuic Argiudoll in 2020 [silt loam to silty-clay-loam, 5.6% sand, 28.6% clay, 65.8% silt, OM 3.33%].

The US dryland and irrigated studies were sown on June 4, 2019, and May 20, 2020. In 2019, the dryland study was replanted on June 29 due to poor emergence after the first sowing. The studies in Argentina were sown on December 5 in 2018 and November 20 in 2019. At all six studies, plots were kept free of weeds, pests, and diseases through recommended chemical control.

The genotypes used in the US were P40A47X (MG 4.0) and P39A58X (MG 3.9) (Corteva Agriscience, Johnston, IA, USA) in 2019 and 2020, respectively. Both varieties are tolerant to glyphosate and dicamba herbicides (RR2X) and have low lodging probability. For the northeast region of Kansas, recommended sowing dates range from May 15 to June 15 along with MG 4 (Ignacio Ciampitti et al., 2022). In addition, recommended seeding rates are between 270 and 355 thousand seeds ha⁻¹ for low-yielding environments and 190 to 285 thousand seeds ha⁻¹ for medium- and high-yielding environments (Carciochi et al., 2019). In Argentina, the genotype AW5815IPRO (MG 5.8, Bayer, Leverkusen, Germany) was used both in 2020 and 2021, it is tolerant to glyphosate and sulfonylureas and has low lodging probability. Recommended sowing dates for Entre Rios considering soybeans as a single crop range from October 20 to December 10, and MG usually range from 4 to 6; lastly, seeding rate recommendations are between 200 and 250 thousand seeds ha⁻¹ in the region (Di Mauro et al., 2018).

2.2.2 Study design

The studies carried out in the US were arranged as a split plot design with three replicates in both 2019 and 2020. In 2019, the main plot treatment factor was planter type with two levels [John Deere (Moline, Illinois, US) Max Emerge planter (ME, 12 rows), and John Deere Exact Emerge Planter (EE, 16 rows)], and the split-plot treatment factor was seeding rate with two levels (160 and 321 thousand seeds ha⁻¹). In 2020 the main plot treatment factor was also planter type with two levels (ME and EE), and the split-plot treatment factor was seeding rate with four levels (160, 215, 270 and 321 thousand seeds ha⁻¹). Planting speed was 7 km h⁻¹ in both studies and years, plots were 24 and 32 rows wide when planted with ME and EE, respectively, with 0.76 m row spacing. Plot length was 80 m in the dryland studies and 160 m in the irrigated

studies. The studies in Argentina were arranged as a single replicate of each seeding rate (100, 230, 360 and 550 thousand seeds ha⁻¹) in both years. Planting speed was 5.5 km h⁻¹ in both years, and plots were 10 rows wide with 0.52 m row spacing and 350 m in length.

All treatment factors in US studies were evaluated with the overall goal of producing substantial variation in the variable of interest, plant-to-plant spatial uniformity, rather than to make an inference of their effect on yield. The Argentinian studies were only used for selection of stand uniformity variables due to the single replicate. Plant spatial uniformity variables were first fitted using the data from US studies (details below), and then the best explanatory metrics were selected to re-fit the relationships combining both data sets from US and Argentina. Finally, sowing dates, maturity groups, and seeding rates evaluated in this study at both locations (Arg and US) were aligned with those recommended for each region.

2.2.3 Data collection and spacing uniformity variables

All treatment factors in US studies were evaluated with the overall goal of producing substantial variation in the variable of interest, plant-to-plant spatial uniformity, rather than to make an inference of their effect on yield. The Argentinian studies were only used for selection of stand uniformity variables due to the single replicate. Plant spatial uniformity variables were first fitted using the data from US studies (details below), and then the best explanatory metrics were selected to re-fit the relationships combining both data sets from US and Argentina. Finally, sowing dates, maturity groups, and seeding rates evaluated in this study at both locations (Arg and US) were aligned with those recommended for each region.

$$ATEI = \frac{Spacing(cm)}{TPS(cm)} \quad \text{Equation 2.1}$$

The ATEI index was designed to account for the proximity of the observed plant spacing to the TPS. Values closer to 1 indicate that the plant spacing is close to the TPS and values that

are below or above 1 indicate that the plant spacing is lower or higher than the TPS, respectively; thereby departing from an ideal plant spacing. Hence, ATEI values greater than 1 depict both i) non-uniform plant-to-plant spacing distribution and ii) plant densities below the target (seeding rate). To further understand the meaning of ATEI, the relative density (rd) was calculated as the ratio between plant density (based on the number of plants in the 2 m segment) and seeding rate.

To account for the unevenness of distance from a plant to both neighboring plants within the row, we used the Evenness Index (EI, dimensionless), calculated as the ratio between the distance to the nearest neighbor (cm) and the plant spacing (cm) of a given plant (Eq. 2). The Evenness Index values range from 0 to 1, a value closer to 1 indicates that a plant is equidistantly spaced to both of its neighboring plants within the row, if zero then those plants are occupying the same position (as doubles). It is important to note that EI does not provide information on the spacing (in distance, cm) or how close the spacing is compared to the TPS, but only describes the unevenness distance of a plant to its neighboring plants within a row.

$$EvennessIndex(EI) = \frac{nearestneighbor(cm)}{Spacing(cm)} \quad \text{Equation 2}$$

In addition, the distance from a plant to its preceding neighboring plant, and the TPS were used to classify the position of each plant into one of eight classes (Fig. 1). Plants were classified in classes ranging from “double” (preceding plant distance < 17.5% of TPS), to “perfect” (preceding plant distance between 83.5 and 116.5% of TPS), and “greater than double skip” (preceding plant distance greater than 215.5% of TPS).

Plant position classification		Range of Theoretical Plant Spacing
Double		< 17%
Plant misplaced by 66%		17 – 50%
Plant misplaced by 33%		50 – 83%
Perfect		83 – 116%
Short skip		116 – 149%
Long skip		149 – 182%
Double skip		182 – 215%
Greater than double skip		> 215%

Figure 2.1. Plant position classification based on the observed spacing to the previous plant within a row compared to the theoretical plant spacing (TPS). The TPS is the targeted (perfect) plant spacing of a plant population given a seeding rate and row width. The difference between TPS and observed spacing to the previous plant was rated as double (< 17.5% of TPS); misplaced by 66% (between 17.5-50.5% of TPS), misplaced by 33% (between 50.5-83.5 % of TPS), perfectly spaced plants (between 83.5-116.5% of TPS), short-skip plant (between 116.5-149.5% of TPS), long-skip plants (between 149.5-182.5% of TPS), double-skip plants (between 182.5-215.5% of TPS), and greater than double-skip plants (>215.5% of TPS).

Plant spatial uniformity metrics (Eq. 1 and 2, Fig. 1) were calculated on a plant-to-plant basis and then summarized at the segment level (hereafter called community-scale). Plant spacing (cm), ATEI, and EI were summarized by calculating their mean (except for plant spacing) and sd. Plant spacing sd and cv (directly proportional to sd) have been used in previous studies (Andrade and Abbate, 2005; Masino et al., 2018b) to estimate spatial variability. Hence, plant spacing sd (cm) was included as the base line variable to compare the performance of the new metrics proposed in this study. Lastly, plant position classes (Fig. 1) were summarized by

calculating the mean percentage of plants in a segment classified within each of the eight categories.

At physiological maturity, 10 (US studies) or 20 (Arg studies) consecutive plants within each segment were hand cut, individually identified, bagged, and oven-dried at 65°C until constant weight was achieved. Grain weight per plant was recorded and is reported on a 13 g kg⁻¹ water content. Yield (Mg ha⁻¹) was calculated at the segment level by summing the grain weight (g) of all harvested plants and correcting it to the segment area.

2.2.4 Statistical analysis and software

Community-scale data from the four US studies (site-years) were combined to model yield and the 13 stand uniformity metrics (ATEI mean, ATEI sd, EI mean, EI sd, Spacing sd (cm), Double, Mis 33, Mis 66, Perfect, Short-skip, Long-skip, Double-skip, >Double-skip) as a function of seeding rate, planter type and their interaction (fixed effects), and block nested in site-year (random effect) (Table 1 and 2). Independent models for each of the 4 US studies were built assessing the effects of planter type, seeding rate, and their interaction (fixed effects), and seeding rate nested in planter type, and in block (random effects) on the same variables previously mentioned (Appendix Table A.1). The models were run using the ‘lmer’ function from *lme4* package in R (R Core Team, 2021). In addition, the US and Arg studies were combined to evaluate the effect of site-year on yield, plant density, and all stand uniformity variables (Appendix Fig. A.1) using the ‘lm’ function from package *stats*. Means separation were performed using Fisher’s LSD (Least Significance Difference) test (alpha= 0.05) with ‘emmeans’ function from package *emmeans*.

Community-scale data from the four US studies were combined and fitted to bivariate linear regression models with yield as the response variable and each of the stand spatial

uniformity variables as the explanatory variable. Significant models ($\alpha=0.05$) were further evaluated by calculating the coefficient of determination (R^2) and root mean squared error (RMSE) (Fig. 2). Models with the lower RMSE and higher R^2 were selected as those that best captured the effect of non-uniform stands on soybean yield. After variables were selected, both US and Arg data sets were combined and the linear regressions between the selected variables and yield were re-fitted to assess the consistency of the relationships when an independent data set was included. Community-scale yield from US and Arg studies was modelled as a function of the selected stand uniformity variable, country (US and Arg), and their interaction (fixed effects) (Fig. 3). The spatial uniformity metric showing the most consistent relationship for both US and Arg studies (i.e., non-significant interaction between stand uniformity metric and country), was selected to continue the analysis. The bivariate linear regression models were run with function *lm*.

Different environmental conditions and seeding rate levels may modify the effect of plant spatial uniformity on yield. To explore this, each of the studies from Arg and US were separated into low- (USDry19 and ArgDry20, mean of 2.7 Mg ha⁻¹), medium- (USIrr19, USDry20 and ArgDry19, mean of 3.0 Mg ha⁻¹), and high- (USIrr20, mean of 4.3 Mg ha⁻¹) yield environments based on the effect of site-year on yield (Appendix Figure A.1). Additionally, the tested seeding rates were separated in low (<200 thousand seeds ha⁻¹), medium (between 200 and 300 thousand seeds ha⁻¹) and high (>300 thousand seeds ha⁻¹) levels based on the current optimal seeding rate for medium yielding environments (235 thousand seeds ha⁻¹, 4 Mg ha⁻¹) (Carciochi et al., 2019) and the extreme values proposed by Suhre et al. (2014) (148 and 445 thousand seeds ha⁻¹). This classification was used to model yield as a function of i) the selected stand uniformity metric, yield environment, and their interaction, and ii) the selected stand uniformity metric, seeding rate

levels, and their interaction. These models were tested to obtain a robust conclusion on the overall effect of yield environment and seeding rate levels, and their interactions (all treated as fixed effects) with plant-to-plant spatial uniformity relative to the response variable, soybean yield. The Akaike information criteria (AIC) was used to compare the full (with interactions) relative to the reduced models (single effects).

2.3 Results

2.3.1. Treatment effects on yield and spatial stand uniformity metrics

Mean cumulative seasonal precipitation was 560 and 630 mm for the USDry19 and USDry20, respectively; and 770 and 450 mm for the Arg19 and Arg20, respectively. Total water availability was 712 mm for USIrr19 (152 mm from irrigation) and 757 mm for USIrr20 (127 mm from irrigation). The effect of the treatments analyzed in this study (planter type and seeding rate) were significant on yield and in 8 out of 13 stand uniformity metrics ($\alpha = 0.05$, Tables 1 and 2). From the four seeding rates evaluated, attention will be given only to the lowest (161 thousand seeds ha^{-1}) and highest (321 thousand seeds ha^{-1}) seeding rates as only those levels were present at all four US studies. Yield was not affected by planter type, but a significant effect of seeding rate was observed. Yield was greater at the low seeding rate (3.3 Mg ha^{-1}) compared to the high seeding rate (3.0 Mg ha^{-1}). In terms of the spatial uniformity metrics, the EE planter generally improved the uniformity between plants compared to ME planter. The EE planter increased the number of perfectly spaced plants (24%), decreased the number of double plants (1.3%), and plants misplaced by 66% (4.3%) compared to the ME planter (12, 7.8, and 13% respectively). In addition, EE achieved a higher mean evenness index (EI) and lower evenness index sd (0.7 and 0.2, respectively) compared to ME (0.5 and 0.3, respectively). In terms of the variables affected by seeding rate, the low seeding rate was associated with a greater

proportion of perfectly spaced plants (23%), fewer double plants (4%), smaller ATEI sd and evenness index sd (0.6 and 0.2, respectively), and an increased mean evenness index (0.6) compared to that associated with the high seeding rate (14 and 7.0% of perfectly spaced and double plants, respectively; 1.1, and 0.3 for ATEI sd and evenness index sd, respectively; and 0.5 for mean evenness index). Thus, the lower seeding rate resulted in more uniform spacing between plants.

Average yield was greatest in the USIrr20 (4.3 Mg ha⁻¹, high yield environment), medium in Arg19, USIrr19 and USDry20 (2.9 to 3.2 Mg ha⁻¹, medium yield environment), and lowest in the Arg20 and USDry19 (2.8 and 2.6 Mg ha⁻¹, respectively, low yield environment). More information on individual studies and treatment effects on yields can be consulted in the supplementary material (Appendix Table A.1, Appendix Fig. A.1).

Table 2.1. Effect of planter type, seeding rate, and their interaction on variables from plant position classification for all US studies. References: percentage of perfectly spaced plants (Perfect), percentage of plants misplaced by 66% (Mis 66), percentage of plants misplaced by 33% (Mis 33), percentage of double plants (Double), percentage of short skips plants (Short-skip), percentage of long skip plants (Long-skip), percentage of double skips plants (Double-skip), and percentage of greater than double skip plants (> Double-skip).

		Perfect	Mis 66	Mis 33	Double	Short	Long	Double	>Double
		(%)	(%)	(%)	(%)	skip (%)	skip (%)	skip (%)	skip (%)
Planter¹	EE	23.8 a ²	4.3 b	15.5	1.3 b	17.0	11.4	8.1	18.9
	ME	11.8 b	13.1 a	12.2	7.8 a	14.1	9.1	8.5	23.7
	<i>p-value</i>	<i><0.001</i>	<i><0.001</i>	<i>0.070</i>	<i><0.001</i>	<i>0.126</i>	<i>0.165</i>	<i>0.812</i>	<i>0.057</i>
Seeding rate	161	22.7 a	6.4	12.7	3.9 ab	15.7	9.0 ab	6.8	22.3
(Thousand seeds ha⁻¹)	215	21.1 a	9.6	10.6	3.9 b	17.0	13.1 a	7.9	17.7
	269	13.2 b	10.0	16.5	3.1 b	13.7	13.1 a	10.0	21.5
	321	14.1 b	8.9	15.6	7.0 a	16.0	5.8 b	8.5	23.8
Planter x Seeding rate	<i>p-value</i>	<i><0.001</i>	<i>0.168</i>	<i>0.108</i>	<i>0.023</i>	<i>0.766</i>	<i>0.003</i>	<i>0.445</i>	<i>0.493</i>
		<i>0.011</i>	<i>0.085</i>	<i>0.414</i>	<i>0.249</i>	<i>0.032</i>	<i>0.122</i>	<i>0.787</i>	<i>0.885</i>

¹Planter type: Max Emerge (ME) and Exact Emerge (EE)

²Letter and ANOVA (analyses of variance) threshold. Different letters mean significant differences among treatments.

Table 2.2. Effect of planter type, seeding rate, and their interaction on yield and stand uniformity variables for all site-years. References: Spacing between plants standard deviation (Spacing sd), achieved versus targeted evenness index mean and standard deviation (ATEI and ATEI sd, respectively), and evenness index mean and standard deviation (EI and EI sd, respectively).

		Yield (Mg ha ⁻¹)	Spacing sd (cm)	ATEI	ATEI sd	EI	EI sd
Planter	EE	3.19	4.0	1.6	0.7	0.7 a	0.2 b
	ME	3.2	4.5	1.5	0.8	0.5 b	0.3 a
<i>p-value</i>		0.948	0.371	0.489	0.383	<0.001	<0.001
Seeding rate (thousand seeds ha⁻¹)	161	3.34 a	5.2	1.5	0.6 b	0.6 a	0.2 b
	215	3.36 ab	4.1	1.5	0.7 b	0.6 a	0.2 b
	269	3.06 ab	3.3	1.5	0.7 b	0.6 ab	0.2 b
	321	3.02 b	4.4	1.7	1.1 a	0.5b	0.3 a
Planter x Seeding rate		<i>p-value</i>	0.04	0.058	0.051	<0.001	<0.001
			0.831	0.978	0.760	0.977	0.229
			0.020				

¹Planter type: Max Emerge (ME) and Exact Emerge (EE)

²Letter and ANOVA threshold. Different letters mean significant differences among treatments.

2.3.2 Spatial stand uniformity effects on soybean yield

Soybean yield from the four US studies (n = 190) was significantly correlated with 7 out of the 13 plant spatial uniformity metrics (alpha = 0.05, Fig. 2). Yield was negatively correlated with ATEI sd ($R^2 = 0.32$), Spacing sd (cm, $R^2 = 0.27$), ATEI mean ($R^2 = 0.16$), > Double-skip plants (% , $R^2 = 0.04$), and Evenness Index (EI, $R^2 = 0.04$) sd. While soybean yield was positively correlated with EI mean ($R^2 = 0.03$) and perfectly spaced plants ($R^2 = 0.02$). Spacing sd and ATEI sd were selected to continue with the analysis due to their greater R^2 and lower RMSE.

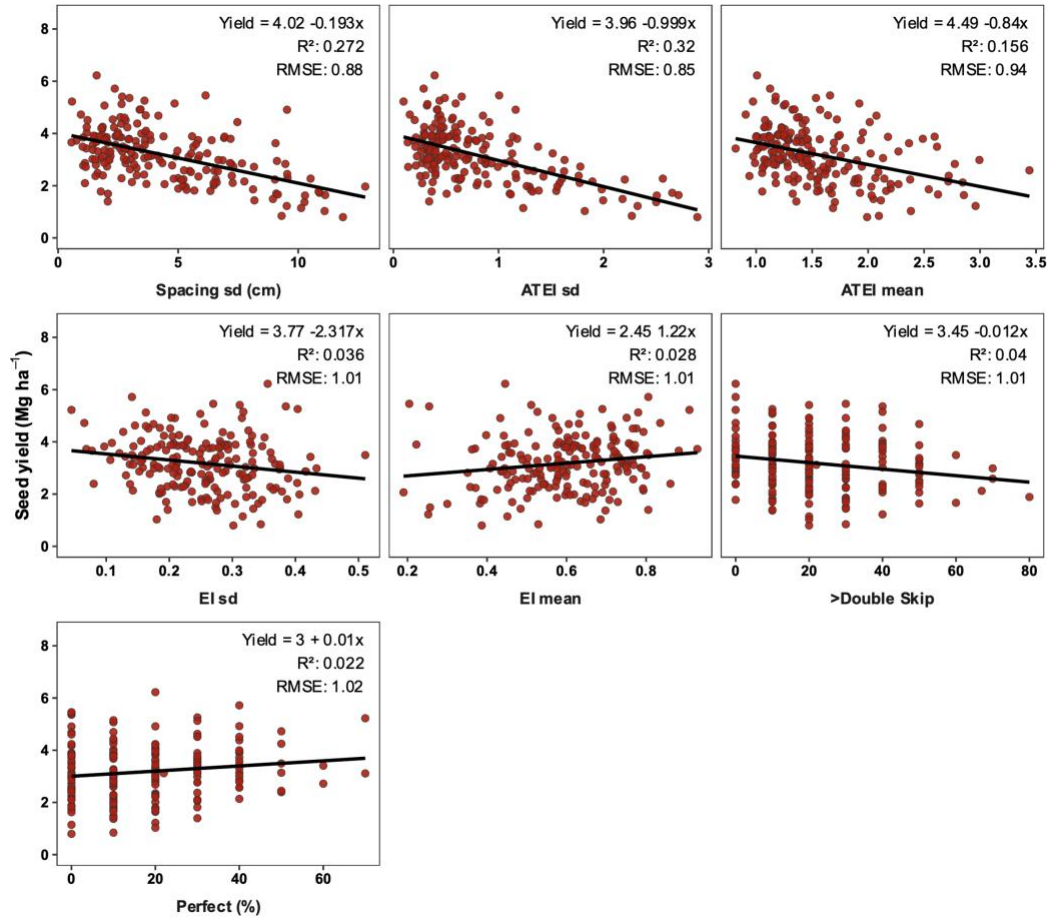


Figure 2.2. Relationship between stand uniformity variables and soybean yield for US studies. ATEI mean and sd: achieved versus targeted evenness index mean and standard deviation, EI mean and sd: evenness index mean and standard deviation, Perfect: percentage of perfectly spaced plants, R²: coefficient of determination, RMSE: root mean square error. All stand uniformity variables presented a significant slope at alpha = 0.05.

Spacing sd and ATEI sd, the selected metrics based on R² and RMSE (Fig. 2), were correlated with yield from all six studies (US and Arg, n = 262). Yield from Arg was significantly correlated with both spacing sd and ATEI sd. However, only ATEI sd presented a stable relationship across datasets of both countries (non-significant interaction). In this sense, for each unit of increase in spacing sd, US showed a greater reduction in yield (110 kg ha⁻¹) compared to Arg (30 kg ha⁻¹), even though the observed range of spacing sd was smaller in US (0.5 to 12.8 cm) than Arg (1.2 to 21.7 cm). This indicates greater variability in the spacing

between plants in Arg did not translate to a greater yield reduction relative to the US dataset. Whereas ATEI sd and yield relationship had a different intercept (due to yield levels US 3.8, and Arg 3.7 Mg ha⁻¹), the slope of yield decreased per unit increase in ATEI sd did not statistically differ between datasets (960 kg ha⁻¹, Fig. 3). Moreover, ATEI sd ranges were similar between the US (0.1 to 2.9) and Arg (0.3 to 2.4). Thus, ATEI sd was selected as the variable that best explained plant spatial uniformity effects on soybean yield.

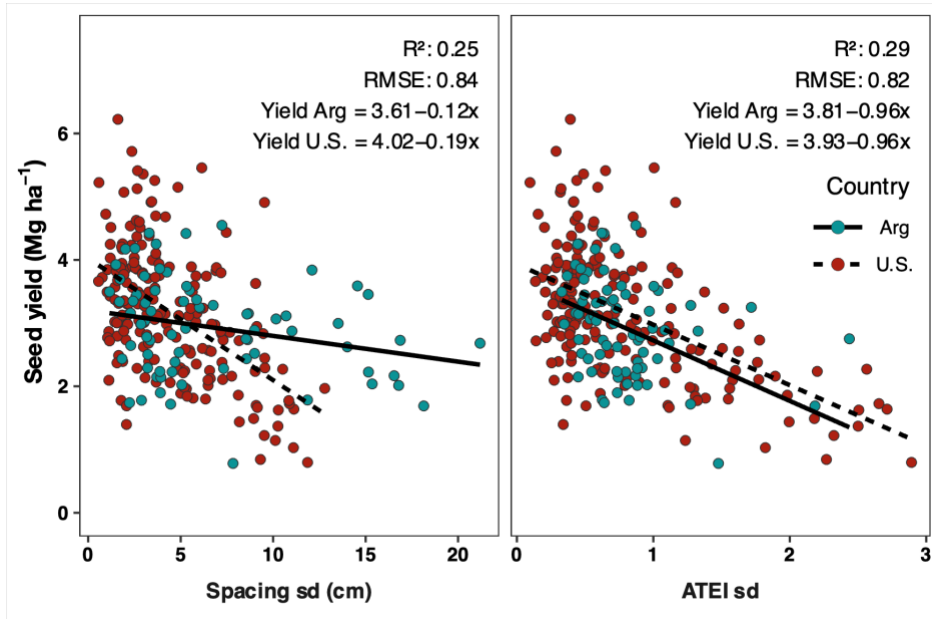


Figure 2.3. Relationship of spacing standard deviation (Spacing sd, cm) and achieved versus targeted evenness index standard deviation (ATEI sd) to soybean yield. Different colors and line types denote different countries (Argentina, Arg - full line, red points; United States, US - dashed line, blue points). R²= coefficient of determination, RMSE= root mean square error.

To assess the relationship between yield and plant spatial uniformity (ATEI sd) at different yield environments, site years were divided into Low (USDry19 and Arg20, 2.7 Mg ha⁻¹), Medium (USIrr19, Arg19, and USDry20, 3.0 Mg ha⁻¹), and High (USIrr20, 4.3 Mg ha⁻¹) yield environments (Appendix Fig. A.1). Increases in ATEI sd negatively impacted yield in both low-

and medium- yield environments (0.75 and 0.83 Mg ha⁻¹ per unit increase in ATEI sd, respectively), while no effect was observed in high-yield environments ($p>0.05$, Fig. 4). Moreover, yield from low- and medium- yield environments showed a strong correlation with ATEI sd (R^2 of 0.32 and 0.26, respectively). Although the relationship between yield and plant spatial uniformity (ATEI sd) did not change with seeding rate levels (Appendix Table A.2), the achieved plant density changed across studies. The achieved plant densities (panels a, c, and e in Fig. 4) were similar across yield environments with a median of 148 thousand plants ha⁻¹. Yield reductions in low- and medium-yield environments were associated with values of ATEI sd >1 , plant densities <200 thousand plants ha⁻¹, and with achieved relative to target densities (bubble sizes, panels b, d, and f in Fig. 4) around 50%. The model performance (measured by AIC) improved by including the effect of yield environment relative to the simpler model (ATEI sd only).

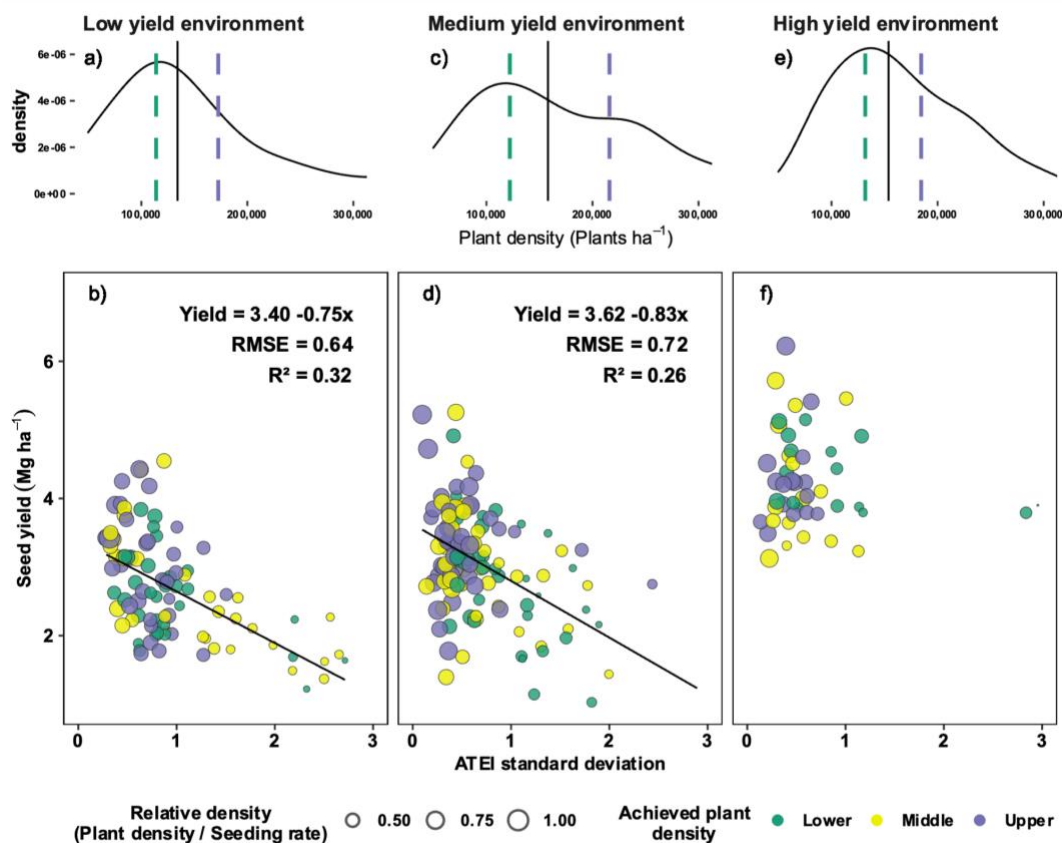


Figure 2.4. Relationship between soybean yield (Mg ha⁻¹) and ATEI sd (achieved versus targeted evenness index standard deviation) in low- (b, mean of 2.7 Mg ha⁻¹), medium- (d, mean of 3.0 Mg ha⁻¹), and high- (f, mean of 4.3 Mg ha⁻¹) yield environments. The size of the bubbles indicates the relative density (plant density to seeding rate ratio), and colors indicate lower (red), middle (blue), and upper (green) thirds of achieved plant density distribution within each yield environment. Panels a, c, and e show the density distribution of the obtained plant density (plants ha⁻¹) by yield environment, full line and dashed lines denotes the median, and lower-upper terciles, respectively. Yield environments are based on ANOVA of site-year versus yield (Appendix Fig. A.1). Interaction between ATEI sd and yield environment was significant ($\alpha = 0.05$). R² = coefficient of determination, RMSE = root mean square error.

2.4 Discussion

This study provides new insights into the relevance of soybean plant-to-plant spacing uniformity, mainly under low- and medium-yield environments and reduced plant densities. Soybean breeding efforts have altered the overall contribution of branches to yield, reducing the dependency on high seeding rates over time (Carpenter and Board, 1997b; Suhre et al., 2014b).

Still, low yielding environments limit the ability of the plants to produce compensatory yield from branches and thus require increased seeding rates (Carciochi et al., 2019; Gaspar et al., 2020). Our results showed that in both high- and low- yield environments, soybean yield was not increased by seeding rates greater than 121 thousand seeds ha⁻¹. Instead, yield variability from low- and medium- yield environments were explained by plant spatial uniformity (ATEI sd).

Plant-to-plant spatial uniformity effects on yield can be understood from the onset of competition at the plant level, which determines yield at the community-scale. Soybean seedlings initiate shade avoidance responses when they sense reductions in the red:far red ratio of the light spectrum (i.e., shading due to increased plant-to-plant competition), causing reductions in branch development, overall growth, and yield per plant (Casal et al., 1990; Green-Tracewicz et al., 2011; Yang et al., 2014). Similarly, a reduction of early-season plant-to-plant spatial uniformity could result in a continuum of shade avoidance responses with differing competitive abilities and yields at the plant level. At a community-level, the ability of soybean plants to produce compensatory yields via branches (Board and Settini, 1986a; Frederick et al., 2001b) is impaired by restrictive environments such as late planting dates (Settini and Board, 1988a) or drought stress (Frederick et al., 2001b). In summary, early-season plant-to-plant uniformity can determine the onset of a differential competitive ability among individual plants (Maddonni and Otegui, 2004) mainly under limited resources and reduced plant densities penalizing yield at a community scale.

The most promising metric of plant spatial uniformity, ATEI sd, explained the highest variability of soybean yield (higher R²) and reduced the error of the estimation (smaller RMSE) across studies from the US and Arg (Figure 3). Furthermore, the indices of stand uniformity presented (Fig. 1, Eq. 1, Eq. 2) were able to detect differences in spatial distribution between

planter types and among seeding rates, whereas the most common uniformity index used in previous literature, spacing sd, did not (Table 1 and 2). A previous study has shown that 77% of spacing sd variability is explained by clusters (3 or more plants together), double plants, short-skips, and long-skips, with the latter making the greatest contribution to its variability (Liu et al., 2004). High spatial stand uniformity of seedlings is dependent not only on a highly uniform sowing pattern, but also on high emergence ratios (Lu et al., 2020). In this sense, ATEI sd reflects non-uniform spacing as well as poor emergence and plant mortality. In summary, ATEI sd is a better predictor of soybean yield and is a useful metric for measuring within-row stand uniformity in future research.

This study is a first attempt at describing the effects of spatial uniformity on soybean yield in environments producing phenotypes that do not express the plant's growth potential, penalizing yield at the community-scale. In agreement, a review from Lu et al. (2020) concluded that plant spatial uniformity resulting from precision sowing is an effective way to increase soybean yields at all seeding rates. Moreover, lowering seeding rates along with improvements in planting quality delivering seed at a uniform spacing and depth can become critical components of modern soybean systems. Farmers could adopt such practices as an insurance against yield losses in poor-yielding conditions. Future studies exploring AOSR in low- and medium- yielding environments should consider plant spatial uniformity and relative density (plant density to seeding rate ratio) (Carciochi et al., 2019; Gaspar et al., 2020) as factors that may confound the yield response to changes in seeding rates. Improved foundational knowledge of key soil and environmental factors affecting early- and within-season plant-to-plant spatial uniformity is still a research gap that should be explored in future research studies. Finally, the study of the differential plant-to-plant ability to produce compensatory yield (Maddonni and

Otegui, 2004) could clarify the influence of canopy characteristics determined early in the season on yield formation under realistic field conditions.

2.5 Conclusion

Our working hypothesis “early-season spatial plant-to-plant uniformity is more relevant for soybean yield under limited resources (low yielding environments) and under low seeding rates (low levels of achieved plant density)” is partially accepted. On one side, early-season plant-to-plant spatial uniformity impacted soybean yields mainly under both low- and medium-yield environments. However, this effect is not only relevant for low seeding rates, but mainly for plant densities below 200 thousand plants ha^{-1} . Plant spatial uniformity under field conditions results from uniformity in the sowing pattern, emergence, and early-establishment of plants. To capture these effects, relating the observed spacing with the TPS (obtaining ATEI) removed the confounding effects of spacing sd. The negative impact on soybean yields due to reduced spatial uniformity and plant establishment suggests an impairment of plants to produce compensatory yields under resource-limited environments. Hence, yield optima in low- and medium-yield environments resulted from improved plant spatial uniformity (ATEI sd <1) and plant densities above 200 thousand plants ha^{-1} .

Chapter 3 - Soybean seed yield, protein, and oil concentration for a modern and old genotype under varying row spacings

**Published in Heliyon.*

Valentina M. Pereyra, Trevor Hefley, P. V. Vara Prasad, Ignacio A. Ciampitti. Soybean seed yield, protein, and oil concentration for a modern and old genotype under varying row spacings. Scientific Reports. 2022, 12:17128. <https://doi.org/10.1016/j.heliyon.2024.e35054>.

Abstract

Progress in soybean (*Glycine max* L.) breeding has led to a reduction in optimal seeding rates due to enhanced branching capacity over time. However, less is known about the changes in canopy architecture between old and modern soybean genotypes at varying row spacing and their impact on yield and seed quality through the main stem and branches. Therefore, this study aimed to i) evaluate yield and seed quality responses of an old and modern soybean genotype at different row spacings and ii) examine the yield and seed quality of branches and the main stem. Trials were conducted in Kansas (United States, US) during 2020 and 2021, comparing two genotypes (old, released in 1980, and modern, released in 2013) at four row spacings (0.19, 0.38, 0.76, and 1.52 m) under rainfed conditions. Seed yield and quality (protein and oil concentrations, %) were assessed at the end of each growing season. In 2021, both genotypes had low and similar yields at all row spacings (averaging 2481 kg ha⁻¹) with 2.5% less protein on branches compared to the main stem. However, 2022 resulted in a high-yielding environment, with the modern yielding 50% more (3584 kg ha⁻¹) than the old (2315 kg ha⁻¹) genotype in narrow row spacings (<0.38 m). Additionally, the modern genotype showed a three-fold greater contribution to yield from branches (1113 kg ha⁻¹) relative to the old genotype (379 kg ha⁻¹).

Despite the high yields observed in narrow rows, the modern genotype maintained protein levels. These results highlight the importance of row spacing as a key management practice for improving yield while maintaining protein levels in high yield conditions.

3.1 Introduction

Soybean accounts for 60% (398 million Mg in 2023) of the global oilseed production (USDA-FASS, 2022), primarily intended for animal feed in the form of soybean meal (76%) and for human consumption primarily as oil (20%) (Fraanje et al., 2020). The overall increase in global soybean production is driven by greater yields per unit area, with no signs of yield stagnation in 76% of the harvested area (Foley et al., 2011; Ray et al., 2012). The United States (US) accounts for 29% of the global production, with farmer yields steadily increasing at 36 kg ha⁻¹ year⁻¹ during the last four decades (USDA-NASS, 2024a). Hence, future progress in genetic gain is promising not only for sustaining yield but also for improving seed composition (Borja Reis et al., 2022a).

On-farm yield improvements over time can be attributed to the continued adoption of modern genotypes and agronomic technologies (Specht et al., 2015). The main agronomic technologies associated with these yield gains include earlier planting dates (approximately 12 days earlier to late April/early May) (Sacks and Kucharik, 2011), narrower row spacings (reduced from 0.76 to 0.38 m), minimum or no-till system supported by herbicide-tolerant genotypes, and new planters that ensure uniform seed placement depth and spacing (Karayel, 2009; Zhai et al., 2020). From a plant architecture standpoint, modern genotypes can set more seeds on the main stem and branches, enhancing the ability to adjust yield at varying plant densities (Suhre et al., 2014a). Nonetheless, little is known about how row spacings affect

the yield allocation to branches and main stem when comparing modern relative to older soybean genotypes

Over the past decade, US growers in the north-central region have increasingly adopted narrow (below 0.76 m) over wide (>0.76 m) row spacing (Specht et al., 2015), with a current adoption rate of 66% of producers using narrow and 34% using wide spacing (USDA-NASS, 2024b). Although yield benefits associated with narrow rows are expected (M. Singh et al., 2023), studies have found inconsistent results (Andrade et al., 2019; De Bruin and Pedersen, 2008; Pedersen and Lauer, 2003; Thompson et al., 2015). Narrow rows can provide yield benefits when the crop growth cycle is shortened, such as with late sowing dates, short maturity groups (MG), or under high temperatures (Rotundo and Westgate, 2009a). In optimum irrigated conditions with fast canopy closure, in the other hand, yield improvements might not be observed (B. Singh et al., 2023). Different row spacings can also alter the contribution of yield from branches and main stems (Norsworthy and Shipe, 2005). For instance, narrow rows show a potential to increase yield through branch growth (Board et al., 1990). However, further research is needed on the implications of varying row spacings for modern soybeans, particularly on the contribution of branches and main stem to yield formation and seed quality or composition, including protein and oil concentrations.

For both old and modern genotypes, seed quality on the main stem consistently shows greater protein and less oil concentrations on the upper relative to the bottom section (Bellaloui & Gillen, 2010; Moro Rosso et al., 2021). Conversely, branches can have similar protein but less oil concentrations compared to the main stem (Moro Rosso et al., 2021). Evaluating these dynamics is particularly timely for formulating management practices intended to improve seed

quality given the opportunity of future markets rewarding quality standards (Borja Reis et al., 2022a).

We hypothesize that modern soybean genotypes with high branching capacity have less yield penalties associated with changes in row spacing but present potential variations in seed quality due to changes in partitioning between main stem and branches. We have conducted a two-year trial with two genotypes (one released in 1980 and one in 2013) at four row spacings (0.19, 0.38, 0.76, and 1.52 m) to evaluate impacts on both yield and seed quality (oil and protein concentrations) on the entire plants and in plant fractions, mainly focusing on main stem, and branches.

3.2 Material and methods

3.2.1 Field conditions and management

Field experiments were conducted in 2020 and 2021 at Kansas State University in Manhattan, Kansas, US (39° 12' 22", -96° 35' 37"). The experiment was arranged in a randomized complete block design in a split-plot layout with three replications. The whole plot consisted of row spacing with four levels (0.19, 0.38, 0.76, and 1.52 m), and the subplot consisted of a two-factor factorial of seeding rate with two levels (148 and 445 thousand seeds ha⁻¹) and genotype with two levels (P3981, “Williams 82” released in the 1980s, herein termed as “old,” and P39T67R, released in 2013, herein termed as “modern”) (Corteva Agriscience, Johnston, Iowa, US), both indeterminate growth with 3.9 MG. For each trial year, the soils were a smolam silt loam in 2020 (2.75% OM, 41% clay, 12% sand, 47% silt) and 2021 (2.70% OM, 25% clay, 29% sand, 46% silt). The plots were 10 ft long by 5 ft wide. The planting dates were June 12 in 2020 and May 25 in 2021, with harvesting dates on October 8 in 2020 and October 7 in 2021. Seasonal precipitation from planting to harvesting was 351 mm in 2020 and 443 mm in

2021. Seeds were not treated nor inoculated before planting, and weeds were manually controlled throughout the season.

3.2.2 Field determinations

At R2 (full flowering) and R5 (beginning of seed filling) growth and development stages (Fehr and Caviness, 1977), an area of 0.81 m² per plot was hand-cut, and the plants were counted and weighed. Three representative plants were selected, weighed and oven-dried (65°C) until constant weight to obtain dry weight. Between R2 and R6 phenological stages, a Go-Pro Hero 5 camera placed at 2 m parallel to the soil was used to take weekly pictures of the canopy. Pictures were processed with Canopeo (Patrignani and Ochsner, 2015) to obtain the fractional of green canopy coverage (percentage basis).

At physiological maturity (R8), an area of 0.81 m² per plot was hand-cut, the plants were then counted, and oven-dried until constant weight to obtain the final seed weight. In addition, three plants were hand-cut, and the seeds from the main stem and branches were separated and oven-dried until constant weight to determine the seed weight. The seeds from the branches and main stem were ground to a particle size of 0.1 mm, and protein and oil concentration (percentage basis) were determined using near-infrared spectroscopy (NIR) with the Perten DA7200 Feed Analyzer (Perten Instruments, Stockholm, Sweden). Seed quality per plant was calculated as a weighted average between the seed quality components (%) and yield (%) from the branch and main stem plant fractions of the three plants harvested at the R8 stage.

3.2.3 Statistical analysis

The growing conditions in 2021 were more favorable and less variable compared to 2020, attributed to increased seasonal precipitation and an earlier planting. The result was lower biomass in R2 and R5 (Table 2), lower canopy coverage (Appendix Figure B.2), and lower

average yields (Figure 1) in 2020 compared to 2021. Hence, we analyzed the data separately for 2020 and 2021 due to the strong “by year” interaction effect. Additionally, density plots of the obtained plant densities were similar for both seeding rates in both years (Appendix Figure B.1). Plant density was close to the target for the lower seeding rate, with 162 and 164 thousand plants ha⁻¹ in 2020 and 2021, respectively. On the other hand, the high seeding rate resulted in a plant density 50% below the target, with 207 and 221 thousand plants ha⁻¹ in 2020 and 2021, respectively, mainly due to challenging early season and planting conditions. Therefore, the seeding rate treatment did not result in a factor impacting the main tested traits (yield and seed quality) and was not considered for the analysis.

Generalized linear models (GLM) were used to explain the variation in yield at the plot level, from main stem and branches, and plant biomass at R2 and R5 stages as a function of row spacing, genotype, and their interaction using a gamma distribution. GLMs were also utilized to model oil and protein concentration (%) in the whole plant and partitioned into main stems and branches as a function of row spacing, genotype, and their interaction, using a beta distribution. GLMs were fitted using the *glm* function from the *stats* package (Wood, 2011) in R programming (R Core Team, 2022). The statistical model is as follows:

$$\begin{aligned}
 & [y_{ij} | \mu_{ij}, \psi] \\
 & g(\mu_{ij}) = x_{1i} \cdot \beta_1 + x_{2i} \cdot \beta_2 + \dots + x_{8i} \cdot \beta_8 \\
 & y_{ij} \sim \text{Gamma}(\mu_{ij}, \psi_{ij}) \\
 & y_{ij} \sim \text{Beta}(\mu_{ij}, \psi_{ij}) \\
 & g(u) = X_i \beta
 \end{aligned}$$

where y_{ij} is the response variable (seed yield, biomass at R2 and R5 phenological stages, kg ha⁻¹; and protein and oil concentration, %) for the i^{th} genotype and the j^{th} row spacing, conditional on

an expected value μ and dispersion ψ . The response variables in kg ha⁻¹ follow a gamma distribution with an identity link function; and the response variables in % follow a beta distribution with a logit link function (i.e., $g(\cdot) = \text{logit}(\cdot)$).

To quantify uncertainty in our estimates, we obtained 999 bootstrapped samples using a second-order balanced design of bootstrap simulations (Graham et al., 1990). This involved sampling at the row spacing by genotype level to respect the treatment structure. We calculated the median and the 95% confidence interval (CI_{2.5-97.5}) as derived quantities for estimation of the expected value (effect size) and its uncertainty. Following guidance from the American Statistical Association regarding the use of p-values (Wasserstein and Lazar, 2016), our statistical modeling aimed not to test for ‘statistical significance’, but rather to quantify the effect size and our uncertainty around those estimates. Therefore, we report effect sizes and CI_{2.5-97.5} rather than test for statistical significance. Additionally, density distributions of the difference between the modern and old genotypes at each row spacing are presented for each model to visualize the difference in effect sizes (Appendix Figures B.3, B.4, B.5, and B.6).

3.3 Results

In the 2020 season, both genotypes had similar yields at row spacings of 0.19, 0.38, and 0.76 m with a median of 2481 kg ha⁻¹, while there was a clear reduction at 1.52 m with a median of 1451 kg ha⁻¹ (Figure 1A). In 2021, similar to 2020, the old genotype had similar yields for row spacing levels of 0.19-0.76 m with a median of 2315 kg ha⁻¹, while 1731 kg ha⁻¹ was recorded at 1.52 m. In contrast, the modern genotype in 2021 presented greater yields, with the narrow rows (0.19 and 0.38 m) producing the highest yields with a median of 3584 kg ha⁻¹ and wider rows (0.76, 1.52 m) yielding 2583 kg ha⁻¹ (Figure 1B, Appendix Figure B.3).

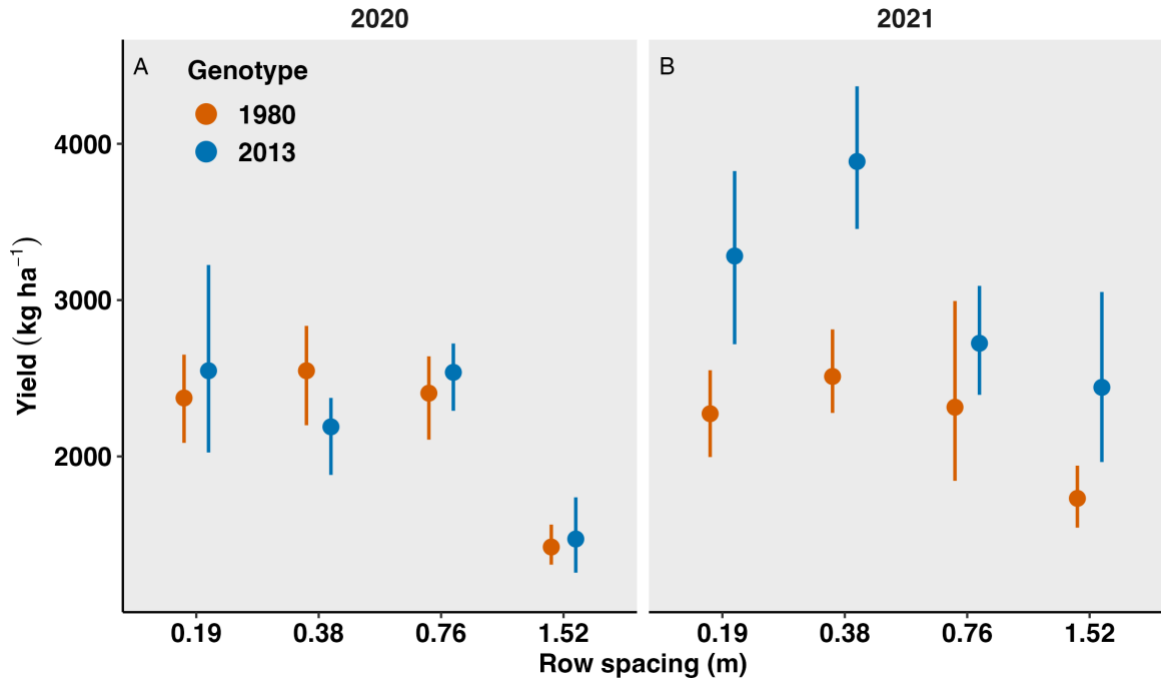


Figure 3.1 Soybean seed yield (kg ha⁻¹) of old (1980) and modern (2013) soybean genotypes at row spacings of 0.19, 0.38, 0.76, and 1.52 m in 2020 (panel A) and 2021 (panel B) growing seasons. The error bars represent the 95% confidence intervals around the estimated expected value.

Overall, plant biomass was higher in 2021 compared to 2020, with no differences observed between genotypes or row spacings (Table 1). Plant biomass at R2 was higher in 2021 (median of 6247 kg ha⁻¹) than in 2020 (median of 3716 kg ha⁻¹). Similarly, biomass at R5 was higher in 2021 (median of 8855 kg ha⁻¹) compared to 2020 (median of 5348 kg ha⁻¹).

Table 3.1 Soybean biomass (kg ha⁻¹) in R2 (full flowering) and R5 (beginning seed filling) phenological stages of old (1980) and modern (2013) genotypes at row spacings of 0.19, 0.37, 0.76, and 1.52 m, in 2020 and 2021. The numbers between parenthesis indicate the 95% confidence interval around the estimated expected value.

Year	Genotype	Row Spacing (m)	Biomass (kg ha ⁻¹)	
			R2	R5
2020	1980	0.19	3926 (3372 - 4493)*	5547 (4280 - 7252)

2021	2013	0.38	3615 (2347 - 5664)	4802 (3637 - 5498)
		0.76	3891 (2882 - 4802)	5925 (5118 - 6831)
		1.52	3817 (3137 - 4439)	4601 (4064 - 5114)
		0.19	3284 (2866 - 3652)	6063 (5672 - 6457)
		0.38	2714 (1970 - 3424)	5222 (3438 - 7107)
		0.76	4437 (3847 - 5090)	5578 (4534 - 6634)
		1.52	2249 (1866 - 2635)	3460 (2726 - 4178)
	1980	0.19	6704 (5316 - 7965)	8718 (7732 - 9846)
		0.38	7639 (5775 - 9687)	8992 (8278 - 9722)
		0.76	5913 (4640 - 7515)	8153 (7039 - 9498)
		1.52	5366 (3910 - 6761)	7006 (5616 - 8635)
	2013	0.19	6581 (5643 - 7642)	9944 (8184 - 11958)
		0.38	8345 (7122 - 9476)	9454 (7369 - 11032)
		0.76	5519 (4048 - 6927)	9095 (8122 - 10359)
		1.52	4627 (3762 - 5783)	8039 (5749 - 11140)

Yield from branches was similar across row spacings and genotypes in 2020, with a median of 44% (Figure 2A). In 2021 however, the modern genotype had a higher contribution from branches (36%) compared to the old genotype (18%) (Figure 2B). There were no differences in yield (kg ha^{-1}) from branches and main stem across genotypes and row spacings in 2020 (Figure 2C). However, in 2021, the main stem showed a similar contribution for both genotypes with greater yields under narrow rows (0.19 and 0.38 m) at a median of 2226 kg ha^{-1} , compared to wide rows (0.76 and 1.52 m) at 1663 kg ha^{-1} . Remarkably, in 2021, the yield derived from branches was three-times greater for the modern compared to the old genotype

(1113 vs. 379 kg ha⁻¹) (Figure 2D, Appendix Figure B.4). Pod number in branches across nodes of the main stem were consistently concentrated in the bottom section (nodes 1 to 6), with more pods observed for the modern genotype in 2021 (Figure 2, insets A1 and B1).

Protein concentration was similar across genotypes and row spacings in 2020, with an overall median of 39.8% (Figure 3A). In 2021, the old genotype had a median protein concentration of 40.1%, while the modern genotype had consistently lower values with a median of 37.1%. Nevertheless, the modern genotype maintained protein levels despite the high yields observed under narrow rows (0.19 and 0.38 m) (Figure 3B). The seed oil concentration in 2020 had an overall mean of 21.1% across treatments (Figure 3C). The old genotype in 2021, similar to 2020, had a median of 20.5%, while the modern genotype tended to have greater oil concentration (median of 21.5%) (Figure 3D, Appendix Figure B.5).

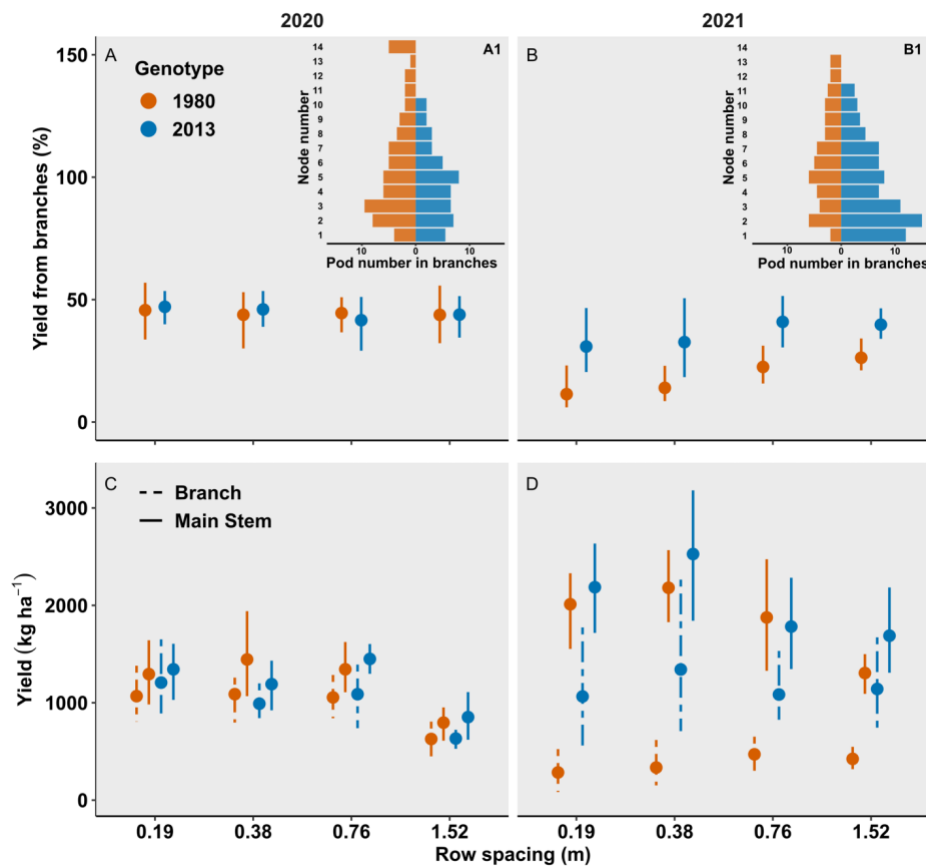


Figure 3.2 Soybean seed yield from branches (%; A and B), and from branches and main stems (kg ha⁻¹; C and D; dashed and full line, respectively) of old (1980) and modern (2013) genotypes at row spacings of 0.19, 0.38, 0.76, and 1.52 m, in 2020 (panels A and C) and 2021 (panels B and D) growing seasons. The error bars represent the 95% confidence intervals around the estimated expected value. Insets A1 and A2 present the median number of pods in branches across nodes of the main stem averaged over all tested row spacings in 2020 and 2021 growing seasons, respectively.

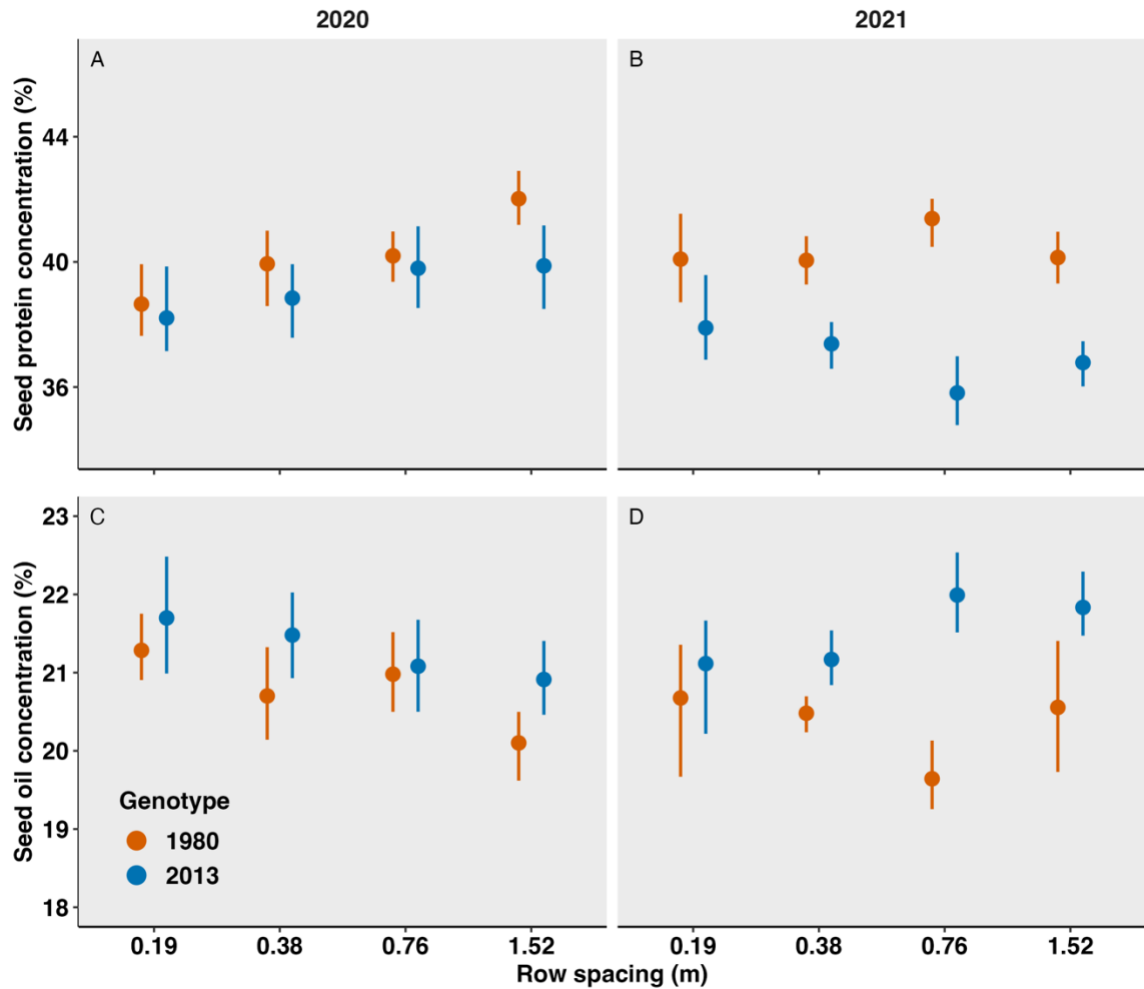


Figure 3.3 Soybean seed oil (panels A and B) and protein (panels C and D) concentration (%) in the whole plant of the old (1980) and modern (2013) genotype at row spacings of 0.19, 0.38, 0.76, and 1.52 m, in 2020 (panels A and C) and 2021 growing seasons (panels B and D). The error bars represents the 95% confidence intervals around the estimated expected value.

No apparent differences were observed in seed quality (or composition) between branches and main stems, except for protein in 2020, where the main stem had a median of 40.8% compared to 38.3% in branches (Figure 4A to 4D, Appendix Figure B.6).

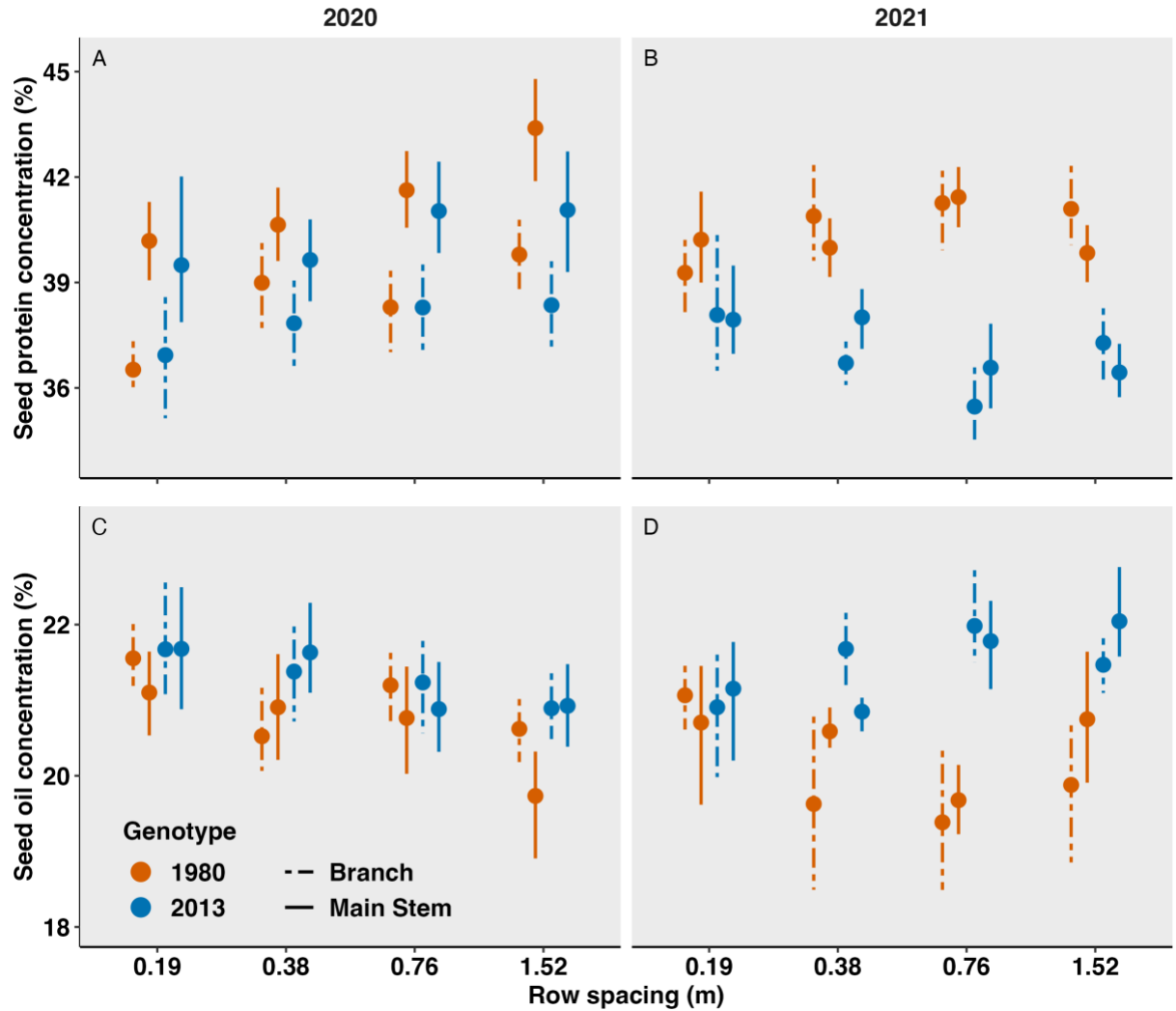


Figure 3.4 Soybean seed oil (panels A and B) and protein (panels C and D) concentrations (%) in branches (dashed lines) and main stems (full line) (C and D) of the old (1980) and modern (2013) genotype at row spacings of 0.19, 0.38, 0.76, and 1.52 m, in 2020 (panels A and C) and 2021 growing seasons (panels B and D). The error bars represents the 95% confidence intervals around the estimated expected value.

3.4 Discussion

This study provides evidence of the documented genetic yield gain for modern soybean genotypes (Rincker et al., 2014b; Rowntree et al., 2013; Suhre et al., 2014a), reflecting greater yield differences under narrow (<0.76 m) relative to wide row spacing (>0.76 m). Our results also expand on the ability of modern genotypes to tolerate plant-to-plant competition under high

plant densities (Cober et al., 2005; Suhre et al., 2014a) to that observed in the intra-row under wide row spacings. Furthermore, the improved allocation of biomass to reproductive structures without major changes in vegetative biomass and time to canopy closure (Koester et al., 2016, 2014; Tamagno et al., 2020) are key factors sustaining yield gains for modern soybeans. Finally, the absence of a yield-protein trade-off in modern soybeans, likely due to high soil NO_3^- and leaf N content late in the season, offers promising economic advantages for the seed quality-driven market (Carciochi et al., 2023a; Ergo et al., 2024a).

Our findings align with previous research (Suhre et al., 2014a) indicating that branches contribute more to the yield of modern soybean genotypes. Branches can drive yield increases particularly when soybeans are planted early (Board and Settini, 1986b), and under narrow row spacings (Herbert and Litchfield, 1982; Settini and Board, 1988b). In this sense, our study confirms that branches tend to concentrate in the bottom and middle section of the canopy when growth is maximized (Herbert and Litchfield, 1982; Settini and Board, 1988b), likely supported by enhanced light penetration (Willcott et al., 1984). Light interception has linearly increased with newer genotype releases (Koester et al., 2014), enhancing seed protein concentration at the base of the main stem (Huber et al., 2016). Therefore, branches are crucial for stabilizing and increasing yields in modern soybean genotypes, significantly impacting seed quality composition (Moro Rosso et al., 2021).

From a seed quality perspective, maintaining protein levels at narrow rows despite higher yields could be linked to a better light distribution (Bellaloui et al., 2012; Huber et al., 2016) and longer green canopy duration (Appendix Figure B.2) relative to wide rows. Furthermore, although research has shown that seed protein concentration can be lower on branches than main stem (Moro Rosso et al., 2021), our results suggest that mainly low-yielding conditions favored

less protein on branches. The latter could be linked to the delayed reproductive development of branches (Munier-Jolain et al., 1994) and the fast senescence of the canopy (as for the 2020 season), suggesting a greater sensitivity to stress conditions on seed from branches. Studying seed quality dynamics of canopy portions (i.e., branch and main stem) could aim to identify farming practices to maximize seed quality and target protein and oil markets (Borja Reis et al., 2022a).

Despite the limitations of this study associated with the number of genotypes tested, these results report on the contribution of row spacing as a key management for achieving better yields for modern soybean genotypes, with implications on seed quality. Further research should explore a larger number of genotypes and investigate in detail changes linked to light interception and biomass conversion efficiencies that could underpin underlying physiological processes. Branches and main stems should be considered as modulators of seed quality (Moro Rosso et al., 2021) with potential variations due to water and temperature stresses at different timings during reproductive development (Rotundo and Westgate, 2009a). Canopy architecture traits, mainly branch angle and leaf shape, are key determinants of canopy closure (Virdi et al., 2023). Hence, branching architecture and light interception could be exploited by breeding to enhance not only yield but also protein concentration.

3.5 Conclusion

Our study revealed that soybean yield gain could be greater at narrow row spacings (<0.76 m) compared to the documented at 0.76 m and reinforces the important contribution from branches to yield. Protein concentrations for the modern genotype were maintained despite high yields under narrow rows. Future research should evaluate a larger number of genotypes and study different canopy architecture traits.

Chapter 4 - Assessing on-farm drivers of soybean seed protein, and oil concentration

**Submitted to Field Crops Research*

Valentina M. Pereyra, Carlos Hernandez, Trevor Hefley, Ajay Sharda, P.V. Vara Prasad, Laila Puntel, Guillermo R. Balboa, David Moseley, Jon Trenton Irby, Eros Francisco, André F. B. Reis, Aaron Prestholt, Joshua McDanel, Matthew Carroll, Shaun N. Casteel, Randall Pearson, Joshua Pristolas, Maninder Pal Singh, Peter Kovács, Michael H. Ostlie, John Fulton, Peter Kyveryga, and Ignacio A. Ciampitti

Abstract

Seed yield is the main goal for soybean (*Glycine max* L.) farmers, while soybean meal with a high protein is the target for the processing industry. Soybean oil, historically considered a byproduct, is now in higher demand due to growing interest in renewable diesel. Understanding how environmental variables influence soybean yield and quality (seed oil and protein concentration) in different regions is essential for developing a more diverse production chain and strong industry. This study aims to i) assess the importance of environmental factors in predicting seed oil and protein concentration and reported yields, ii) identify regions with different yield and seed quality levels, and iii) explore key factors linked to these variables across regions. A standardized protocol was implemented to collect seed and soil samples from 235 fields in 13 US states during 2022 and 2023. We collected yield and crop management data via survey and retrieved crop phenological stages using crop growth model simulations (DSSAT) to summarize weather data during vegetative, reproductive pre seed-filling, and reproductive seed-filling stages. Weather, soil, and seasonal variables predicted yield and seed oil concentration

with better accuracy (R^2 0.56 and 0.39, respectively) than for seed protein concentration (R^2 0.09). Greater seed protein concentration in the northern region (39.5 %) compared to the Corn Belt region (38.2 %) is consistent with recent studies showing improved quality in northern soybeans. In the north, low temperatures during seed fill, linked to late planting dates or long-maturity groups, constrained yield and seed oil concentration. While in the south, high temperatures in the early season significantly limited yield where late sowing (after June 1st) shortened the total season length. This study offers new insights into the variation of on-farm seed quality (seed protein and oil concentrations), utilizing a large dataset from 235 fields in 13 US states during 2022 and 2023 growing seasons. The findings of this study reveal regional differences in soybean quality, with greater seed protein concentration, but lower seed oil concentration in the north relative to the other regions and offers insights on environmental factors associated. These findings serve as a reference for on-farm seed quality variation, aiding price determination and decision-making to enhance yield and seed quality across US soybean regions.

4.1 Introduction

The United States (US) produced 39% of the global soybeans (397 million Mg) in 2023, with a record volume of seed crushed (55% of the production) and exported as soybean meal (13.2 million Mg). Additionally, the production capacity of renewable diesel is expected to grow from 2.6 to 22 billion L year⁻¹ between 2022 and 2025 (EIA, 2023; USDA FAS, 2024). The US soybean meal, although with less protein concentration, has nutritional advantages over Brazil and Argentina (Thakur and Hurburgh, 2007; USSEC, 2021) due to a better amino acid profile and increased carbohydrates. However, the high regional variability in protein levels presents a

challenge for price differentiation among farmers and merchandisers. Regional and international beliefs and perceptions about quality ultimately define market requirements and price differentials received by farmers (Bullock et al., 2024; Hertsgaard et al., 2019; William et al., 2020).

Historical data indicates that the southern US states exhibit greater protein concentrations (Assefa et al., 2018; Hurburgh et al., 1990; Rotundo et al., 2016) and higher price differentials (Hertsgaard et al., 2019) than in the Corn Belt and northern regions. However, in the northern region, protein levels are not consistently lower than the Corn Belt with a high year-to-year variability (Cerrudo et al., 2023; Rotundo et al., 2016), which highlights the challenge of generalizing a seed quality level and the need to define regional differences more accurately. The availability of quantitative information on regional and intra-regional quality variation could prove beneficial for informed decision-making and develop a more structured soybean pricing formation. Indeed, a substantial body of research pointed to the need to integrate weather, soil, and crop management data to further understand and predict more accurately the composition of soybean seeds at a regional level (Assefa et al., 2019; Borja Reis et al., 2022b; Rotundo et al., 2016).

Environmental factors modulate the regional variation in seed composition due to water availability, temperature, and soil nutrients for a given genotype. Drought stress during the seed filling period usually leads to higher seed protein concentration due to reduced seed size and less detrimental effect on seed protein relative to oil and carbohydrate accumulation (Rotundo and Westgate, 2009b). Similarly, heat stress has been observed to increase protein concentration, though this phenomenon is not directly related to increased protein synthesis (Egli and Bruening, 2004; Pipolo et al., 2004; Rotundo and Westgate, 2010, 2009b). Oil accumulation is highly

dependent on current photoassimilates and is therefore negatively affected by heat and drought stress (Rotundo and Westgate, 2009b), which shortens the seed fill duration. Soil nutrients availability, particularly nitrogen (N) during the seed-filling period can restrict protein concentration when the N from biological fixation and mineralization are unable to meet the plant N demand (Cafaro La Menza et al., 2020, 2017; Carciochi et al., 2023b; Moro Rosso et al., 2023). The impact of environmental factors has been examined in controlled studies at limited regional scales. However, significant variability exists between farms, and the assessment of changes in seed quality at the regional scale remains a challenging and ongoing area of research (Hernandez et al., 2023).

Our research gathered data on farmer fields across 13 US states during the 2022 and 2023 growing seasons. We followed a standardized protocol across all farms to collect representative seed and soil samples in-situ for analysis. In addition, we retrieved relevant crop management and yield data via surveys and linked all the data with seasonal weather variables to develop a large on-farm database, which included a total of 235 soybean fields. Following this rationale, our objectives were to i) assess the importance of environmental factors in predicting seed oil and protein concentration and reported yields, ii) identify regions with different yield and seed quality levels, and iii) explore key environmental factors linked to these variables across regions to further understand seed oil and protein concentration differences across defined geographical regions.

4.2 Material and Methods

4.2.1 On-farm database on soybean yield, management practices, soil, and seed quality

Two years (2022 and 2023) of data were collected from 235 farmer fields across 13 US states in collaboration with local universities including North Dakota, South Dakota, Michigan,

Ohio, Iowa Soybean Association, Nebraska, Kansas, Missouri, Illinois, Indiana, Louisiana, Mississippi, and Alabama. We collected data on location, soybean yield (reported at 130 g kg⁻¹ moisture), crop management (i.e., sowing date, harvest date, seeding rate, row spacing, maturity group (MG), irrigation), and factors limiting yield (i.e. diseases, late planting date, flood) via online surveys. Field size ranged from 2.8 to 455 ha with a median field size of 24.3 ha. Additionally, we collected in-situ soil and seed samples through a standardized protocol and analyzed them in a single lab at Kansas State University. Before harvest, we selected between 6 to 12 sampling points distributed across the whole field for collecting the plants within 54 cm of one row and forming a representative seed sample per field. Seed protein and oil concentration (reported on a dry basis) were analyzed in whole seed samples using a DA 7250TM near-infrared spectroscopy (NIRS) analyzer (Pertin Instruments, Hägersten, Sweden). We calculated the median of oil and protein from all seed samples to characterize each field. For soil samples, we took four cores (0-60 cm depth) at each sampling point, mixed homogeneously, and obtained one composite soil sample per field. Lastly, we analyzed soil samples for texture (sand, silt, clay, %), organic matter (OM, %), cation exchange capacity (CEC, meq 100 g soil⁻¹), pH, NO₃⁻ (mg kg⁻¹), and P (mg kg⁻¹).

4.2.2 Weather and crop modelling

We retrieved daily weather data on precipitation, temperature, and radiation from the Iowa Environmental Mesonet (<https://mesonet.agron.iastate.edu/request/coop/fe.phtml>). We used the nearest weather station to each field at a distance less than 40 km. We gathered relative humidity from the NASA-POWER project (<https://power.larc.nasa.gov>). We calculated vapor pressure deficit as a proxy for drought stress and reported precipitation plus irrigation (mm) for irrigated fields. The Decision Support System for Agrotechnology Transfer (DSSAT) model

(Hoogenboom et al., 2019; Jones et al., 2003) was used to estimate the growth stages (Fehr and Caviness, 1977) from VE (emergence) to R7 (physiological maturity) based on site-specific weather, soil, and management. The observed season length presented an adequate level of agreement (MAE = 9.7 days) compared to the season length derived from DSSAT model (Appendix Fig. C.1). We used the simulated phenology to summarize weather variables for the periods: vegetative (v, from VE to beginning of flowering [R1]), reproductive pre-seed filling (r, from R1 to full pod [R5]), and reproductive seed filling (s, from R5 to R7). In addition, we calculated the date of the year (doy) of R1, R5, and R7 for their reported correlation with seed quality (Salmerón et al., 2022). Finally, we calculated the number of days during seed fill (s) with maximum temperatures above 30°C as in (Rotundo and Westgate, 2010) and, similarly, with minimum temperatures below 15°C to account for heat and cold stress, respectively.

4.2.3 Data analysis

4.2.3.1 Variable importance assessment

We explored the association between seasonal characteristics, weather, and soil variables with the response variables (yield, oil, and protein concentration) using the tree-boosting system XGBoost (Chen and Guestrin, 2016). We chose XGBoost due to its ability to capture non-linear relationships, capture complex interactions, and ranking the importance of each predictor adds interpretability to the model. We excluded highly correlated covariates with similar meaning that did not convey extra information using the Pearson's correlation coefficient (r) ($|r| > 0.75$) (Appendix Fig. C.2). For instance, sand, silt, and clay were all highly correlated, and we kept only the latter for its affinity to CEC (also dropped) and influence on fertility. Similarly, MG, planting date, radiation, and season length were highly correlated, hence, we only kept MG and season length due to its easier interpretation and comparison. Weather variables in the periods v,

r, and s, although correlated, were not dropped since we wanted to explore their association with the response variables. We did not consider the nutrient fertilizer applied (N, P, K, S) due to an unbalanced number of observations in each class. We also dropped NO_3^- due to its high in-season natural dynamic and late sampling date (harvest). We selected 21 continuous variables and presented a complete overview of soil, season characteristics, and weather variables for each state in Table 4.1.

We generated 13 stratified random samples for model evaluation at the year and state level ($n=25$, 11% for testing), with the remaining data reserved for model fitting ($n=210$, 89% for training). For each training set, we performed a grid search to find the best combination of hyperparameters. The hyperparameters optimized were: i) *nrounds*, the number of boosted decision trees; ii) *eta*, the learning rate or step size shrinkage; iii) *maxdepth*, the maximum depth of a tree, where greater depth increases complexity; and iv) *subsample*, the ratio of training instances used per tree. We fixed *gamma* (the minimum loss reduction required for further partitioning on a tree node) to 0 and *colsample_bytree* at 0.7 (70% of features randomly selected) while leaving other parameters at their default. We evaluated the model performance across the 13 testing sets using the concordance correlation coefficient (CCC), mean absolute error (MAE), and coefficient of determination (R^2). We selected the optimal hyperparameter combination based on the highest CCC, which reflects precision and accuracy. Finally, we calculated the feature importance on each training set as the average gain in prediction across all splits the feature is used in (Chen and Guestrin, 2016). For an intuitive explanation of how the feature importance is calculated, we refer the reader to Chen & Guestrin, 2016, and XGBoost tutorials (XGBoost Developers, 2022). We ran the model with the function *xgboost* and obtained the

feature importance with *xgb.importance* from the package *xgboost* (Chen et al., 2014) in R Studio (Posit team, 2023).

4.2.3.2 Regional influence and relationships

We aimed to identify the most important variables for predicting yield, protein, and oil in soybean, exploring a region spanning latitudes 30.5°N to 47.7°N, and longitudes 101.2°W to 82.4°W. This region exhibits considerable variability in growing conditions related to soil, weather, and season characteristics, with many variables highly correlated to latitude and longitude (Supplementary Fig. 2). To capture this regional variability, we applied the spatial fuzzy c-means (SFCM) clustering algorithm (Cai et al., 2007) using total radiation (MJ m^{-2}) and precipitation (mm), median vapor pressure deficit (VPD, kPa) and temperature ($^{\circ}\text{C}$) during the growing season, yield (kg ha^{-1}), protein and oil (%). We chose SFCM because it allows each observation to have a probability of belonging to a cluster rather than a fixed assignment and incorporates a spatial constraint for better space segmentation. We set the number of clusters to three based on the Silhouette (Rousseeuw, 1987), the Xie and Benni indexes (Xie and Beni, 1991), and experts criteria to allow for meaningful regional interpretation. We used the *SFCMeans* function from the *geocmeans* package (Jérémy Gelb, 2021) with 10 nearest neighbors, a fuzziness coefficient of 1.5, and alpha (influence of spatial autocorrelation) of 1.5.

Finally, we explored the relationship between the three most important variables predicting yield, oil, and protein, using the clusters to detect regional effects and derive more interpretable results. We fit generalized additive models as flexible smoothed effects via thin-plate regression splines using a Gaussian distribution and calculated 95% confidence interval around the estimated expected values. We used the function *gam* from the *mgcv* R package (Wood, 2017, 2003) and evaluated the model performance using R^2 and MAE. All the

visualizations were made using *ggplot2* (Hadley Wickham, 2016) and *ggpubr* (Alboukadel Kassambara, 2023) in R Studio (R Core Team, 2024).

4.3 Results

A total of 235 farmer fields were included in the database, comprising yield, seed protein and oil concentrations, soil, and crop management data. We collected 43% of the fields in 2022 and 57% in 2023 (Fig. 1). The distribution of values for yield, seed oil and protein concentrations were similar in both years. Overall, yield resulted in a median of 4.04 Mg ha⁻¹ ranging between 1.88 and 5.38 Mg ha⁻¹ as defined by the 95th percentile (P2.5 and P97.5). Oil concentration ranged between 17.9 and 22.9% with a median of 20.8%. Protein concentration had a median of 38.6% and ranged between 35.5 and 41.8%. A complete overview of weather, soil, and season characteristics for each state is presented in Table 1.

Table 4.1 Summary of season characteristics, soil, and management variables for each state showing the median and range of observations for each variable.

	State															Factor
	AL	IA	IL	IN	KS	LA	MI	MO	MS	ND	NE	OH	SD			
vDays	38, [35, 60]	46, [20, 63]	48, [34, 59]	45, [36, 58]	42, [36, 56]	34, [29, 36]	47, [13, 71]	52, [29, 64]	38, [32, 46]	32, [26, 41]	47, [38, 60]	48, [12, 63]	37, [31, 41]		Season	
rDays	33, [25, 37]	29, [19, 37]	32, [26, 36]	27, [24, 31]	24, [22, 37]	33, [31, 41]	26, [21, 29]	34, [25, 38]	38, [27, 42]	22, [18, 27]	28, [24, 37]	27, [21, 31]	26, [21, 32]		Season	
sDays	41, [35, 50]	39, [34, 44]	43, [40, 45]	41, [36, 44]	34, [29, 44]	43, [39, 59]	39, [30, 43]	43, [34, 50]	44, [31, 50]	35, [28, 38]	36, [26, 43]	42, [40, 44]	36, [31, 40]		Season	
MG	5, [4, 7]	3, [2, 4]	4, [3, 4]	3, [2, 4]	4, [4, 5]	5, [5, 6]	2, [2, 3]	4, [2, 5]	5, [5, 6]	0, [0, 1]	3, [2, 4]	3, [3, 4]	2, [1, 2]		Season	
Clay (%)	18, [8, 56]	28, [17, 33]	18, [13, 38]	18, [8, 28]	29, [2, 42]	18, [8, 39]	11, [6, 27]	24, [12, 41]	46, [10, 78]	20, [12, 35]	27, [6, 35]	26, [16, 40]	26, [15, 28]		Soil	
OM (%)	2.5, [0.8, 5.0]	3.6, [2.4, 5.2]	2.4, [1.9, 5.3]	2.9, [1.8, 4.9]	2.5, [0.8, 3.1]	1.5, [1.0, 2.6]	2.7, [1.6, 4.4]	1.8, [1.0, 3.7]	2.8, [2.1, 3.7]	3.5, [2.7, 5.0]	2.4, [0.9, 3.6]	3.2, [2.3, 4.1]	2.7, [2.0, 3.6]		Soil	
pH	5.8, [5.1, 6.7]	6.3, [5.2, 7]	6, [5.6, 6.5]	6.2, [5.1, 6.9]	6.7, [4.9, 8.2]	6.6, [4.9, 8]	6.4, [5.6, 7.1]	6.2, [4.4, 7]	6.8, [5.8, 7.4]	7.1, [6.5, 8.1]	6.8, [6, 8.2]	6.7, [5.6, 7.1]	7.4, [5.8, 7.7]		Soil	
P (ppm) ^{1,2}	69, [11, 350]	38, [9, 97]	30, [7, 176]	37, [14, 211]	27, [15, 79]	22, [8, 54]	40, [13, 189]	22, [11, 60]	40, [19, 82]	56, [8, 96]	19, [12, 53]	31, [13, 68]	12, [5, 41]		Soil	
vPrec (mm)	137, [37, 244]	156, [68, 289]	113, [70, 195]	101, [28, 244]	146, [28, 326]	88, [61, 135]	109, [62, 227]	178, [73, 287]	189, [75, 304]	95, [56, 135]	132, [80, 277]	177, [54, 298]	85, [60, 115]		Weather	
vT _{min} (°C)	19.4, [12.9, 22.6]	15.4, [12.3, 18.8]	16.6, [14.2, 21.0]	15.3, [13.3, 17.8]	20.4, [15.1, 21.4]	20.2, [16.3, 23.4]	13.3, [11.2, 18.6]	16.9, [12.9, 20.1]	19.6, [15.3, 22.5]	13.9, [11.0, 15.6]	14.4, [10.5, 16.9]	16.4, [14.5, 19.0]	16.7, [14.1, 18.3]		Weather	
vT _{max} (°C)	31.5, [25.3, 35.0]	27.9, [26.4, 31.1]	28.6, [27.1, 33.3]	28.5, [26.3, 30.6]	33.0, [27.7, 35.5]	32.1, [28.1, 34.1]	27.1, [25.2, 29.5]	27.9, [24.0, 32.5]	30.9, [28.5, 34.1]	27.4, [25.4, 28.1]	28.4, [27.0, 32.0]	28.8, [27.2, 30.6]	28.2, [25.6, 30.5]		Weather	
vPPD (kPa)	1.0, [0.5, 1.5]	0.8, [0.6, 1.1]	0.7, [0.6, 0.9]	0.7, [0.5, 0.9]	1.4, [0.8, 1.9]	0.8, [0.5, 1.1]	0.6, [0.6, 0.8]	0.7, [0.4, 1.0]	0.7, [0.5, 1.6]	0.8, [0.7, 0.9]	1.2, [0.8, 1.6]	0.7, [0.6, 0.8]	1.0, [0.8, 1.2]		Weather	
rPrec (mm)	134, [27, 266]	44, [3, 175]	225, [58, 286]	102, [41, 368]	36, [25, 106]	106, [45, 154]	72, [30, 124]	94, [16, 339]	102, [75, 203]	33, [10, 53]	76, [22, 70]	80, [28, 155]	75, [40, 131]		Weather	
rT _{min} (°C)	22.2, [18.0, 23.1]	17.9, [16.7, 21.6]	19.6, [17.0, 22.9]	17.6, [14.8, 20.7]	19.1, [17.5, 21.0]	23.0, [21.3, 23.7]	16.1, [14.3, 21.5]	20.0, [14.6, 21.4]	22.4, [20.8, 23.5]	13.8, [12.0, 17.3]	17.6, [15.1, 19.7]	19.0, [15.7, 20.5]	18.6, [17.8, 20.0]		Weather	
rT _{max} (°C)	32.9, [30.5, 35.6]	30.4, [29.0, 34.8]	30.6, [29.2, 34.7]	28.9, [26.8, 31.6]	34.1, [31.0, 35.4]	33.8, [32.2, 35.3]	28.6, [27.0, 34.6]	31.6, [29.5, 35.2]	34.3, [32.5, 34.9]	27.9, [26.7, 29.0]	32.2, [29.8, 34.2]	29.1, [26.7, 31.4]	31.0, [30.3, 31.5]		Weather	
rVPD (kPa)	1.0, [0.8, 1.4]	1.2, [0.8, 1.4]	0.9, [0.6, 1.1]	0.8, [0.6, 0.9]	1.7, [1.1, 1.8]	1.1, [0.9, 1.3]	0.8, [0.6, 1.2]	1.0, [0.7, 1.4]	1.4, [1.1, 1.5]	0.8, [0.8, 1.1]	1.5, [1.1, 1.8]	0.7, [0.6, 0.8]	1.2, [1.1, 1.6]		Weather	
sPrec (mm)	171, [59, 342]	104, [52, 219]	90, [74, 183]	121, [50, 168]	42, [33, 153]	306, [112, 390]	100, [40, 192]	116, [38, 190]	192, [56, 323]	41, [9, 68]	48, [15, 127]	110, [71, 169]	124, [82, 172]		Weather	
sT _{min} (°C)	20.6, [16.3, 22.7]	15.2, [14.4, 21.8]	17.4, [15.2, 20.7]	15.2, [12.8, 19.2]	15.5, [13.0, 18.5]	23.0, [22.0, 23.8]	14.2, [10.8, 23.0]	17.4, [13.8, 22.3]	22.0, [19.0, 23.1]	12.6, [10.3, 14.7]	15.1, [11.7, 17.3]	15.9, [11.2, 23.5]	15.6, [14.7, 16.5]		Weather	
sT _{max} (°C)	30.5, [28.8, 33.3]	28.2, [27.3, 33.1]	28.5, [27.0, 32.3]	27.1, [25.6, 29.6]	31.0, [28.6, 34.7]	33.1, [31.7, 34.1]	26.6, [23.5, 34.5]	29.7, [27.7, 33.8]	32.3, [31.4, 34.4]	27.0, [26.3, 27.7]	31.4, [29.7, 32.2]	27.0, [23.9, 34.5]	28.2, [27.6, 28.6]		Weather	
sVPD (kPa)	0.9, [0.8, 1.0]	1.0, [0.7, 1.5]	0.8, [0.6, 0.9]	0.8, [0.6, 0.9]	1.5, [1.2, 1.8]	0.8, [0.6, 1.1]	0.6, [0.5, 1.1]	1.0, [0.7, 1.2]	1.1, [0.8, 1.4]	0.8, [0.8, 1.0]	1.5, [1.1, 1.7]	0.7, [0.6, 1.0]	1.0, [0.9, 1.2]		Weather	
sDaysT _{max} >30	28, [15, 46]	13, [6, 40]	13, [4, 35]	6, [2, 26]	24, [14, 34]	46, [29, 55]	4, [2, 34]	29, [10, 41]	39, [25, 46]	9, [4, 12]	24, [15, 30]	6, [2, 42]	10, [6, 14]		Weather	
sDaysT _{min} <15	1, [0, 17]	21, [0, 26]	11, [1, 23]	23, [3, 31]	14, [4, 26]	0, [0, 2]	25, [0, 38]	12, [0, 27]	0, [0, 6]	28, [24, 36]	20, [10, 30]	18, [0, 36]	13, [9, 20]		Weather	
Number of days between V6-R1, R1-R5, and R5-R67 (Olay, and Olay, respectively); maturity group (MG), Precipitation (Prec), minimum and maximum temperature (T _{min} and T _{max} , respectively), vapor pressure deficit (VPD), number of days during seed fill with maximum temperatures > 30°C, (sDaysT _{max} >30), number of days during seed fill with minimum temperatures < 15°C, (sDaysT _{min} <15).																
¹ Weather was simulated in vegetative (v), reproductive pre-seed filling (r), and reproductive seed filling (s).																
² Growth stages were simulated with DSSAT.																

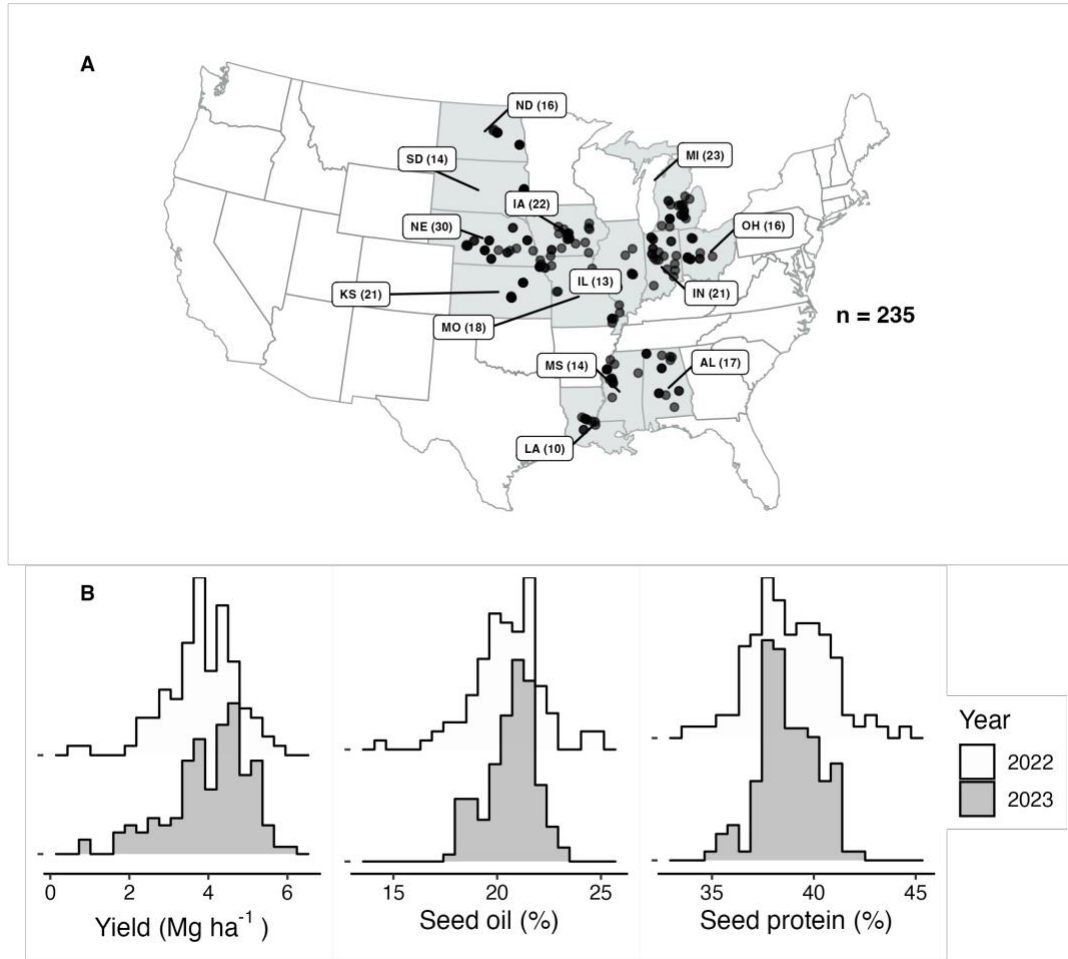


Figure 4.1. Location of the surveyed farmer fields (circles) in the US during the 2022 and 2023 growing seasons, showing the total per state (A). Histogram displaying yield (Mg ha^{-1}), seed oil and protein concentrations (%) in both 2022 and 2023 growing seasons (B).

The xgboost algorithm demonstrated superior accuracy in predicting yield and oil concentration (R^2 of 0.56 and 0.39, respectively) compared to protein (R^2 of 0.09) (Supplementary fig. 3). The predicted expected value exhibited a MAE of 0.52 Mg ha^{-1} for yield, 0.91% for oil, and 1.29% for protein. The variable importance (feature gain) was important to assess the relative contribution of each variable to the overall accuracy obtained by the xgboost algorithm (Fig. 2). In general, weather variables were found to be important for prediction, although the specific timing of their influence varied depending on the response variable under

consideration. For instance, yield was primarily affected by $T_{\max}$ during early-season, maturity group, and vapor pressure deficit (VPD) during seed filling. In contrast, seed oil concentration was mainly influenced by minimum temperatures throughout the reproductive phase and by the duration of the early reproductive phase (R1-R5). Protein concentration, despite exhibiting relatively low predictive power, was related to precipitation throughout the entire growing season.

We have identified three main regions via clustering as the south, center (Corn Belt), and north of the US soybean producing regions (Fig. 3). Yield, as expected, was $\sim 0.82 \text{ Mg ha}^{-1}$ (22%) greater in the Corn Belt (4.44 Mg ha^{-1}) relative to the northern and southern regions (3.58 and 3.67 Mg ha^{-1} , respectively), although more variability was observed in the south. Oil concentration was greater in the Corn Belt and the south both presenting a median of 21.1%, and, as usually reported, lower in the north (19.0%). Protein concentration was lowest in the Corn Belt (38.2%) where higher yields were observed, compared to the northern and southern regions with a median of 39.5 and 39.4%, respectively. We further calculated the performance metrics from the xgboost algorithm by region. Yield was best predicted in the Corn Belt region (R^2 0.67, CCC 0.79) followed by the northern (R^2 0.35, CCC 0.58), and with low performance in the southern (R^2 0.06, CCC 0.23) region. Oil prediction performance was highest in the north (R^2 0.34, CCC 0.53), followed by the Corn Belt (R^2 0.18, CCC 0.39) and lowest in the south (R^2 0.09, CCC 0.29). Lastly, the prediction performance of protein was low with better metrics for the south (R^2 0.05, CCC 0.20) than for the north and center (R^2 0.03 and 0.02, CCC 0.18 and 0.12, respectively). A complete overview of the season, soil, and weather by cluster is described in Table 2.

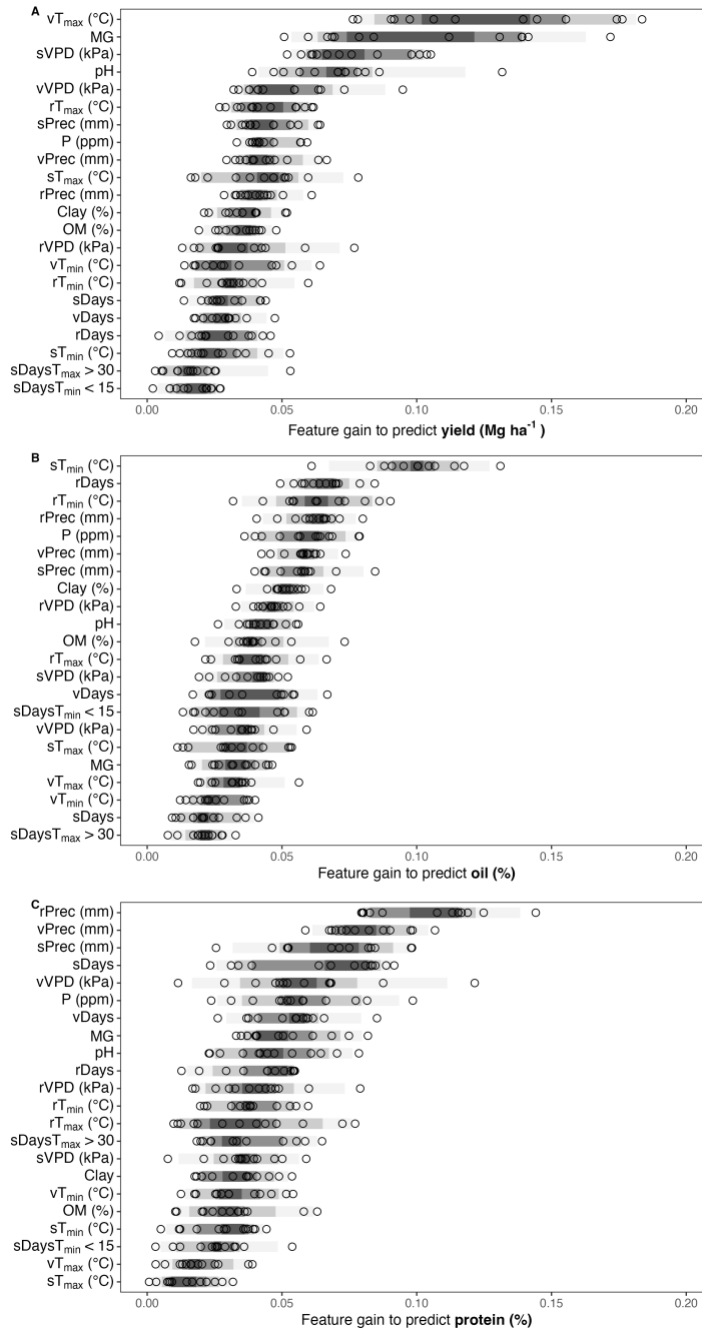


Figure 4.2. Variable importance (feature gain) of XGBoost model to predict yield (Mg ha⁻¹, A), oil (%), B), and protein (%), C) concentration based on each of the thirteen-training sets (each point). Shaded areas indicate the .25, .5, .75, and .95 quantiles. Maximum and minimum temperature (T_{max} and T_{min} , respectively), maturity group (MG), precipitation (Prec), vapor pressure deficit (VPD), number of days during seed fill with maximum temperatures > 30°C (sDays $T_{max} > 30$), soil organic matter (OM). Weather was summarized for the vegetative (v), reproductive pre-seed filling (r), and reproductive seed filling (s) periods. Crop growth stages were simulated with the Decision the Support System for Agrotechnology Transfer (DSSAT) model.

Table 4.2 Summary of season characteristics, weather, and soil variables for each cluster showing the median and the range (between brackets) of observations for each variable.

	Cluster			Factor
	North	Center	South	
vDays	40, [13, 71]	47, [12, 64]	38, [20, 60]	Season
rDays	25, [18, 32]	29, [21, 38]	33, [19, 42]	Season
sDays	36, [28, 43]	41, [26, 50]	39, [29, 59]	Season
MG	2, [0, 4]	3, [2, 5]	5, [2, 7]	Season
Clay (%)	17, [6, 35]	25, [3, 41]	27, [2, 78]	Soil
OM (%)	3.1, [1.6, 5.0]	2.8, [0.9, 5.3]	2.5, [0.8, 5.2]	Soil
pH	6.6, [5.6, 8.1]	6.3, [4.4, 8.2]	6.4, [4.9, 8.2]	Soil
P (ppm)	33, [5, 189]	25, [6, 211]	40, [8, 350]	Soil
vPrec (mm)^{1,2}	97, [56, 227]	143, [28, 298]	137, [28, 326]	Weather
vT_{min} (°C)	14.0, [11.0, 18.6]	15.5, [10.5, 21.0]	20.0, [12.9, 23.4]	Weather
vT_{max} (°C)	27.4, [25.2, 29.6]	28.4, [24.0, 33.3]	31.4, [25.3, 35.5]	Weather
vVPD (kPa)	0.8, [0.6, 1.1]	0.8, [0.4, 1.6]	1.0, [0.5, 1.9]	Weather
rPrec (mm)	60, [10, 131]	82, [3, 368]	80, [25, 266]	Weather
rT_{min} (°C)	16.1, [12.0, 21.5]	18.2, [14.6, 22.9]	21.8, [17.5, 23.7]	Weather
rT_{max} (°C)	28.5, [26.7, 34.6]	30.6, [26.8, 35.2]	33.8, [30.5, 35.6]	Weather
rVPD (kPa)	0.8, [0.6, 1.3]	1.0, [0.6, 1.8]	1.4, [0.8, 1.8]	Weather
sPrec (mm)	86, [9, 192]	100, [15, 190]	128, [33, 390]	Weather
sT_{min} (°C)	14.0, [10.3, 23.0]	15.8, [11.5, 23.5]	20.6, [13.0, 23.8]	Weather
sT_{max} (°C)	26.9, [22.9, 34.5]	28.8, [24.1, 34.5]	32.1, [28.6, 34.7]	Weather
sVPD (kPa)	0.8, [0.5, 1.1]	1.0, [0.6, 1.7]	1.0, [0.6, 1.8]	Weather
sDaysT_{max}>30	7, [2, 34]	15, [2, 42]	32, [14, 55]	Weather
sDaysT_{min}<15	26, [0, 38]	18, [0, 36]	2, [0, 26]	Weather

Number of days between VE-R1, R1-R5, and R5-R7 (vDays, rDays, and sDays, respectively), maturity group (MG), Precipitation (Prec), minimum and maximum temperature (T_{min} and T_{max}, respectively), vapor pressure deficit (VPD), number of days during seed fill with maximum temperatures > 30°C, (sDaysT_{max}>30), number of days during seed fill with minimum temperatures < 15°C, (sDaysT_{min}<15).

¹ Weather was summarized in vegetative (v), reproductive pre-seed filling (r), and reproductive seed filling (s).

² Growth stages were simulated with DSSAT.

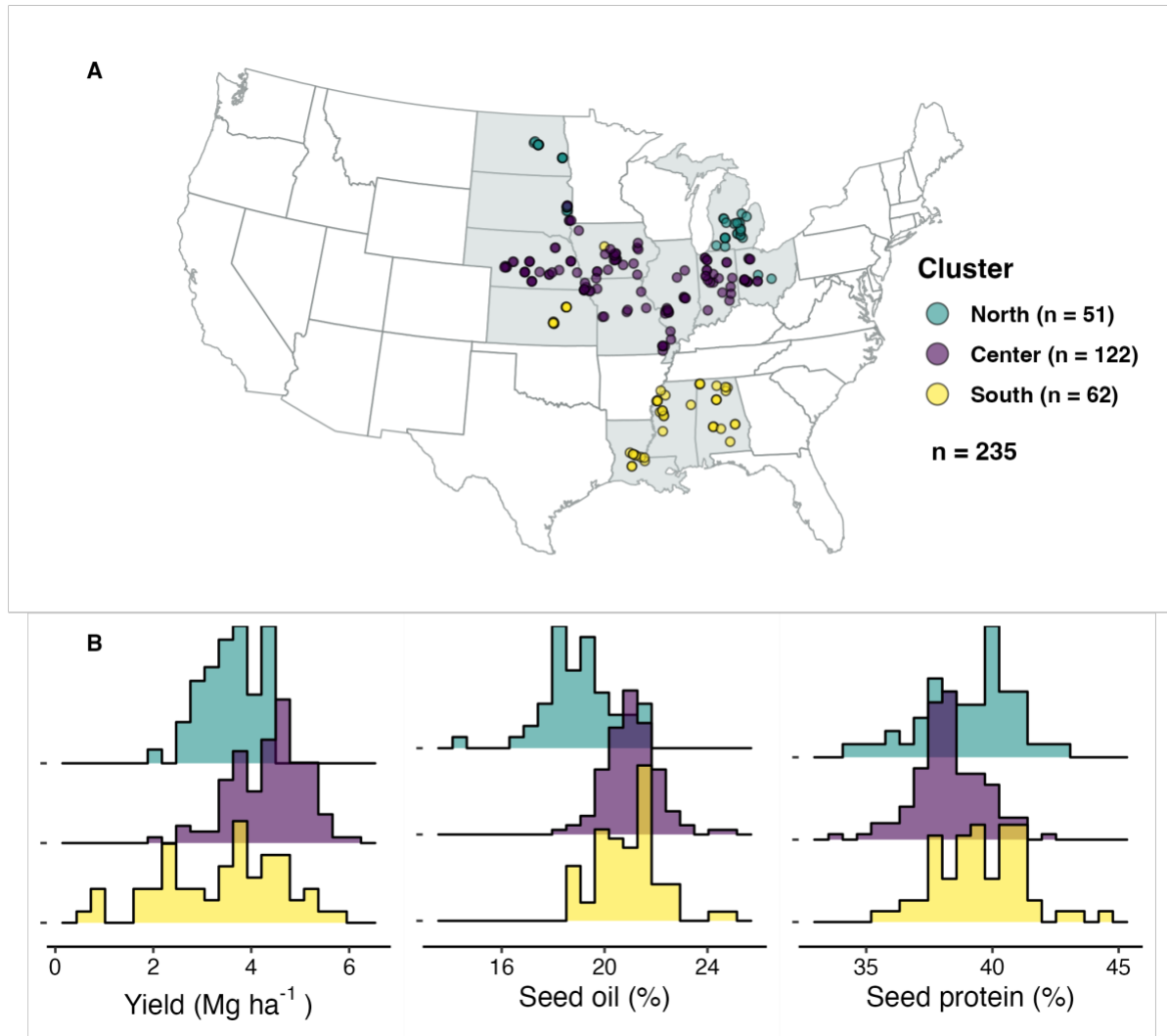


Figure 4.3 Location of the surveyed farmer fields (circles) colored according to the cluster classification, showing the total per cluster (A). Histogram displaying yield (Mg ha⁻¹), protein, and oil concentration (%) for each cluster (B).

Yield, oil and protein concentrations were significantly correlated with key variables influencing prediction accuracy (Fig. 2), with some of these variables exhibiting a strong link to regional differences (Fig. 4). High temperatures during early vegetative stages significantly limited yield, particularly in the southern region (Fig. 4A) where late planted soybeans (after June 1st) exacerbated a shortening of the total season length (Supplementary Fig. 4). The maturity group showed a regional pattern; with a positive yield response to longer crop maturity

(and longer seasons) in the northern region, whereas in the southern region, the tendency was for higher yields with shorter crop maturity (MG 4) (Fig. 4B). No discernible correlation was identified between yield and VPD during seed fill (Fig. 4C). Regarding oil, a positive correlation was observed with increasing minimum temperatures during the pre-seed-fill and seed-fill periods and a northern-to-southern trend. Higher minimum temperatures in pre-seed-fill were linked to an extended R1-R5 period, which was also positively associated with oil concentration (Fig. 4 D - F). Lastly, a positive association, albeit weak, was observed between protein concentration and precipitation in late reproductive stages (Fig. 4 G, H, I).

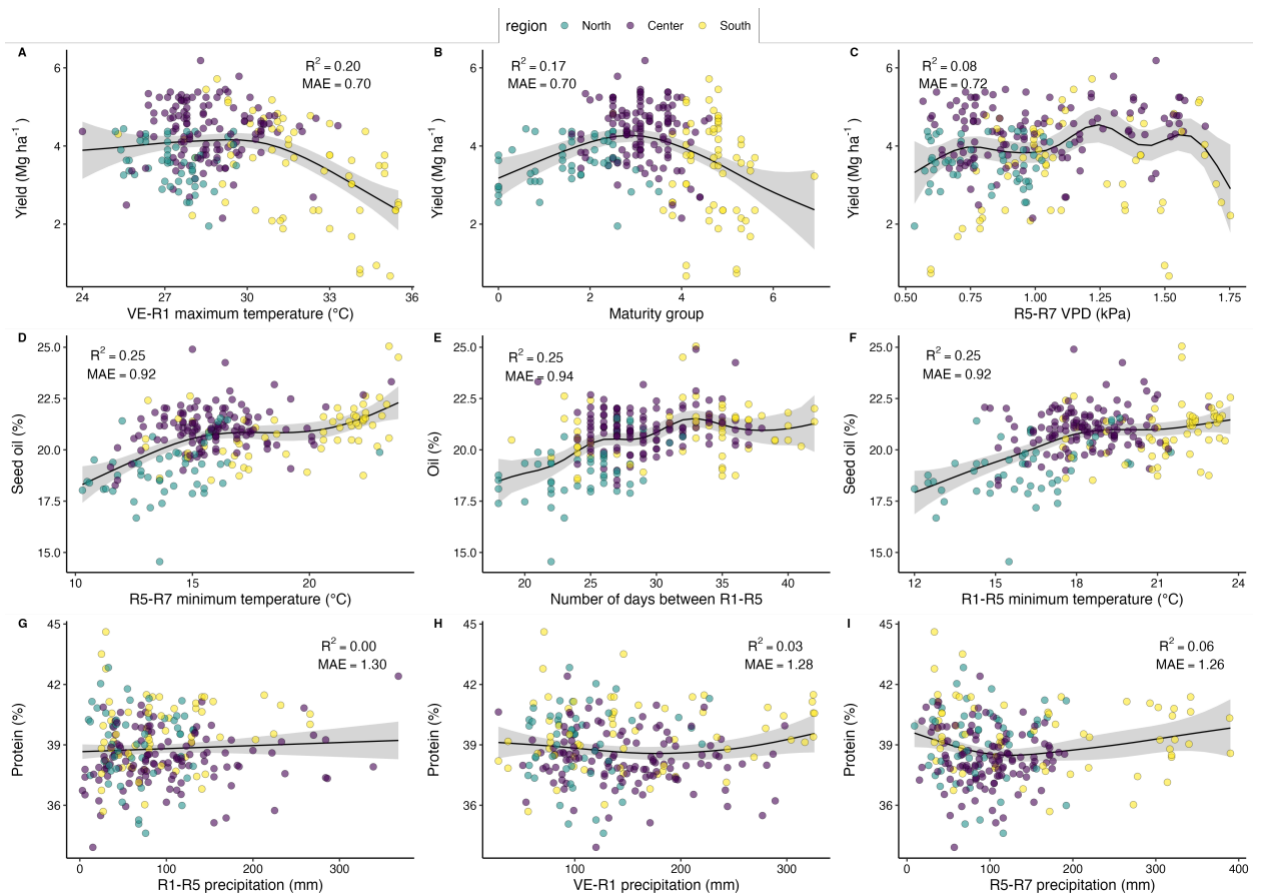


Figure 4.4. Soybean yield in relation to maximum temperature in VE-R1 (A), maturity group (B), and vapor pressure deficit (VPD, kPa) between R5-R7 (C). Additionally, soybean oil concentration in relation to minimum temperature between R5-R7 (D), number of days between

R1-R5 (E), and minimum temperature between R1-R5 (F). Lastly, protein concentration in relation to precipitation (mm) between R1-R5 (G), VE-R1 (H), and R5-R7 (I). The colors represent three clustered regions: North (green), Center (Corn Belt, purple), and South (yellow). Growth stages were simulated with DSSAT. The relationships were fitted using generalized additive models. The coefficient of determination (R^2) and mean absolute error (MAE) are shown.

4.4 Discussion

Our study encompasses a broad range of soil, weather, and crop management scenarios from 235 soybean fields spanning 13 US states, collectively accounting for over three-fourths of the country's production (85 million Mg) (USDA, NASS, 2023). The sampling protocol for seed and soil employed in this study ensured consistency in the characterization of each field through a representative sample ($n = 6\text{--}12$ within each farmer field). This approach represents a novel and unique method for capturing regional variations in seed quality parameters. Although our sample size may not capture every US soybean-producing region, this study is, to our knowledge, among the first to fuse field-derived data on yield, seed quality, and soil with weather information to identify the main factors driving variability in soybean crop. This is relevant to farmers and the soybean crushing industry, whose primary objectives are yield and protein concentration, respectively, as well as to the oil industry, where an unparalleled surge in demand is projected to result in a ninefold increase in renewable diesel production in the US to 22 billion L in 2024 (Egli, 2024; EIA, 2023).

Environmental variables (weather and soil) play a key role defining soybean yield and oil and protein concentrations, as evidenced by our study, which categorized different regions across the north, center (Corn Belt), and south of the US. Lower protein in the Corn Belt region have been widely documented although with high variability (Hurburgh, 1994; Rotundo et al., 2016). These protein levels have been attributed to higher overall yields (Assefa et al., 2018; Brumm

and Hurburgh, 2006), although high technification (irrigation and N management) is narrowing the yield/protein trade-off (Carciochi et al., 2023b). Conversely, the higher protein concentration observed in the northern regions compared to the Corn Belt is consistent with recent findings (Bullock et al., 2024; Cerrudo et al., 2023; Seth L. Naeve, 2023) showing a smaller difference in quality than usually reported in the past decades (Hurburgh, 1994; Hurburgh et al., 1990). This has significant implications for the price penalties that are typically applied (Bullock et al., 2024; William et al., 2020) to northern soybeans. In the southern region, elevated protein levels are commonly observed (Assefa et al., 2019; Rotundo et al., 2016), although with considerable heterogeneity and interannual variability. Regional variation is related to environmental variables that influence crop performance, including temperature, radiation, precipitation (Rotundo and Westgate, 2009b), and soil nutrients (De Borja Reis et al., 2021). Our study encompassed a broad area with diverse environmental gradients, enabling us to identify relationships that might not be discernible in smaller study regions.

The relationships observed between yield, oil, protein, and environmental descriptors were strongly influenced by regional gradients. High temperatures during the vegetative stages can accelerate development to reproductive stages, thereby reducing vegetative growth and shortening the overall season length, which directly impacts yield (Board and Modali, 2005). This effect is particularly evident in the south, where delayed planting dates have been shown to negatively affect soybean yield to a greater extent than in the central US region (Egli and Cornelius, 2009; Salmerón et al., 2016), likely related to higher temperatures. For oil concentration, low temperatures during reproductive stages often result from late planting dates and the selection of longer maturity groups, which shift the seed set and filling to environments with reduced solar radiation (Salmerón et al., 2022). Additionally, nearly 50% of vegetative

growth occurs between R1 and R5 (Egli and Leggett, 1973), further compromising seed filling for yield and oil synthesis. Consequently, selecting an appropriate crop maturity to ensure that reproductive growth aligns with optimal environmental conditions is crucial for yield and oil determination particularly in northern US regions (Egli and Cornelius, 2009; Salmerón et al., 2022). Protein, however, largely depends on N remobilization from vegetative biomass, N fixation, and mineral absorption during seed fill (Moro Rosso et al., 2023). Therefore, the environmental descriptors utilized in this study do not reflect the formation process for protein. In this regard, low precipitation or heat stress during reproductive stages can result in low yields and high protein (Rotundo and Westgate, 2009b). Conversely, high precipitation can prolong leaf area duration and photosynthesis, thereby sustaining N fixation and N remobilization (Kehoe et al., 2022) leading to an increase in protein levels (Carciochi et al., 2023b; Ergo et al., 2024b) as we observed in the southern region.

In summary, longer growing seasons with optimal planting dates were positively associated to increases in yield and oil concentration. The earlier planting of soybeans in the US can be hampered by erratic wet and/or cold springs as evidenced during the last decades (Deines et al., 2023). Hence, there is an opportunity to refine the planting date and maturity selection not only for yield (Mourtzinis et al., 2019; Siler and Singh, 2022), but also for oil productivity. Some limitations of our study are linked to the lack of i) in-season biomass data to understand physiological changes in crop growth and seed quality formation, ii) field-variability in seed quality and soil variables (instead of using the average), iii) temporal variation of soil N supply to investigate the impact on crop seed quality, and iv) more locally based weather stations. Further steps are required to account for the correlation between the environmental descriptors we employed with space such as in Rotundo et al., (2016), and also including topography and

within-field variability (Kravchenko and Bullock, 2002). Future studies might focus on understanding spatial variation within a farmer's field on these main soybean seed quality attributes. However, in order to do so, high-resolution spatial data will be required from both weather (at the local level) and soil parameters (Carroll et al., 2024). Finally, this work demonstrates the value of multi-institutional collaboration in gathering in-situ data with a consistent protocol and allowing for inference and comparison.

4.5 Conclusion

We used a robust protocol to collect representative seed, soil samples, and management data in 235 soybean farmer fields across 13 US states. We combined growth model simulations to summarize weather data during key crop phenological stages. The prediction of yield and oil concentration exhibited greater accuracy than that of protein concentration when seasonal variables related to weather, soil, and crop growth were considered. Yield, protein, and oil levels were within the ranges usually reported for soybean in the regions explored. However, higher protein levels in the north suggest a narrowing in the quality gap of soybeans between this region and the Corn Belt. The southern US yield might be limited by delayed planting dates that expose early stages to high temperatures and shorten the season length. In the northern regions, yield and oil are mainly limited by low temperatures during seed set and seed filling, which occur primarily with long maturity groups or late planting dates.

Chapter 5 - Soybean seed protein and oil concentration prediction in the US during 2022 and 2023

**Not submitted*

Valentina M. Pereyra*, Carlos Hernandez*, Trevor Hefley, Ajay Sharda, P.V. Vara Prasad, Laila Puntel, Guillermo R. Balboa, David Moseley, Jon Trenton Irby, Eros Francisco, André F. B. Reis, Aaron Prestholt, Joshua McDanel, Matthew Carroll, Shaun N. Casteel, Randall Pearson, Maninder Pal Singh, Peter Kovács, Michael H. Ostlie, John Fulton, Peter Kyveryga, and Ignacio A. Ciampitti*

Abstract

Soybean seed oil and protein concentration are important traits on driving the market demand for human consumption, biofuels, and animal feed. Yet, the strong influence of the environment (i.e., weather, soil), management (i.e., planting date), and genotype (i.e., maturity group) challenge the prediction and classification of seed quality across and within regions. In this study, we aimed at integrating satellite data and ground-truth seed samples to build a predictive model of soybean seed oil and protein concentration. We collected 1972 seed samples across 235 fields in 13 US states during 2022 and 2023 growing seasons. We modelled the green chlorophyll vegetation index (GCVI) derived from Sentinel-2 satellite, during the growing season and calculated the start of the season (emergence), the peak (around beginning of seed formation), and end of the growing season. We gathered soil data (organic matter, pH, and clay content), and summarized GCVI and weather (radiation, precipitation, minimum and maximum temperature) during three periods. We used these variables for training the XGBoost algorithm to predict seed protein and oil concentration. The final model for oil concentration explained 57%

of the variation (R^2) with a mean absolute error (MAE) of 0.76%, while the prediction of protein concentration explained 47% of the variation with a MAE of 1.10%. Lastly, we implemented the model in all fields of the US during the same growing seasons using the USDA soybean crop data layer to mask the predictors at a 30 m spatial resolution. This represents the first attempt on soybean at predicting seed quality over a large area and providing insights on weather, soil, and crop seasonal characteristics derived from satellite data (GCVI).

5.1 Introduction

The demand for soybeans has been predominantly driven by their use in animal feed, with a significant emphasis on enhancing protein concentration to achieve higher yields of soybean meal. Soybean oil, traditionally used for human consumption (Ritchie, 2021), is now receiving an increased attention due to recent policies promoting the use of biodiesel from vegetable oils in the transportation sector (Egli, 2024; EIA, 2023). The increasing demand for biofuel has intensified the debate surrounding the competing interests of food versus fuel, particularly as policies are implemented to phase out key crops such as palm oil (European Union, 2018). Accurately predicting seed quality within and across regions is therefore crucial to meet evolving market demands, diversify the oilseed production where possible, and leverage the competition for crop land use between food and fuel.

The quality of soybeans exhibits considerable variation both within and between regions, resulting from the complex interplay of genetic, environmental, and management factors (GxExM). During the seed filling phase, oil synthesis is predominantly influenced by weather conditions (e.g., temperature and drought stress), which directly impact current photosynthesis. In contrast, protein synthesis is more susceptible to the dynamics of nitrogen fixation and

remobilization from plant tissues. Although the impact of GxExM on soybean quality has been extensively investigated in controlled experiments and at the local scale, the challenge remains to translate this knowledge into large-scale regional predictions. In this context, satellite data present promising opportunities for leveraging the high spatial resolution in soil, weather, and crop growth through the use of vegetation indices (Hernandez et al., 2023; Sarkar et al., 2023).

The prediction of soybean quality using satellite data can provide crucial insights for inference, particularly if we can discern the ecophysiological processes that drive variation. For example, time series of vegetation indices, such as the green chlorophyll vegetation index (GCVI), have been demonstrated to predict protein and oil concentration during the peak at late reproductive stages (i.e., seed fill) (Hernandez et al., 2023; Sarkar et al., 2024, 2023). Time series of vegetation indices inform about crop phenology and vegetation dynamics thereby enhancing our understanding of seasonal weather and environmental variations (Gao and Zhang, 2021; Lou et al., 2023; Shen et al., 2022). Moreover, the incorporation of weather covariates has been demonstrated to enhance the prediction of planting dates for corn and soybeans relative to the use of vegetation indices alone (Deines et al., 2023). However, the foundation of a scalable model hinges on the acquisition of extensive ground-truth data from production-scale fields, capturing the spatial variability. Consequently, there is a necessity and an opportunity to capitalize on the high temporal and spatial resolution of satellite data to refine the prediction of soybean oil and protein concentration.

In this study, we developed a robust model to predict soybean oil and protein concentration using 1972 seed samples collected over 235 farmer fields in 13 US states during 2022 and 2023 growing seasons. We used GCVI time series obtained from Sentinel-2 to derive the peak and length of the growing season. We summarized weather and GCVI data across three

stages during the growing season and combined soil variables to optimize the XGBoost algorithm to predict soybean protein and oil concentration. Finally, we used the crop data layer (CDL) (Boryan et al., 2011) to extend the predictions of seed quality to all soybean fields across the US.

5.2 Material and methods

5.2.1 On-farm database on soybean yield, management practices, soil, and seed quality

Two years (2022 and 2023) of data were collected from 235 farmer fields across 13 U.S. states in collaboration with local universities: North Dakota, South Dakota, Michigan, Ohio, Iowa, Nebraska, Kansas, Missouri, Illinois, Indiana, Louisiana, Mississippi, and Alabama. We collected data on location, sowing date, harvest date, seeding rate, row spacing, and maturity group (MG) via online surveys. Field size ranged from 2.8 to 455 ha with a median field size of 24.3 ha. For sampling seeds, we selected between 6 to 12 sampling points distributed across the extension of each field and collected the plants within 54 cm of one row to form a representative seed sample per field. Seed protein and oil concentration (reported on a dry basis) were analyzed in whole seeds using a DA 7250TM near-infrared spectroscopy (NIRS) analyzer (Perten Instruments, Hägersten, Sweden). We collected a total of 1972 seed samples across 235 fields that formed our ground truth data for model training. We then used the model to predict quality in all soybean fields of the US during 2022 and 2023. We present an overview of the workflow implemented from data collection to model prediction in Figure 5.1.

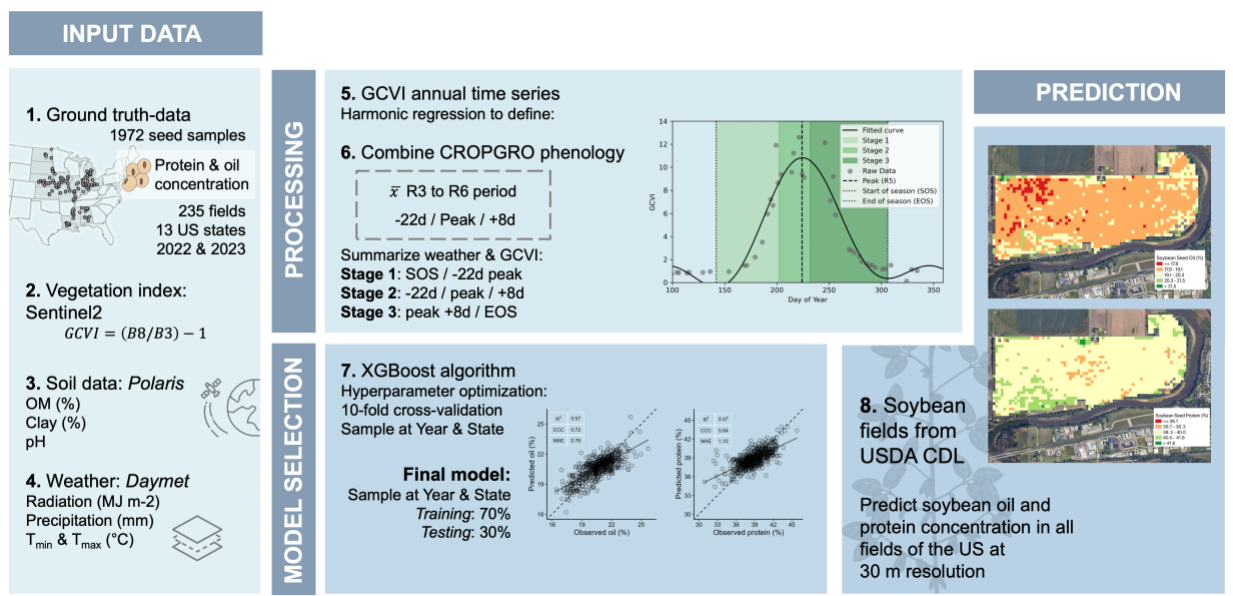


Figure 5.1. Workflow overview. Green Chlorophyll Vegetation Index (GCVI), soil organic matter (OM), minimum and maximum temperature ($T_{\min}$ & $T_{\max}$), beginning of pod (R3) and full seed (R6) phenological stages obtained from CROPGRO simulation model, coefficient of determination (R^2), concordance correlation coefficient (CCC), mean absolute error (MAE), crop data layer (CDL).

5.2.2 Satellite data

We downloaded soybean polygons from the USDA's Crop-land Data Layer (CDL) (Boryan et al., 2011) defined at 30 m resolution for 2022 and 2023. We used the CDL to mask all the raster data to the region of interest.

From Sentinel-2 satellite mission (ESA, 2021) we downloaded images with a resolution of 10 m for near infrared band (NIR, B8) and green (B3) with a 5-day revisit frequency. We filter cloudy observations using **XX**. We downloaded observations from April 1 until October 10 in 2022 and 2023, and calculated the Green Chlorophyll Vegetation Index (GCVI) (Gitelson et al., 2005, 2003) defined as:

$$GCVI = NIR/GREEN - 1 \quad \text{Equation 5.1}$$

We used GCVI to monitor crop greenness during the growing season because does not saturate as the Normalized Difference Vegetation Index (Viña et al., 2011) and characterizes crop

phenology and productivity in soybeans. We smoothed the time series observations for each pixel using a harmonic regression defined as:

$$f(t) = c + \sum_{k=1}^2 [a_k \cos(2\pi\omega kt) + b_k \sin(2\pi\omega kt)] \quad \text{Equation 5.2}$$

Where f is the response variable (GCVI) as a function of t , the day of the year, we set the number of harmonic terms k to 2, and ω to 2. Examples of a fitted harmonic curve can be found in Appendix fig D.1.

For each pixel, the fitted harmonic regression was employed to extract the day of the maximum GCVI (peak) and the minimum preceding and succeeding the peak. The peak of GCVI is closely correlated with the maximum leaf area index and the onset of seed formation (R5). The minimum and maximum values were employed to indicate the start (emergence) and conclusion (harvest time) of the season (SOS and EOS, respectively). The overall agreement between the estimated dates of emergence, peak (R5), and harvest time obtained with the Decision Support System for Agrotechnology Transfer (DSSAT) and those obtained with GCVI is illustrated in Appendix Fig. D.2. The mean interval between the R3 and R6 stages as estimated by DSSAT was 30 days, with a range of approximately 22 days prior to (SD = 12 days) and 8 days after (SD = 12 days) the peak determined by GCVI. Accordingly, this range was employed to delineate three stages throughout the season: between the SOS and 22 days prior to the peak (Stage 1), 22 days prior until 8 days after the peak (Stage 2, aimed at capturing the critical period for yield determination), and the remaining period until EOS (Stage 3).

We retrieved daily weather data on precipitation, maximum and minimum temperature, and radiation from Daymet (Thornton et al., 2022) with a 1 km x 1 km resolution using Google Earth Engine. For each pixel, we summarized the accumulated precipitation and radiation, and the median of temperature and GCVI during the Stage 1, 2, and 3. To retrieve soil data, we used Polaris to obtain a 30 m resolution probabilistic map on soil organic matter (%), clay (%), and

pH (Chaney et al., 2019) summarizing each at 30 cm depth. We calculated a total of 18 continuous predictors and summarized those for each state of data collection in Table 5.1.

5.2.3 Model Evaluation and prediction

We obtained all predictors of GCVI, weather, and soil at the sampling point locations (n=1972) in Google Earth Engine. We then worked in R (R Core Team, 2024) to implement and train the XGBoost algorithm (Chen and Guestrin, 2016) to predict soybean oil and protein concentration. For model training, we implemented a 10-fold cross-validation sampling at the state-year level and evaluated its performance on the held-out data. We optimized the hyperparameters: (i) nrounds, the number of boosted decision trees; (ii) eta, the learning rate or step size shrinkage; (iii) maxdepth, the maximum depth of a tree, where greater depth increases complexity; and (iv) subsample, the ratio of training instances used per tree. The minimum loss reduction required for further partitioning on a tree node (gamma) was fixed at 0, and the colsample_bytree parameter was set to 0.7 (70% of features randomly selected). All other parameters were left at their default values. We selected the optimal hyperparameter combination based on the overall model performance on the 10 held-out (testing) datasets. We used the concordance correlation coefficient (CCC) to identify the best combination, which reflects precision and accuracy. Finally, we calculated the feature importance on each training set during cross-validation as the average gain in prediction across all splits in which the feature was used.(Chen and Guestrin, 2016).

We used the best combination of hyperparameters to build the final model for predicting soybean oil and protein concentration. We split the ground truth data in training (70%) and testing (30%) sampling again at the State-Year level, and evaluated the prediction performance using CCC, mean absolute error (MAE), and coefficient of determination (R^2). We applied this

model in Google Earth Engine to all pixels classified as soybeans in the US based on the CDL for 2022 and 2023 growing seasons.

5.3 Results

We collected a total of 1972 seed samples over 235 farmer fields in 13 US states, with 41% and 59% collected during 2022 and 2023, respectively (Figure 5.2). The ranges observed for oil and protein were similar in 2022 and 2023 (Figure 5.2 B). Soybean oil concentration had an overall median of 20.7% and ranged between 19.8 and 23.5%, while protein concentration averaged 38.7% and ranged between 35.1 and 42.9% (95% quantile). A summary of protein, oil, GCVI, weather, and soil variables is presented in Table 5.1.

Table 5.1. Summary of season characteristics, soil, and management variables for each state showing the median and range (between brackets) of observations for each variable.

	State													
	LA	MS	AL	KS	MO	NE	IA	IN	IL	OH	MI	SD	ND	
Protein (%)	39.7 [35.1, 43.1]	38.6 [36.1, 43.8]	39.7 [33.5, 44.0]	39.7 [32.8, 46.7]	38.7 [33.9, 42.8]	38.7 [30.8, 43.3]	37.6 [30.9, 41.8]	38.6 [33.3, 43.8]	38.0 [34.9, 41.9]	38.4 [31.1, 42.3]	38.8 [31.1, 44.1]	39.1 [32.4, 42.4]	39.9 [34.7, 45.3]	
Oil (%)	21.8 [17.1, 27.0]	21.3 [18.6, 23.1]	21.0 [17.1, 23.5]	20.3 [16.5, 24.8]	20.8 [18.1, 23.5]	21.2 [17.0, 26.0]	21.0 [18.4, 23.8]	21.1 [18.0, 24.1]	21.5 [19.4, 23.6]	20.4 [16.0, 25.4]	19.0 [16.5, 22.6]	20.2 [17.5, 23.7]	18.7 [15.4, 23.1]	
SOS	122, [75, 171]	134 [101, 150]	137 [62, 189]	163 [127, 197]	138.5 [119, 154]	146 [126, 187]	145 [122, 174]	146 [63, 160]	141 [129, 160]	150 [115, 170]	152 [104, 188]	147.5 [51, 171]	158 [64, 215]	
Peak	204, [165, 238]	210 [193, 234]	219, [185, 285]	243 [210, 260]	218, [198, 230]	223 [209, 266]	223 [208, 246]	227 [192, 245]	220 [206, 237]	229 [204, 253]	229 [196, 272]	220 [164, 235]	229 [191, 300]	
EOS	288.5, [235, 345]	287 [266, 327]	303 [257, 358]	309 [282, 339]	297 [274, 340]	301 [286, 355]	301 [283, 327]	308 [271, 343]	298 [279, 316]	307 [284, 347]	303 [280, 350]	294 [233, 335]	298 [245, 342]	
GCVI S1	1.4 [0.4, 3.0]	2.1 [0.8, 6.5]	1.5 [0.7, 8.4]	1.6 [0.9, 7.9]	1.4 [0.2, 3.4]	1.2 [0.8, 4.1]	1.7 [1.1, 8.6]	1.6 [0.8, 5.1]	1.9 [0.9, 3.8]	1.0 [0.7, 3.4]	1.3 [0.6, 5.4]	1.6 [0.9, 6.2]	1.1 [0.5, 2.1]	
GCVI S2	5.1 [1.4, 13.7]	10.5 [1.8, 13.4]	6.6 [0.2, 14.2]	8.9 [1.4, 13.5]	9.8 [1.6, 12.9]	10.9 [1.7, 15.7]	9.7 [3.7, 17.2]	8.9 [0.1, 16.2]	11.9 [5.7, 14.5]	10.1 [0.9, 14.2]	10.0 [0.9, 14.2]	10.7 [2.6, 13.7]	7.0 [0.9, 13.3]	
GCVI S3	1.5 [0.4, 6.0]	2.0 [1.2, 5.3]	2.0 [1.2, 6.3]	2.2 [1.2, 5.2]	2.1 [1.1, 5.9]	2.1 [1.1, 4.5]	2.2 [1.0, 6.3]	1.6 [0.9, 3.8]	2.1 [1.1, 3.1]	1.7 [0.8, 9.8]	2.0 [0.9, 9.8]	2.3 [1.5, 4.7]	2.5 [0.8, 7.5]	
Prec S1 (mm)	182 [67, 437]	173 [129, 321]	287 [119, 619]	183 [36, 305]	100 [38, 291]	165 [64, 307]	164 [76, 264]	181 [77, 452]	166 [104, 305]	195 [135, 330]	181 [42, 414]	120 [80, 271]	130 [38, 273]	
Prec S2 (mm)	63 [0, 275]	92 [21, 434]	105 [13, 252]	38 [8, 129]	156 [59, 438]	64 [5, 167]	75 [4, 173]	97 [43, 259]	145 [66, 202]	102 [38, 136]	98 [36, 234]	101 [48, 141]	46 [0, 91]	
Prec S3 (mm)	237 [51, 528]	123 [49, 481]	166 [47, 417]	88 [33, 152]	130 [42, 280]	100 [31, 234]	126 [51, 214]	151 [86, 250]	148 [97, 218]	109 [49, 181]	174 [92, 238]	69 [37, 259]	67 [6, 191]	
Radn S1 (MJ m ⁻²)	1116 [741, 1726]	1109 [942, 1563]	1158 [820, 2173]	1103 [648, 1332]	1228 [1013, 1470]	1183 [982, 1531]	1170 [997, 1580]	1174 [964, 2662]	1176 [1024, 1329]	1192 [993, 1914]	1212 [772, 2284]	1129 [757, 2981]	1027 [722, 2643]	
Radn S2 (MJ m ⁻²)	575 [461, 622]	552 [434, 580]	557 [428, 607]	561 [475, 609]	516 [483, 584]	583 [470, 624]	578 [495, 644]	546 [517, 611]	547 [517, 577]	554 [485, 639]	550 [385, 599]	578 [534, 659]	547 [126, 663]	
Radn S3 (MJ m ⁻²)	1179 [832, 1959]	1150 [865, 1483]	1124 [416, 1656]	819 [570, 1251]	1051 [900, 1444]	987 [439, 1228]	966 [626, 1197]	946 [784, 1439]	1042 [811, 1177]	920 [718, 1197]	787 [331, 1482]	917 [668, 1577]	741 [314, 1353]	
T _{max} S1 (°C)	32.2 [27.5, 35.3]	32.0 [29.3, 33.3]	30.6 [25.9, 33.9]	32.9 [29.3, 36.0]	30.0 [27.8, 31.4]	28.5 [27.0, 32.3]	28.4 [26.7, 30.5]	28.0 [21.9, 30.5]	29.2 [28.7, 31.6]	27.5 [25.0, 29.6]	27.1 [22.6, 28.8]	27.9 [14.1, 30.3]	27.3 [15.5, 28.8]	
T _{max} S2 (°C)	35.5 [32.5, 38.5]	33.8 [31.9, 35.2]	32.8 [28.0, 35.9]	33.0 [30.2, 35.9]	31.0 [28.7, 32.3]	30.8 [27.7, 32.4]	29.1 [27.3, 31.5]	28.0 [26.6, 30.2]	29.8 [28.5, 31.4]	28.0 [26.7, 29.4]	27.3 [21.9, 28.8]	28.1 [25.9, 30.6]	27.0 [14.6, 28.2]	
T _{max} S3 (°C)	32.0 [25.9, 36.7]	31.6 [25.9, 34.3]	27.8 [20.5, 34.7]	24.6 [18.7, 31.7]	27.4 [23.4, 30.2]	25.7 [16.8, 28.4]	23.9 [16.9, 27.2]	22.0 [16.9, 27.4]	24.8 [22.4, 28.9]	21.7 [16.0, 26.4]	20.0 [11.3, 26.8]	23.2 [15.6, 29.1]	20.3 [7.9, 27.4]	
T _{min} S1 (°C)	20.0 [14.6, 24.4]	20.4 [16.8, 22.1]	18.8 [13.2, 21.8]	20.0 [15.9, 20.8]	16.5 [14.4, 18.2]	15.1 [13.5, 18.6]	15.4 [13.5, 17.8]	15.0 [9.2, 17.9]	16.8 [15.2, 19.8]	15.6 [12.6, 17.8]	13.9 [10.3, 17.0]	15.2 [1.7, 17.5]	14.7 [2.7, 15.9]	
T _{min} S2 (°C)	23.4 [21.3, 24.8]	23.1 [22.3, 23.7]	21.8 [12.8, 22.7]	17.4 [15.6, 21.0]	20.2 [18.1, 21.5]	17.2 [10.3, 19.1]	16.9 [15.3, 18.8]	16.6 [14.8, 20.0]	19.4 [17.0, 20.6]	16.8 [15.1, 18.4]	15.3 [10.9, 17.3]	15.1 [13.5, 17.3]	14.1 [1.3, 15.5]	
T _{min} S3 (°C)	20.4 [12.1, 23.4]	18.6 [11.9, 21.0]	14.9 [6.6, 22.0]	8.9 [3.6, 15.9]	14.3 [9.7, 18.1]	9.4 [0.4, 13.1]	9.9 [3.1, 13.4]	9.2 [4.3, 15.8]	11.7 [9.2, 16.5]	9.8 [4.0, 13.7]	8.8 [1.8, 15.1]	9.1 [1.9, 16.6]	6.2 [-2.4, 13.7]	
Clay (%)	17.5 [12.2, 60.4]	48.4 [11.5, 54.1]	16.9 [4.7, 52.6]	30.8 [5.0, 37.2]	23.9 [8.8, 45.1]	24.5 [5.7, 41.6]	28.9 [17.4, 40.6]	18.7 [3.2, 35.1]	22.5 [13.3, 42.6]	24.6 [10.5, 37.0]	12.3 [3.4, 44.8]	25.7 [18.2, 30.0]	19.3 [15.5, 54.8]	
OM (%)	1.0 [0.9, 1.2]	1.1 [0.8, 1.4]	0.9 [0.8, 1.4]	1.3 [0.6, 1.7]	1.1 [0.8, 1.6]	1.2 [0.7, 1.6]	1.5 [1.0, 2.3]	1.2 [0.7, 2.0]	1.5 [0.9, 2.0]	1.2 [1.0, 1.7]	1.2 [0.7, 4.8]	1.6 [1.3, 1.9]	1.6 [1.4, 1.9]	
pH	6.5 [5.1, 7]	5.2 [5.1, 6]	5.2 [4.7, 7.2]	5.8 [5.7, 7.6]	6.1 [5.3, 7.3]	6.3 [5.6, 7.2]	6.2 [5.6, 7.8]	6.1 [5.4, 7]	6.2 [5.4, 6.6]	6.2 [5.4, 6.5]	6.2 [5.4, 7.5]	6.7 [6.4, 7.7]	7.3 [7.1, 7.9]	

We obtained the time series of the green chlorophyll vegetation index (GCVI) during the growing season and calculated the day of the year for start of season, Peak and end of season (SOS, Peak and EOS, respectively). Then we divided the growing season into three stages (S1, S2, and S3) between emergence and end of season. Precipitation (Prec), maximum and minimum temperature (T_{max} and T_{min}, respectively).

We obtained the time series of the green chlorophyll vegetation index (GCVI) during the growing season and calculated the day of the year for start of season (SOS, Peak, and EOS, respectively). We then divided the growing season it in three stages (S1, S2, and S3) between emergence and end of season. Precipitation (Prec), radiation (Radn), maximum and minimum temperature (T_{max} and T_{min}, respectively).

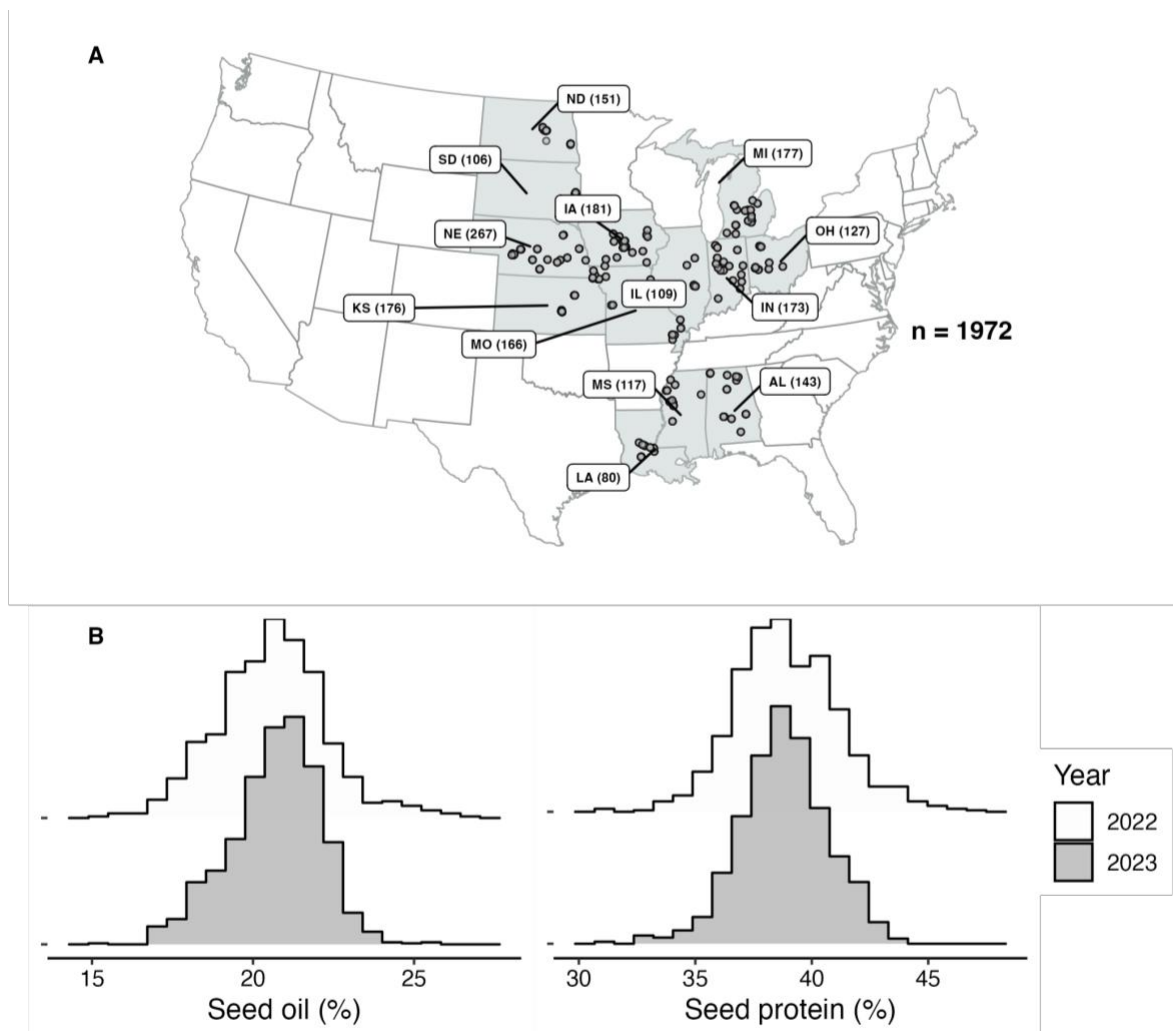


Figure 5.2. Location of the farmer fields (circles) in the US during the 2022 and 2023 growing seasons, showing the total per state (A). Histogram displaying seed oil and protein concentrations (%) in both 2022 and 2023 growing seasons (B).

The final model predicting soybean oil concentration explained 57% of the variation (R^2) with a MAE of 0.76%, while the prediction of protein concentration explained 47% of the variation with a MAE of 1.10% (Figure 5.3 A and B). The performance of the best combination of hyperparameters across the held-out datasets during model training is shown in Appendix Fig. D.4. The three most significant variables for predicting oil were identified as the minimum temperature during the critical period surrounding the peak of GCVI (stage 2), radiation in the

late season (stage 3), and maximum temperature during stage 2 (Figure 5.1-C). In contrast, for protein, the GCVI during stage 1 and 2, and precipitation in stage 3 were determined to be the most influential predictors (Figure 5.1-D).

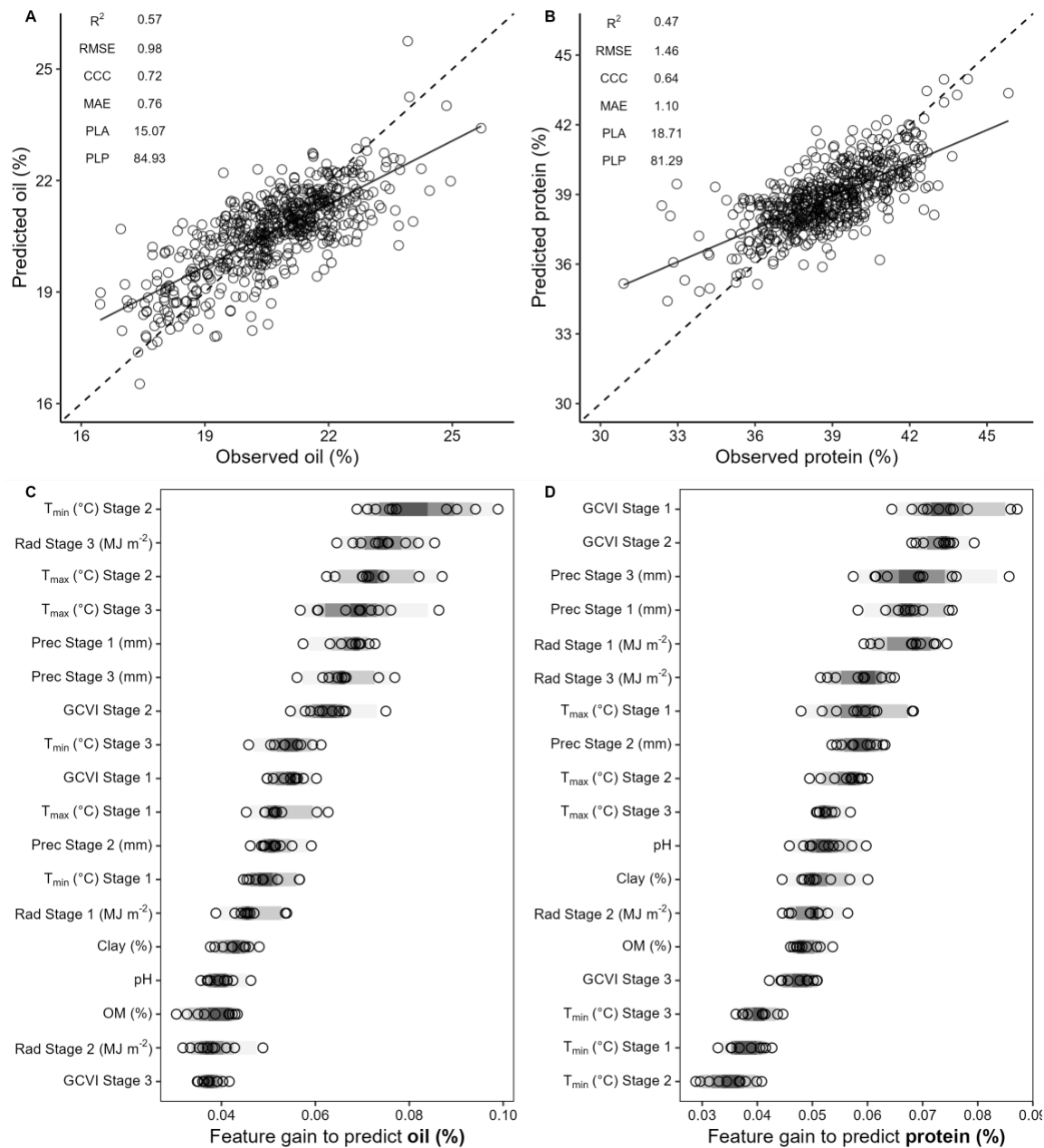


Figure 5.3. Comparison of observed versus predicted oil (%), A) and protein concentration (%), B) of soybean on the testing dataset (30%) of the final model. The dots represent observed values from a farmer field, the dashed line shows the 1:1 relationship, and the solid line denotes the overall relationship between the observed and predicted data. Coefficient of determination (R^2), root mean square error (RMSE), concordance correlation coefficient (CCC), mean absolute error

(MAE), percentage lack of accuracy (PLA), and percentage lack of precision (PLP) are shown. Variable importance (feature gain) of XGBoost model for oil (C) and protein (D) concentration for each of the 10 training datasets from the cross-validation with the best combination of hyperparameters. Shaded areas indicate .25, .50, .75, and .95 quantiles.

As an example, we show the prediction of protein and oil in soybean fields close to Manhattan, Kansas during 2023 in Figure 5.4.

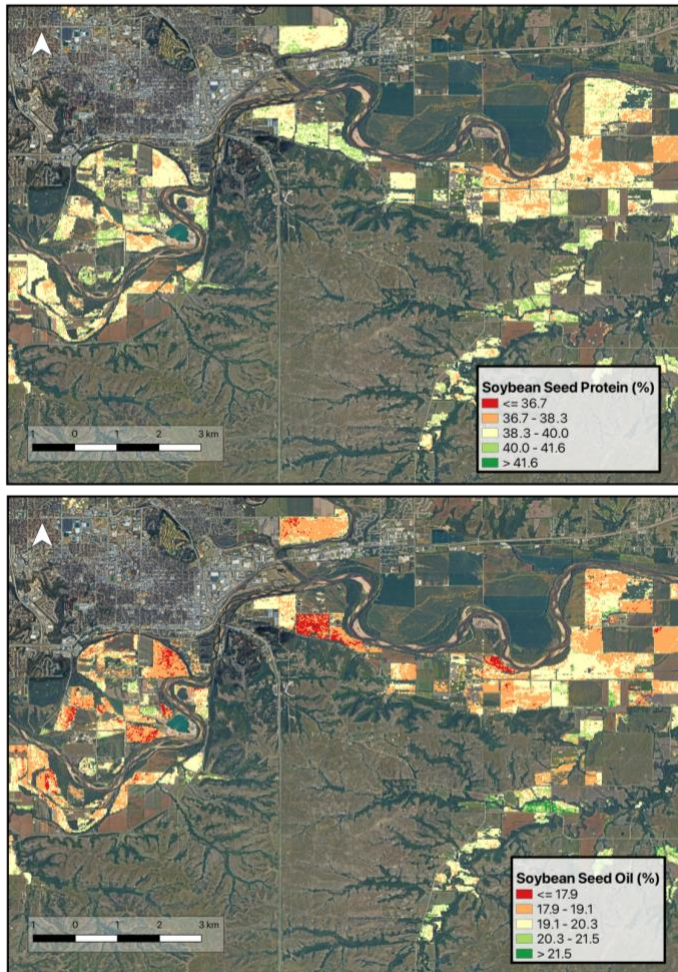


Figure 5.4. Example map showing the prediction of seed oil and protein in soybean fields near Manhattan, KS during 2023.

5.4 Discussion

These results mark a significant advancement in predicting soybean seed quality with high accuracy and inference into the factors driving variability across large regions at fine spatial

scales. To the best of our knowledge, this is the first study to integrate ground-truth data with satellite information on weather, soil, and crop growth characterization (i.e., GCVI) to predict soybean seed quality across a large region. This approach markedly enhances spatial resolution and prediction accuracy in comparison to previous endeavors employing interpolation of observations or the utilization of vegetation indices in isolation (Hernandez et al., 2023; Rotundo et al., 2016; Sarkar et al., 2024, 2023). The observed variability in oil content is similar to other studies showing a 17.6% of variability around the mean (Assefa et al., 2019; Cerrudo et al., 2023; Rotundo et al., 2016). However, we explored a wider range of protein variability than is typically reported, with a mean value of 20.8% and a standard deviation of 17.8%. This likely reflects a broader range of environmental conditions during the growing season. These insights can assist farmers and merchandisers in optimizing production and supply chains to meet diverse quality demands. Moreover, the ability to predict oil concentration can directly influence policy decisions about energy production. As the demand for biofuels continues to grow, soybeans yields are increasing globally, with no evidence of stagnation in 85% of global areas (Gerber et al., 2024; Ray et al., 2012). However, challenges remain regarding the expansion and conversion of land for soybean production (Heimann et al., 2023; Marin et al., 2022; Ritchie, 2021).

In general, oil concentration is primarily influenced by temperature and the duration (accumulated radiation) of the early reproductive and seed fill period (Rotundo and Westgate, 2009a), while protein concentration is more related to N remobilization from plant tissues and N fixation during the seed fill period. Vegetation indices, such as GCVI, can be employed to monitor the phenology and leaf area of soybeans, thereby reflecting crop growth dynamics across the growing season (Lou et al., 2023). Consequently, GCVI may serve to indicate the nitrogen

status of the entire crop and assist in the forecasting of seed protein concentration in soybeans (Hernandez et al., 2023; Sarkar et al., 2023). The integration of weather data with vegetation indices has consistently demonstrated superior performance in prediction frameworks, exceeding the accuracy of models that rely on a single data source (Cai et al., 2019; Deines et al., 2023; Li et al., 2023). Lastly, discrepancies in seed quality between full-season and double-cropped soybeans are frequently observed (Hernandez et al., 2023), underscoring the necessity to refine this framework to connect management strategies directed at specific seed quality levels.

A few limitations of our study are linked to the use of harmonic regression to extract phenology features, which can result in an under- or overestimation of the beginning and end of the growing season (Beck et al., 2006), and assume a symmetric relationship during the growing and senescence stages. Future efforts should focus on i) increasing the ground truth data size including more observations and years to link temporal variability on protein and oil concentration, ii) using of historical weather data to predict average regional quality, iii) combining yield and seed quality predictions to better understand the trade-off between these two attributes and extrapolate estimates of oil and protein production to county, state, and national statistics. Similar approaches applied to other oilseed species could help to better assess the production and diversification of species and end-uses. Coordinated efforts between research and policymakers is needed (Mehrabi et al., 2022), especially to adapt policies that aim for sustainability and diversification of the global oilseed production system (Cadillo-Benalcazar et al., 2021; Többen et al., 2024).

5.5 Conclusion

We integrated a comprehensive ground truth dataset of on-farm soybean seed quality with satellite data on weather, soil, and GCVI to forecast soybean seed protein concentration across the United States during 2022 and 2023. The resulting predictions exhibited a mean absolute error (MAE) of 0.76% for oil and 1.10% for protein, representing an improvement over previous effort. This work illustrates the potential of collaborative research initiatives that integrate observed data across extensive regions. Future endeavors should extend these approaches to yield predictions, enabling estimates of seed oil and protein production across diverse geographical regions. Connecting research and policy efforts will be pivotal to address sustainability concerns and address shortcomings in both hard and soft infrastructure for fostering consensus.

Chapter 6 - Conclusions

Soybean plant arrangement, due to the seeding rate and row spacing combination, lays the initial configuration for plant growth and development. Appropriate combinations depend on the genotype and the environments. Continuous advances in high yielding cultivars and machinery necessitate ongoing reevaluation of basic management strategies affecting growth, yield, and seed quality. Additionally, understanding the factors that influence soybean protein and oil concentration is crucial for addressing global market prospects with informed and science-based decision-making. Our research expands the understanding of soybean growth plasticity under different seeding rates and row spacing in modern systems. Moreover, our results set the foundation for exploring seed quality variations within and across regions in soybeans.

The first objective of this dissertation was to evaluate the effect of plant-to-plant spatial uniformity on soybean yield across various seeding rate and yield environments. In chapter 2, we found that non-uniform plant spacing affected yield similar across seeding rates from 100 to 550 seeds m^{-2} . Contrary to our hypothesis that plants at lower seeding rates have less compensation under non-uniform spacing, we found that yield compensation was limited at all seeding rates. However, non-uniform spacing impacted yield in low- and medium-yield environments ($< 3 \text{ Mg ha}^{-1}$), but not in higher yield environments. Moreover, we developed a novel metric that correlates observed plant spacing with the theoretical spacing (calculated from seeding rate and row spacing) and demonstrated that this metric has a stronger correlation with yield than the standard deviation of spacing used in prior research. These results are valuable for managing soybeans in environments that can limit growth and compensation for yield (i.e., express phenotypic plasticity).

The second objective of this dissertation was to assess the impact of different row spacing and their effects on yield and seed quality, at plant level and also on the overall variation in main stems and branches for both an old and a modern soybean genotype. In chapter 3, we found that the yield improvement of the modern variety was greater under narrow rows (<0.76 m) relative to the already documented gain under wider rows. As anticipated, the yield improvement was predominantly attributable to an increased number of pods and yield on branches. In addition, seed protein concentration in the modern genotype was maintained, despite the high yields. Lastly, we found that in low yielding conditions, protein concentration on branches was lower than on the main stem, suggesting that photo-assimilate dominance on the main stem limit protein deposition on branches. These findings reinforce the benefits of narrow row spacing for improving yield in modern soybean, particularly in high-yielding conditions, without sacrificing protein concentration.

The third objective was to evaluate the role of weather, soil, and management factors in predicting soybean oil and protein concentration, and to identify regions associated with distinct seed quality levels. In chapter 4, we collected in situ seed and soil samples from 235 soybean fields across 13 US states. We averaged the seed quality of 6 to 12 samples per field allowing for a representative estimation of the mean. A model of phenology was constructed, and weather variables were summarized during key stages of soybean growth. The results demonstrated that weather, soil, and management variables exhibited a greater degree of accuracy in predicting yield and oil content than protein content. Three main regions were identified -north, center (Corn Belt) and south, within which seed quality ranges fell within the expected ranges except for higher protein in the north than in the Corn Belt region suggesting a narrowing of the quality gap. The yield and oil content in the southern region were constrained by delayed planting dates,

which exposed the early stages to elevated temperatures and shortened the growing season. In contrast, the northern regions were limited by low temperatures during seed formation, due to the long maturity groups or late planting dates. These findings offer important benchmarks for soybean seed quality across regions and enhance our understanding of how management and environmental factors influence regional variations.

The fourth objective of this dissertation was to build a robust predictive model for soybean oil and protein concentration across US soybean fields. In chapter 5, we integrated satellite data on weather, soil, along with the green chlorophyll vegetation index (GCVI) and derived the growing season length at a 1 x 1 km resolution. We used the same data from chapter 4 although each seed sample collected was used as an observation allowing for a $n=1973$. Using the XGBoost algorithm, we predicted oil with an R^2 of 0.54 and a MAE of 0.75%, while protein predictions had a R^2 of 0.45 and a MAE = 1.13%. Seed quality predictions across all US soybean fields showed higher protein concentration in northern and southern regions with lower levels on the Corn Belt. While oil was higher in the Corn Belt and southern region. This effort substantially improved the spatial resolution and accuracy of prediction compared to existing remote sensing and interpolation efforts to map seed quality. Finally, this work demonstrates the potential of multi-institutional collaboration by bringing multiple farmers across all of the main producing region of the US.

Limitations and Future Research

The main limitations on the finding of our studies can be attributed to the lack of variation in planting dates, locations, and genotypes for Chapters 2 and 3. For instance, different planting dates could have introduced substantial variation in environmental conditions affecting growth and development likely expanding possible recommendations on seeding rate and row

spacing choice. Similarly, in Chapter 3, only two genotypes were tested over two contrasting growing seasons, which constrains the inference of our results. Nevertheless, the detailed measurements collected throughout the season allowed for meaningful interpretations. In Chapters 4 and 5, limitations were related to the yield data collected via survey rather than in-situ measurements. In addition, in-season biomass collection could have provided physiological perspectives of yield and seed quality formation.

Future research efforts on soybean should focus on yield and seed quality improvement in the context of climate change, the development of tailored varieties to meet current market demands, and an investigation into the capacity of legumes to fix N₂. Recent studies have demonstrated that elevated CO₂ can increase yield and protein concentration, though oil levels may decline (Digrado et al., 2024). However, high nodulation and balanced carbon allocation may offset this effect maintaining oil levels while increasing yield and protein concentration (Zhong et al., 2024). Additionally, the imperative need to reduce greenhouse gas emissions to revert catastrophic events due to climate change tackled growing interest in renewable diesel from vegetable oils. In the near future, developing high oil-low protein, high-yielding cultivars may become a primary research objective to enhance oil production per unit area (Egli, 2024).

References

- Alboukadel Kassambara, 2023. ggpubr: “ggplot2” Based Publication Ready Plots.
- Andrade, F.H., Abbate, P.E., 2005. Response of Maize and Soybean to Variability in Stand Uniformity. *Agron. J.* 97, 1263–1269. <https://doi.org/10.2134/agronj2005.0006>
- Andrade, J.F., Rattalino Edreira, J.I., Mourtzinis, S., Conley, S.P., Ciampitti, I.A., Dunphy, J.E., Gaska, J.M., Glewen, K., Holshouser, D.L., Kandel, H.J., Kyveryga, P., Lee, C.D., Licht, M.A., Lindsey, L.E., McClure, M.A., Naeve, S., Nafziger, E.D., Orlowski, J.M., Ross, J., Staton, M.J., Thompson, L., Specht, J.E., Grassini, P., 2019. Assessing the influence of row spacing on soybean yield using experimental and producer survey data. *Field Crops Res.* 230, 98–106. <https://doi.org/10.1016/j.fcr.2018.10.014>
- Assefa, Y., Bajjalieh, N., Archontoulis, S., Casteel, S., Davidson, D., Kovács, P., Naeve, S., Ciampitti, I.A., 2018. Spatial Characterization of Soybean Yield and Quality (Amino Acids, Oil, and Protein) for United States. *Sci. Rep.* 8, 14653. <https://doi.org/10.1038/s41598-018-32895-0>
- Assefa, Y., Purcell, L.C., Salmeron, M., Naeve, S., Casteel, S.N., Kovács, P., Archontoulis, S., Licht, M., Below, F., Kandel, H., Lindsey, L.E., Gaska, J., Conley, S., Shapiro, C., Orlowski, J.M., Golden, B.R., Kaur, G., Singh, M., Thelen, K., Laurenz, R., Davidson, D., Ciampitti, I.A., 2019. Assessing Variation in US Soybean Seed Composition (Protein and Oil). *Front. Plant Sci.* 10, 298. <https://doi.org/10.3389/fpls.2019.00298>
- Beck, P.S.A., Atzberger, C., Høgda, K.A., Johansen, B., Skidmore, A.K., 2006. Improved monitoring of vegetation dynamics at very high latitudes: A new method using MODIS NDVI. *Remote Sens. Environ.* 100, 321–334. <https://doi.org/10.1016/j.rse.2005.10.021>
- Bellaloui, N., Gillen, A.M., 2010. Soybean seed protein, oil, fatty acids, N, and S partitioning as affected by node position and cultivar differences. *Agric. Sci.* 01, 110–118. <https://doi.org/10.4236/as.2010.13014>
- Bellaloui, N., Smith, J.R., Gillen, A.M., Fisher, D.K., Mengistu, A., 2012. Effect of Shade on Seed Protein, Oil, Fatty Acids, and Minerals in Soybean lines Varying in Seed Germinability in the Early Soybean Production System. *Am. J. Plant Sci.* 03, 84–95. <https://doi.org/10.4236/ajps.2012.31008>
- Board, J.E., Harville, B.G., Saxton, A.M., 1990. Branch Dry Weight in Relation to Yield Increases in Narrow-Row Soybean. *Agron. J.* 82, 540–544. <https://doi.org/10.2134/agronj1990.00021962008200030021x>
- Board, J.E., Modali, H., 2005. Dry Matter Accumulation Predictors for Optimal Yield in Soybean. *Crop Sci.* 45, 1790–1799. <https://doi.org/10.2135/cropsci2004.0602>
- Board, J.E., Settimi, J.R., 1986a. Photoperiod Effect before and after Flowering on Branch Development in Determinate Soybean ¹. *Agron. J.* 78, 995–1002. <https://doi.org/10.2134/agronj1986.00021962007800060013x>
- Board, J.E., Settimi, J.R., 1986b. Photoperiod Effect before and after Flowering on Branch Development in Determinate Soybean ¹. *Agron. J.* 78, 995–1002. <https://doi.org/10.2134/agronj1986.00021962007800060013x>
- Borja Reis, A.F., Rosso, L., Davidson, D., Kovács, P., Purcell, L.C., Below, F.E., Casteel, S., Kandel, H.J., Naeve, S., Archontoulis, S.V., Ciampitti, I.A., 2022a. Soybean management for seed composition: The perspective of U.S. farmers. *Agron. J.* 114, 2608–2617. <https://doi.org/10.1002/agj2.21082>

- Borja Reis, A.F., Rosso, L., Davidson, D., Kovács, P., Purcell, L.C., Below, F.E., Casteel, S., Kandel, H.J., Naeve, S., Archontoulis, S.V., Ciampitti, I.A., 2022b. Soybean management for seed composition: The perspective of U.S. farmers. *Agron. J.* 114, 2608–2617. <https://doi.org/10.1002/agj2.21082>
- Boryan, C., Yang, Z., Mueller, R., Craig, M., 2011. Monitoring US agriculture: the US Department of Agriculture, National Agricultural Statistics Service, Cropland Data Layer Program. *Geocarto Int.* 26, 341–358. <https://doi.org/10.1080/10106049.2011.562309>
- Brumm, T.J., Hurburgh, C.R., 2006. Changes in long-term soybean compositional patterns. *J. Am. Oil Chem. Soc.* 83, 981–983. <https://doi.org/10.1007/s11746-006-5056-4>
- Bullock, D.W., Wilson, W.W., Thompson, R., 2024. Nongrade quality factor influence on origin soybean basis: Evidence from North Dakota. *J. Agric. Appl. Econ. Assoc.* 3, 228–241. <https://doi.org/10.1002/jaa2.110>
- Cadillo-Benalcazar, J.J., Bukkens, S.G.F., Ripa, M., Giampietro, M., 2021. Why does the European Union produce biofuels? Examining consistency and plausibility in prevailing narratives with quantitative storytelling. *Energy Res. Soc. Sci.* 71, 101810. <https://doi.org/10.1016/j.erss.2020.101810>
- Cafaro La Menza, N., Monzon, J.P., Lindquist, J.L., Arkebauer, T.J., Knops, J.M.H., Unkovich, M., Specht, J.E., Grassini, P., 2020. Insufficient nitrogen supply from symbiotic fixation reduces seasonal crop growth and nitrogen mobilization to seed in highly productive soybean crops. *Plant Cell Environ.* 43, 1958–1972. <https://doi.org/10.1111/pce.13804>
- Cafaro La Menza, N., Monzon, J.P., Specht, J.E., Grassini, P., 2017. Is soybean yield limited by nitrogen supply? *Field Crops Res.* 213, 204–212. <https://doi.org/10.1016/j.fcr.2017.08.009>
- Cai, W., Chen, S., Zhang, D., 2007. Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation. *Pattern Recognit.* 40, 825–838. <https://doi.org/10.1016/j.patcog.2006.07.011>
- Cai, Y., Guan, K., Lobell, D., Potgieter, A.B., Wang, S., Peng, J., Xu, T., Asseng, S., Zhang, Y., You, L., Peng, B., 2019. Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. *Agric. For. Meteorol.* 274, 144–159. <https://doi.org/10.1016/j.agrformet.2019.03.010>
- Carciochi, W.D., Grassini, P., Naeve, S., Specht, J.E., Mamo, M., Seymour, R., Nygren, A., Mueller, N., Sivits, S., Proctor, C., Rees, J., Whitney, T., La Menza, N.C., 2023a. Irrigation increases on-farm soybean yields in water-limited environments without a trade-off in seed protein concentration. *Field Crops Res.* 304, 109163. <https://doi.org/10.1016/j.fcr.2023.109163>
- Carciochi, W.D., Grassini, P., Naeve, S., Specht, J.E., Mamo, M., Seymour, R., Nygren, A., Mueller, N., Sivits, S., Proctor, C., Rees, J., Whitney, T., La Menza, N.C., 2023b. Irrigation increases on-farm soybean yields in water-limited environments without a trade-off in seed protein concentration. *Field Crops Res.* 304, 109163. <https://doi.org/10.1016/j.fcr.2023.109163>
- Carciochi, W.D., Schwalbert, R., Andrade, F.H., Corassa, G.M., Carter, P., Gaspar, A.P., Schmidt, J., Ciampitti, I.A., 2019. Soybean Seed Yield Response to Plant Density by Yield Environment in North America. *Agron. J.* 111, 1923–1932. <https://doi.org/10.2134/agronj2018.10.0635>

- Carpenter, A.C., Board, J.E., 1997a. Branch Yield Components Controlling Soybean Yield Stability across Plant Populations. *Crop Sci.* 37, 885–891.
<https://doi.org/10.2135/cropsci1997.0011183X003700030031x>
- Carpenter, A.C., Board, J.E., 1997b. Growth Dynamic Factors Controlling Soybean Yield Stability across Plant Populations. *Crop Sci.* 37, 1520–1526.
<https://doi.org/10.2135/cropsci1997.0011183X003700050018x>
- Carroll, M.E., Riera, L.G., Miller, B.A., Dixon, P.M., Ganapathysubramanian, B., Sarkar, S., Singh, A.K., 2024. Leveraging soil mapping and machine learning to improve spatial adjustments in plant breeding trials. *Crop Sci.* csc2.21336.
<https://doi.org/10.1002/csc2.21336>
- Casal, J.J., Sanchez, R.A., Gibson, D., 1990. The significance of changes in the red/far-red ratio, associated with either neighbour plants or twilight, for tillering in *Lolium multiflorum* Lam. *New Phytol.* 116, 565–572. <https://doi.org/10.1111/j.1469-8137.1990.tb00540.x>
- Cerrudo, A., Miller-Garvin, J., Naeve, S.L., 2023. Western Hemisphere quality and production capacity of soybean protein. *Front. Sustain. Food Syst.* 7, 1223921.
<https://doi.org/10.3389/fsufs.2023.1223921>
- Chaney, N.W., Minasny, B., Herman, J.D., Nauman, T.W., Brungard, C.W., Morgan, C.L.S., McBratney, A.B., Wood, E.F., Yimam, Y., 2019. POLARIS Soil Properties: 30-m Probabilistic Maps of Soil Properties Over the Contiguous United States. *Water Resour. Res.* 55, 2916–2938. <https://doi.org/10.1029/2018wr022797>
- Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System, in: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Presented at the KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, San Francisco California USA, pp. 785–794. <https://doi.org/10.1145/2939672.2939785>
- Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., Yuan, J., 2014. xgboost: Extreme Gradient Boosting. <https://doi.org/10.32614/CRAN.package.xgboost>
- Cober, E.R., Morrison, M.J., Ma, B., Butler, G., 2005. Genetic Improvement Rates of Short-Season Soybean Increase with Plant Population. *Crop Sci.* 45, 1029–1034.
<https://doi.org/10.2135/cropsci2004.0232>
- Collins, F.I., Cartter, J.L., 1956. Variability in Chemical Composition of Seed From Different Portions of the Soybean Plant ¹. *Agron. J.* 48, 216–219.
<https://doi.org/10.2134/agronj1956.00021962004800050006x>
- Cox, W.J., Cherney, J.H., 2011. Location, Variety, and Seeding Rate Interactions with Soybean Seed-Applied Insecticide/Fungicides. *Agron. J.* 103, 1366–1371.
<https://doi.org/10.2134/agronj2011.0129>
- David A. Hennessy, Yuyuan Che and Hongli Feng, 2019. Rising seed costs & a tale of two very different seeding rates. Michigan State University - Michigan Farm News.
- De Borja Reis, A.F., Moro Rosso, L., Purcell, L.C., Naeve, S., Casteel, S.N., Kovács, P., Archontoulis, S., Davidson, D., Ciampitti, I.A., 2021. Environmental Factors Associated With Nitrogen Fixation Prediction in Soybean. *Front. Plant Sci.* 12, 675410.
<https://doi.org/10.3389/fpls.2021.675410>
- De Bruin, J.L., Pedersen, P., 2008. Effect of Row Spacing and Seeding Rate on Soybean Yield. *Agron. J.* 100, 704–710. <https://doi.org/10.2134/agronj2007.0106>

- Deines, J.M., Swatantran, A., Ye, D., Myers, B., Archontoulis, S., Lobell, D.B., 2023. Field-scale dynamics of planting dates in the US Corn Belt from 2000 to 2020. *Remote Sens. Environ.* 291, 113551. <https://doi.org/10.1016/j.rse.2023.113551>
- Di Mauro, G., Cipriotti, P.A., Gallo, S., Rotundo, J.L., 2018. Environmental and management variables explain soybean yield gap variability in Central Argentina. *Eur. J. Agron.* 99, 186–194. <https://doi.org/10.1016/j.eja.2018.04.012>
- Digrado, A., Montes, C.M., Baxter, I., Ainsworth, E.A., 2024. Seed quality under elevated CO₂ differs in soybean cultivars with contrasting yield responses. *Glob. Change Biol.* 30, e17170. <https://doi.org/10.1111/gcb.17170>
- Egli, D.B., 2024. Expanding the availability of soybean oil for renewable diesel with a high oil–low protein cultivar. *Crop Sci.* 64, 39–43. <https://doi.org/10.1002/csc2.21133>
- Egli, D.B., Bruening, W.P., 2004. Water stress, photosynthesis, seed sucrose levels and seed growth in soybean. *J. Agric. Sci.* 142, 1–8. <https://doi.org/10.1017/S0021859604004095>
- Egli, D.B., Cornelius, P.L., 2009. A Regional Analysis of the Response of Soybean Yield to Planting Date. *Agron. J.* 101, 330–335. <https://doi.org/10.2134/agronj2008.0148>
- Egli, D.B., Leggett, J.E., 1973. Dry Matter Accumulation Patterns in Determinate and Indeterminate Soybeans¹. *Crop Sci.* 13, 220–222. <https://doi.org/10.2135/cropsci1973.0011183X001300020021x>
- EIA, 2023. Domestic renewable diesel capacity could more than double through 2025.
- Ergo, V.V., Veas, R.E., Vega, C.R.C., Lascano, R., Carrera, C.S., 2024a. Ecophysiological mechanisms underlying the positive relationship between seed protein concentration and yield in soybean under field heat and drought stress. *J. Agron. Crop Sci.* 210, e12703. <https://doi.org/10.1111/jac.12703>
- Ergo, V.V., Veas, R.E., Vega, C.R.C., Lascano, R., Carrera, C.S., 2024b. Ecophysiological mechanisms underlying the positive relationship between seed protein concentration and yield in soybean under field heat and drought stress. *J. Agron. Crop Sci.* 210, e12703. <https://doi.org/10.1111/jac.12703>
- ESA, 2021. Copernicus Sentinel-2 (processed by the European Space Agency ESA). https://doi.org/10.5270/S2_-6eb6imz
- European Union, 2018. Directive (EU) 2018/2001 of the European Parliament and of the Council on the promotion of the use of energy from renewable sources (recast). Official Journal of the European Union.
- Fehr, W.R., Caviness, C.E., 1977. Stages of Soybean Development 12.
- Fernandez-Cornejo, J., Wechsler, S., Livingston, M., Mitchell, L., 2014. Genetically Engineered Crops in the United States. SSRN Electron. J. <https://doi.org/10.2139/ssrn.2503388>
- Foley, J.A., Ramankutty, N., Brauman, K.A., Cassidy, E.S., Gerber, J.S., Johnston, M., Mueller, N.D., O’Connell, C., Ray, D.K., West, P.C., Balzer, C., Bennett, E.M., Carpenter, S.R., Hill, J., Monfreda, C., Polasky, S., Rockström, J., Sheehan, J., Siebert, S., Tilman, D., Zaks, D.P.M., 2011. Solutions for a cultivated planet. *Nature* 478, 337–342. <https://doi.org/10.1038/nature10452>
- Fraanje, W., Garnett, T., Breewood, H., 2020. Soy: food, feed, and land use change. Food Climate Research Network. <https://doi.org/10.56661/47e58c32>
- Frederick, J.R., Camp, C.R., Bauer, P.J., 2001a. Drought-Stress Effects on Branch and Mainstem Seed Yield and Yield Components of Determinate Soybean. *Crop Sci.* 41, 759–763. <https://doi.org/10.2135/cropsci2001.413759x>

- Frederick, J.R., Camp, C.R., Bauer, P.J., 2001b. Drought-Stress Effects on Branch and Mainstem Seed Yield and Yield Components of Determinate Soybean. *Crop Sci.* 41, 759–763. <https://doi.org/10.2135/cropsci2001.413759x>
- Gao, F., Zhang, X., 2021. Mapping Crop Phenology in Near Real-Time Using Satellite Remote Sensing: Challenges and Opportunities. *J. Remote Sens.* 2021, 2021/8379391. <https://doi.org/10.34133/2021/8379391>
- Gaso, D.V., De Wit, A., De Bruin, S., Puntel, L.A., Berger, A.G., Kooistra, L., 2023. Efficiency of assimilating leaf area index into a soybean model to assess within-field yield variability. *Eur. J. Agron.* 143, 126718. <https://doi.org/10.1016/j.eja.2022.126718>
- Gaspar, A.P., Mitchell, P.D., Conley, S.P., 2015. Economic Risk and Profitability of Soybean Fungicide and Insecticide Seed Treatments at Reduced Seeding Rates. *Crop Sci.* 55, 924–933. <https://doi.org/10.2135/cropsci2014.02.0114>
- Gaspar, A.P., Mourtzinis, S., Kyle, D., Galdi, E., Lindsey, L.E., Hamman, W.P., Matcham, E.G., Kandel, H.J., Schmitz, P., Stanley, J.D., Schmidt, J.P., Mueller, D.S., Nafziger, E.D., Ross, J., Carter, P.R., Varenhorst, A.J., Wise, K.A., Ciampitti, I.A., Carciochi, W.D., Chilvers, M.I., Hauswedell, B., Tenuta, A.U., Conley, S.P., 2020. Defining optimal soybean seeding rates and associated risk across North America. *Agron. J.* 112, 2103–2114. <https://doi.org/10.1002/agj2.20203>
- Gaspar, A.P., Mueller, D.S., Wise, K.A., Chilvers, M.I., Tenuta, A.U., Conley, S.P., 2017. Response of Broad-Spectrum and Target-Specific Seed Treatments and Seeding Rate on Soybean Seed Yield, Profitability, and Economic Risk. *Crop Sci.* 57, 2251–2262. <https://doi.org/10.2135/cropsci2016.11.0967>
- Gerber, J.S., Ray, D.K., Makowski, D., Butler, E.E., Mueller, N.D., West, P.C., Johnson, J.A., Polasky, S., Samberg, L.H., Siebert, S., Sloat, L., 2024. Global spatially explicit yield gap time trends reveal regions at risk of future crop yield stagnation. *Nat. Food* 5, 125–135. <https://doi.org/10.1038/s43016-023-00913-8>
- Gitelson, A.A., Gritz †, Y., Merzlyak, M.N., 2003. Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves. *J. Plant Physiol.* 160, 271–282. <https://doi.org/10.1078/0176-1617-00887>
- Gitelson, A.A., Viña, A., Ciganda, V., Rundquist, D.C., Arkebauer, T.J., 2005. Remote estimation of canopy chlorophyll content in crops. *Geophys. Res. Lett.* 32, 2005GL022688. <https://doi.org/10.1029/2005GL022688>
- Graham, R.L., Hinkley, D.V., John, P.W.M., Shi, S., 1990. Balanced Design of Bootstrap Simulations. *J. R. Stat. Soc. Ser. B Methodol.* 52, 185–202. <https://doi.org/10.1111/j.2517-6161.1990.tb01781.x>
- Green-Tracewicz, E., Page, E.R., Swanton, C.J., 2011. Shade Avoidance in Soybean Reduces Branching and Increases Plant-to-Plant Variability in Biomass and Yield Per Plant. *Weed Sci.* 59, 43–49. <https://doi.org/10.1614/WS-D-10-00081.1>
- Hadley Wickham, 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.
- Heimann, T., Argueyrolles, R., Reinhardt, M., Schuenemann, F., Söder, M., Delzeit, R., 2023. Phasing out palm and soy oil biodiesel in the EU : What is the benefit? *GCB Bioenergy* 16, e13115. <https://doi.org/10.1111/gcbb.13115>
- Herbert, S.J., Litchfield, G.V., 1982. Partitioning Soybean Seed Yield Components ¹. *Crop Sci.* 22, 1074–1079. <https://doi.org/10.2135/cropsci1982.0011183X002200050044x>

- Hernandez, C.M., Correndo, A., Kyveryga, P., Prestholt, A., Ciampitti, I.A., 2023. On-farm soybean seed protein and oil prediction using satellite data. *Comput. Electron. Agric.* 212, 108096. <https://doi.org/10.1016/j.compag.2023.108096>
- Hertsgaard, D.J., Wilson, W.W., Dahl, B., 2019. Costs and risks of testing and blending for essential amino acids in soybeans. *Agribusiness* 35, 265–280. <https://doi.org/10.1002/agr.21576>
- Hoogenboom, G., Porter, C.H., Boote, K.J., University of Florida, USA, Shelia, V., University of Florida, USA, Wilkens, P.W., International Fertilizer Development Center, USA, Singh, U., International Fertilizer Development Center, USA, White, J.W., USDA-ARS, USA, Asseng, S., University of Florida, USA, Lizaso, J.I., Universidad Politécnica de Madrid, Spain, Moreno, L.P., University of Florida, USA, Pavan, W., Universidade de Passo Fundo, Brazil, Ogoshi, R., University of Hawaii, USA, Hunt, L.A., University of Guelph, Canada, Tsuji, G.Y., University of Hawaii, USA, Jones, J.W., University of Florida, USA, 2019. The DSSAT crop modeling ecosystem, in: *Advances in Crop Modeling for a Sustainable Agriculture*. Burleigh Dodds Science Publishing, Cambridge, United Kingdom, pp. 173–216. <https://doi.org/10.19103/AS.2019.0061.10>
- Hosseini, M., McNairn, H., Merzouki, A., Pacheco, A., 2015. Estimation of Leaf Area Index (LAI) in corn and soybeans using multi-polarization C- and L-band radar data. *Remote Sens. Environ.* 170, 77–89. <https://doi.org/10.1016/j.rse.2015.09.002>
- Huber, S.C., Li, K., Nelson, R., Ulanov, A., DeMuro, C.M., Baxter, I., 2016. Canopy position has a profound effect on soybean seed composition. *PeerJ* 4, e2452. <https://doi.org/10.7717/peerj.2452>
- Hurburgh, C.R., 1994. Long-term soybean composition patterns and their effect on processing. *J. Am. Oil Chem. Soc.* 71, 1425–1427. <https://doi.org/10.1007/BF02541367>
- Hurburgh, C.R., Brumm, T.J., Guinn, J.M., Hartwig, R.A., 1990. Protein and oil patterns in U.S. and world soybean markets. *J. Am. Oil Chem. Soc.* 67, 966–973. <https://doi.org/10.1007/BF02541859>
- Ignacio Ciampitti, Adrian Correndo, Emmanuela van Versendaal, 2022. Soybean planting date and maturity group selection (No. Issue 902), *Agronomy eUpdate* April 14th, 2022. Kansas State University.
- Jérémy Gelb, 2021. Package Geocmeans, an R package to perform Spatial Fuzzy C-means.
- Jones, J.W., Hoogenboom, G., Porter, C.H., Boote, K.J., Batchelor, W.D., Hunt, L.A., Wilkens, P.W., Singh, U., Gijsman, A.J., Ritchie, J.T., 2003. The DSSAT cropping system model. *Eur. J. Agron.* 18, 235–265. [https://doi.org/10.1016/S1161-0301\(02\)00107-7](https://doi.org/10.1016/S1161-0301(02)00107-7)
- Joshi, D.R., Clay, S.A., Sharma, P., Rekabdarkolae, H.M., Kharel, T., Rizzo, D.M., Thapa, R., Clay, D.E., 2023. Artificial intelligence and satellite-based remote sensing can be used to predict soybean (*Glycine max*) yield. *Agron. J.* 116, 917–930. <https://doi.org/10.1002/agj2.21473>
- Karayel, D., 2009. Performance of a modified precision vacuum seeder for no-till sowing of maize and soybean. *Soil Tillage Res.* 104, 121–125. <https://doi.org/10.1016/j.still.2009.02.001>
- Koester, R.P., Nohl, B.M., Diers, B.W., Ainsworth, E.A., 2016. Has photosynthetic capacity increased with 80 years of soybean breeding? An examination of historical soybean cultivars: Photosynthesis in historical soybean cultivars. *Plant Cell Environ.* 39, 1058–1067. <https://doi.org/10.1111/pce.12675>

- Koester, R.P., Skoneczka, J.A., Cary, T.R., Diers, B.W., Ainsworth, E.A., 2014. Historical gains in soybean (*Glycine max* Merr.) seed yield are driven by linear increases in light interception, energy conversion, and partitioning efficiencies. *J. Exp. Bot.* 65, 3311–3321. <https://doi.org/10.1093/jxb/eru187>
- Kravchenko, A.N., Bullock, D.G., 2002. Spatial Variability of Soybean Quality Data as a Function of Field Topography: I. Spatial Data Analysis. *Crop Sci.* 42, 804–815. <https://doi.org/10.2135/cropsci2002.8040>
- Lee, C.D., Egli, D.B., TeKrony, D.M., 2008. Soybean Response to Plant Population at Early and Late Planting Dates in the Mid-South. *Agron. J.* 100, 971–976. <https://doi.org/10.2134/agronj2007.0210>
- Li, Y., Zeng, H., Zhang, M., Wu, B., Zhao, Y., Yao, X., Cheng, T., Qin, X., Wu, F., 2023. A county-level soybean yield prediction framework coupled with XGBoost and multidimensional feature engineering. *Int. J. Appl. Earth Obs. Geoinformation* 118, 103269. <https://doi.org/10.1016/j.jag.2023.103269>
- Liu, W., Tollenaar, M., Stewart, G., Deen, W., 2004. Within-Row Plant Spacing Variability Does Not Affect Corn Yield. *Agron. J.* 96, 6.
- Lou, Z., Wang, F., Peng, D., Zhang, X., Xu, J., Zhu, X., Wang, Y., Shi, Z., Yu, L., Liu, G., Xie, Q., Dou, C., 2023. Combining shape and crop models to detect soybean growth stages. *Remote Sens. Environ.* 298, 113827. <https://doi.org/10.1016/j.rse.2023.113827>
- Lu, P., Jiang, B., Weiner, J., 2020. Crop spatial uniformity, yield and weed suppression, in: *Advances in Agronomy*. Elsevier, pp. 117–178. <https://doi.org/10.1016/bs.agron.2019.12.003>
- Maddonni, G.A., Otegui, M.E., 2004. Intra-specific competition in maize: early establishment of hierarchies among plants affects final kernel set. *Field Crops Res.* 85, 1–13. [https://doi.org/10.1016/S0378-4290\(03\)00104-7](https://doi.org/10.1016/S0378-4290(03)00104-7)
- Marin, F.R., Zanon, A.J., Monzon, J.P., Andrade, J.F., Silva, E.H.F.M., Richter, G.L., Antolin, L.A.S., Ribeiro, B.S.M.R., Ribas, G.G., Battisti, R., Heinemann, A.B., Grassini, P., 2022. Protecting the Amazon forest and reducing global warming via agricultural intensification. *Nat. Sustain.* 5, 1018–1026. <https://doi.org/10.1038/s41893-022-00968-8>
- Mark Jeschke and Nanticha Lutt, 2016. Row Width in Soybean Production. Pioneer.
- Masino, A., Rugeroni, P., Borrás, L., Rotundo, J.L., 2018a. Spatial and temporal plant-to-plant variability effects on soybean yield. *Eur. J. Agron.* 98, 14–24. <https://doi.org/10.1016/j.eja.2018.02.006>
- Masino, A., Rugeroni, P., Borrás, L., Rotundo, J.L., 2018b. Spatial and temporal plant-to-plant variability effects on soybean yield. *Eur. J. Agron.* 98, 14–24. <https://doi.org/10.1016/j.eja.2018.02.006>
- Mehrabi, Z., Delzeit, R., Ignaciuk, A., Levers, C., Braich, G., Bajaj, K., Amo-Aidoo, A., Anderson, W., Balgah, R.A., Benton, T.G., Chari, M.M., Ellis, E.C., Gahi, N.Z., Gaupp, F., Garibaldi, L.A., Gerber, J.S., Godde, C.M., Grass, I., Heimann, T., Hirons, M., Hoogenboom, G., Jain, M., James, D., Makowski, D., Masamha, B., Meng, S., Monprapussorn, S., Müller, D., Nelson, A., Newlands, N.K., Noack, F., Oronje, M., Raymond, C., Reichstein, M., Rieseberg, L.H., Rodriguez-Llanes, J.M., Rosenstock, T., Rowhani, P., Sarhadi, A., Seppelt, R., Sidhu, B.S., Snapp, S., Soma, T., Sparks, A.H., Teh, L., Tigchelaar, M., Vogel, M.M., West, P.C., Wittman, H., You, L., 2022. Research priorities for global food security under extreme events. *One Earth* 5, 756–766. <https://doi.org/10.1016/j.oneear.2022.06.008>

- Moore, S.H., 1991. Uniformity of Plant Spacing Effect on Soybean Population Parameters. *Crop Sci.* 31, 1049–1051. <https://doi.org/10.2135/cropsci1991.0011183X003100040041x>
- Moro Rosso, L.H., de Borja Reis, A.F., Ciampitti, I.A., 2021. Vertical Canopy Profile and the Impact of Branches on Soybean Seed Composition. *Front. Plant Sci.* 12, 725767. <https://doi.org/10.3389/fpls.2021.725767>
- Moro Rosso, L.H., De Borja Reis, A.F., Tamagno, S., Correndo, A.A., Prasad, P.V.V., Ciampitti, I.A., 2023. Temporal variation of soil N supply defines N fixation in soybeans. *Eur. J. Agron.* 144, 126745. <https://doi.org/10.1016/j.eja.2023.126745>
- Mourtzinis, S., Specht, J.E., Conley, S.P., 2019. Defining Optimal Soybean Sowing Dates across the US. *Sci. Rep.* 9, 2800. <https://doi.org/10.1038/s41598-019-38971-3>
- Munier-Jolain, N., Ney, B., Duthion, C., 1994. Reproductive Development of an Indeterminate Soybean as Affected by Morphological Position. *Crop Sci.* 34, 1009–1013. <https://doi.org/10.2135/cropsci1994.0011183X003400040033x>
- Naeve, S.L., Huerd, S.C., 2008. Year, Region, and Temperature Effects on the Quality of Minnesota's Soybean Crop. *Agron. J.* 100, 690–695. <https://doi.org/10.2134/agronj2007.0204>
- Norsworthy, J.K., Shipe, E.R., 2005. Effect of Row Spacing and Soybean Genotype on Mainstem and Branch Yield. *Agron. J.* 97, 919–923. <https://doi.org/10.2134/agronj2004.0271>
- OECD, FAO-UN, 2024. OECD-FAO Agricultural Outlook 2024-2033, OECD-FAO Agricultural Outlook. OECD. <https://doi.org/10.1787/4c5d2cfb-en>
- Patrignani, A., Ochsner, T.E., 2015. Canopeo: A Powerful New Tool for Measuring Fractional Green Canopy Cover. *Agron. J.* 107, 2312–2320. <https://doi.org/10.2134/agronj15.0150>
- Pedersen, P., Lauer, J.G., 2003. Corn and Soybean Response to Rotation Sequence, Row Spacing, and Tillage System. *Agron. J.* 95, 965–971. <https://doi.org/10.2134/agronj2003.9650>
- Pipolo, A.E., Sinclair, T.R., Camara, G.M.S., 2004. Effects of temperature on oil and protein concentration in soybean seeds cultured *in vitro*. *Ann. Appl. Biol.* 144, 71–76. <https://doi.org/10.1111/j.1744-7348.2004.tb00318.x>
- Posit team, 2023. RStudio: Integrated Development Environment for R.
- R Core Team, 2024. R: A Language and Environment for Statistical Computing.
- R Core Team, 2022. R: A Language and Environment for Statistical Computing.
- Ray, D.K., Ramankutty, N., Mueller, N.D., West, P.C., Foley, J.A., 2012. Recent patterns of crop yield growth and stagnation. *Nat. Commun.* 3, 1293. <https://doi.org/10.1038/ncomms2296>
- Rincker, K., Nelson, R., Specht, J., Sleper, D., Cary, T., Cianzio, S.R., Casteel, S., Conley, S., Chen, P., Davis, V., Fox, C., Graef, G., Godsey, C., Holshouser, D., Jiang, G.-L., Kantartzi, S.K., Kenworthy, W., Lee, C., Mian, R., McHale, L., Naeve, S., Orf, J., Poysa, V., Schapaugh, W., Shannon, G., Uniatowski, R., Wang, D., Diers, B., 2014a. Genetic Improvement of U.S. Soybean in Maturity Groups II, III, and IV. *Crop Sci.* 54, 1419–1432. <https://doi.org/10.2135/cropsci2013.10.0665>
- Rincker, K., Nelson, R., Specht, J., Sleper, D., Cary, T., Cianzio, S.R., Casteel, S., Conley, S., Chen, P., Davis, V., Fox, C., Graef, G., Godsey, C., Holshouser, D., Jiang, G.-L., Kantartzi, S.K., Kenworthy, W., Lee, C., Mian, R., McHale, L., Naeve, S., Orf, J., Poysa, V., Schapaugh, W., Shannon, G., Uniatowski, R., Wang, D., Diers, B., 2014b. Genetic

- Improvement of U.S. Soybean in Maturity Groups II, III, and IV. *Crop Sci.* 54, 1419–1432. <https://doi.org/10.2135/cropsci2013.10.0665>
- Ritchie, H., 2021. Drivers of Deforestation. Our World Data.
- Rotundo, J.L., Miller-Garvin, J.E., Naeve, S.L., 2016. Regional and Temporal Variation in Soybean Seed Protein and Oil across the United States. *Crop Sci.* 56, 797–808. <https://doi.org/10.2135/cropsci2015.06.0394>
- Rotundo, J.L., Westgate, M.E., 2010. Rate and Duration of Seed Component Accumulation in Water-Stressed Soybean. *Crop Sci.* 50, 676–684. <https://doi.org/10.2135/cropsci2009.05.0240>
- Rotundo, J.L., Westgate, M.E., 2009a. Meta-analysis of environmental effects on soybean seed composition. *Field Crops Res.* 110, 147–156. <https://doi.org/10.1016/j.fcr.2008.07.012>
- Rotundo, J.L., Westgate, M.E., 2009b. Meta-analysis of environmental effects on soybean seed composition. *Field Crops Res.* 110, 147–156. <https://doi.org/10.1016/j.fcr.2008.07.012>
- Rousseeuw, P.J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20, 53–65. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)
- Rowntree, S.C., Suhre, J.J., Weidenbenner, N.H., Wilson, E.W., Davis, V.M., Naeve, S.L., Casteel, S.N., Diers, B.W., Esker, P.D., Specht, J.E., Conley, S.P., 2013. Genetic Gain $\times$ Management Interactions in Soybean: I. Planting Date. *Crop Sci.* 53, 1128–1138. <https://doi.org/10.2135/cropsci2012.03.0157>
- S. D. Kachman, J. A. Smith, 1995. Alternative Measures of Accuracy in Plant Spacing for Planters Using Single Seed Metering. *Trans. ASAE* 38, 379–387. <https://doi.org/10.13031/2013.27843>
- Sacks, W.J., Kucharik, C.J., 2011. Crop management and phenology trends in the U.S. Corn Belt: Impacts on yields, evapotranspiration and energy balance. *Agric. For. Meteorol.* 151, 882–894. <https://doi.org/10.1016/j.agrformet.2011.02.010>
- Salmerón, M., Bourland, F.M., Buehring, N.W., Earnest, L., Fritschi, F.B., Gbur, E.E., Golden, B.R., Hathcoat, D., Lofton, J., McClure, A.T., Miller, T.D., Neely, C., Shannon, G., Udeigwe, T.K., Verbree, D.A., Vories, E.D., Wiebold, W.J., Purcell, L.C., 2022. Regional analysis of planting date and cultivar maturity recommendations that improve soybean oil yield and meal protein concentration. *Front. Plant Sci.* 13, 954111. <https://doi.org/10.3389/fpls.2022.954111>
- Salmerón, M., Gbur, E.E., Bourland, F.M., Buehring, N.W., Earnest, L., Fritschi, F.B., Golden, B.R., Hathcoat, D., Lofton, J., McClure, A.T., Miller, T.D., Neely, C., Shannon, G., Udeigwe, T.K., Verbree, D.A., Vories, E.D., Wiebold, W.J., Purcell, L.C., 2016. Yield Response to Planting Date Among Soybean Maturity Groups for Irrigated Production in the US Midsouth. *Crop Sci.* 56, 747–759. <https://doi.org/10.2135/cropsci2015.07.0466>
- Sarkar, S., Sagan, V., Bhadra, S., Fritschi, F.B., 2024. Spectral enhancement of PlanetScope using Sentinel-2 images to estimate soybean yield and seed composition. *Sci. Rep.* 14, 15063. <https://doi.org/10.1038/s41598-024-63650-3>
- Sarkar, S., Sagan, V., Bhadra, S., Rhodes, K., Pokharel, M., Fritschi, F.B., 2023. Soybean seed composition prediction from standing crops using PlanetScope satellite imagery and machine learning. *ISPRS J. Photogramm. Remote Sens.* 204, 257–274. <https://doi.org/10.1016/j.isprsjprs.2023.09.010>
- Seth L. Naeve, J.C., 2023. United States soybean quality annual report 2023. ASA, USB, USSEC.

- Settimi, J.R., Board, J.E., 1988a. Photoperiod and Planting Date Effects on the Spatial Distribution of Branch Development in Soybean. *Crop Sci.* 28, 259–263. <https://doi.org/10.2135/cropsci1988.0011183X002800020016x>
- Settimi, J.R., Board, J.E., 1988b. Photoperiod and Planting Date Effects on the Spatial Distribution of Branch Development in Soybean. *Crop Sci.* 28, 259–263. <https://doi.org/10.2135/cropsci1988.0011183X002800020016x>
- Sharma, S., Kaur, A., Bansal, A., Gill, B.S., 2013. Positional effects on soybean seed composition during storage. *J. Food Sci. Technol.* 50, 353–359. <https://doi.org/10.1007/s13197-011-0341-0>
- Shen, Y., Zhang, X., Yang, Z., 2022. Mapping corn and soybean phenometrics at field scales over the United States Corn Belt by fusing time series of Landsat 8 and Sentinel-2 data with VIIRS data. *ISPRS J. Photogramm. Remote Sens.* 186, 55–69. <https://doi.org/10.1016/j.isprsjprs.2022.01.023>
- Shi, G., Chavas, J.-P., 2009. Pricing of Herbicide-Tolerant Soybean Seeds: A Market-Structure Approach 8.
- Siler, T.B., Singh, M.P., 2022. Optimal soybean maturity group selection is influenced by planting date in northern production systems. *Crop Sci.* 62, 2462–2475. <https://doi.org/10.1002/csc2.20829>
- Singh, B., Kaur, G., Quintana-Ashwell, N.E., Singh, G., Lo, T.H., Nelson, K.A., 2023. Row spacing and irrigation management affect soybean yield, water use efficiency and economics. *Agric. Water Manag.* 277, 108087. <https://doi.org/10.1016/j.agwat.2022.108087>
- Singh, M., Thapa, R., Singh, N., Mirsky, S.B., Acharya, B.S., Jhala, A.J., 2023. Does narrow row spacing suppress weeds and increase yields in corn and soybean? A meta-analysis. *Weed Sci.* 71, 520–535. <https://doi.org/10.1017/wsc.2023.50>
- Specht, J.E., Diers, B.W., Nelson, R.L., de Toledo, J.F.F., Torrion, J.A., Grassini, P., 2015. Soybean, in: Smith, S., Diers, B., Specht, J., Carver, B. (Eds.), *CSSA Special Publications*. American Society of Agronomy and Soil Science Society of America, Madison, WI, USA, pp. 311–355. <https://doi.org/10.2135/cssaspecpub33.c12>
- Stivers, R.K., Swearingin, M.L., 1980. Soybean Yield Compensation with Different Populations and Missing Plant Patterns ¹. *Agron. J.* 72, 98–102. <https://doi.org/10.2134/agronj1980.00021962007200010019x>
- Suhre, J.J., Weidenbenner, N.H., Rowntree, S.C., Wilson, E.W., Naeve, S.L., Conley, S.P., Casteel, S.N., Diers, B.W., Esker, P.D., Specht, J.E., Davis, V.M., 2014a. Soybean Yield Partitioning Changes Revealed by Genetic Gain and Seeding Rate Interactions. *Agron. J.* 106, 1631–1642. <https://doi.org/10.2134/agronj14.0003>
- Suhre, J.J., Weidenbenner, N.H., Rowntree, S.C., Wilson, E.W., Naeve, S.L., Conley, S.P., Casteel, S.N., Diers, B.W., Esker, P.D., Specht, J.E., Davis, V.M., 2014b. Soybean Yield Partitioning Changes Revealed by Genetic Gain and Seeding Rate Interactions. *Agron. J.* 106, 1631–1642. <https://doi.org/10.2134/agronj14.0003>
- Tamagno, S., Sadras, V.O., Orteza, O.A., Ciampitti, I.A., 2020. Allometric analysis reveals enhanced reproductive allocation in historical set of soybean varieties. *Field Crops Res.* 248, 107717. <https://doi.org/10.1016/j.fcr.2020.107717>
- Thakur, M., Hurburgh, C.R., 2007. Quality of US Soybean Meal Compared to the Quality of Soybean Meal from Other Origins. *J. Am. Oil Chem. Soc.* 84, 835–843. <https://doi.org/10.1007/s11746-007-1107-8>

- Thompson, N.M., Larson, J.A., Lambert, D.M., Roberts, R.K., Mengistu, A., Bellaloui, N., Walker, E.R., 2015. Mid-South Soybean Yield and Net Return as Affected by Plant Population and Row Spacing. *Agron. J.* 107, 979–989. <https://doi.org/10.2134/agronj14.0453>
- Thornton, M.M., Shrestha, R., Wei, Y., Thornton, P.E., Kao, S.-C., 2022. Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 4 R1. <https://doi.org/10.3334/ORNLDAAAC/2129>
- Többen, J., Stöver, B., Reuschel, S., Distelkamp, M., Lutz, C., 2024. Sustainability implications of the EU's bioeconomy transition along global supply chains. *J. Clean. Prod.* 461, 142565. <https://doi.org/10.1016/j.jclepro.2024.142565>
- USDA - ERS, 2021a. Commodity Costs and Returns. United States Department of Agriculture - Economic Research Service.
- USDA - ERS, 2021b. Operating costs for soybean production vary by region of the United States. United States Department of Agriculture - Economic Research Service.
- USDA FAS, 2024. Soybean top producing countries.
- USDA, NASS, 2023. Crop Production (October 2023). USDA, National Agricultural Statistics Service.
- USDA-FAS, 2024. Oilseeds: World Markets and Trade. United States Department of Agriculture Foreign, Agricultural Service.
- USDA-FASS, 2022. USDA/Foreign Agricultural Service (FAS) Production, Supply, and Distribution (PS&D) database. [WWW Document]. URL <https://apps.fas.usda.gov/psdonline/app/index.html#/app/home>
- USDA-NASS, 2024a. Crop Production Historical Track Records 04/10/2024. United States Department of Agriculture, National Agriculture Statistics Service.
- USDA-NASS, 2024b. Crop Production 2023 Summary. *Crop Prod.*
- USSEC, 2021. U.S. Soybean meal quality advantage. U.S. Soybean Export Council (USSEC).
- Viña, A., Gitelson, A.A., Nguy-Robertson, A.L., Peng, Y., 2011. Comparison of different vegetation indices for the remote assessment of green leaf area index of crops. *Remote Sens. Environ.* 115, 3468–3478. <https://doi.org/10.1016/j.rse.2011.08.010>
- Virdi, K.S., Sreekanta, S., Dobbels, A., Haaning, A., Jarquin, D., Stupar, R.M., Lorenz, A.J., Muehlbauer, G.J., 2023. Branch angle and leaflet shape are associated with canopy coverage in soybean. *Plant Genome*. <https://doi.org/10.1002/tpg2.20304>
- Vogel, J.T., Liu, W., Olhoft, P., Crafts-Brandner, S.J., Pennycooke, J.C., Christiansen, N., 2021. Soybean Yield Formation Physiology – A Foundation for Precision Breeding Based Improvement. *Front. Plant Sci.* 12, 719706. <https://doi.org/10.3389/fpls.2021.719706>
- Wasserstein, R.L., Lazar, N.A., 2016. The ASA Statement on *p* -Values: Context, Process, and Purpose. *Am. Stat.* 70, 129–133. <https://doi.org/10.1080/00031305.2016.1154108>
- Willcott, J., Herbert, S.J., Zhi-Yi, L., 1984. Leaf area display and light interception in short-season soybeans. *Field Crops Res.* 9, 173–182. [https://doi.org/10.1016/0378-4290\(84\)90023-6](https://doi.org/10.1016/0378-4290(84)90023-6)
- William, W., Dahl, B., Hertsgaard, D., 2020. Soybean quality differentials, blending, testing and spatial arbitrage. *J. Commod. Mark.* 18, 100095. <https://doi.org/10.1016/j.jcomm.2019.100095>
- Wood, S.N., 2017. Generalized Additive Models: An Introduction with R, 2nd ed. Chapman and Hall/CRC.

- Wood, S.N., 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models: Estimation of Semiparametric Generalized Linear Models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 73, 3–36.
<https://doi.org/10.1111/j.1467-9868.2010.00749.x>
- Wood, S.N., 2003. Thin-plate regression splines. *J. R. Stat. Soc. B* 65, 95–114.
- XGBoost Developers, 2022. XGBoost Tutorials, Introduction to Boosted Trees.
- Xie, X.L., Beni, G., 1991. A validity measure for fuzzy clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* 13, 841–847. <https://doi.org/10.1109/34.85677>
- Xu, C., Li, R., Song, W., Wu, T., Sun, S., Han, T., Wu, C., 2021. High Density and Uniform Plant Distribution Improve Soybean Yield by Regulating Population Uniformity and Canopy Light Interception. *Agronomy* 11, 1880.
<https://doi.org/10.3390/agronomy11091880>
- Yang, F., Huang, S., Gao, R., Liu, W., Yong, T., Wang, X., Wu, X., Yang, W., 2014. Growth of soybean seedlings in relay strip intercropping systems in relation to light quantity and red:far-red ratio. *Field Crops Res.* 155, 245–253.
<https://doi.org/10.1016/j.fcr.2013.08.011>
- Zhai, C., Long, J., Taylor, R., Weckler, P., Wang, N., 2020. Field scale row unit vibration affecting planting quality. *Precis. Agric.* 21, 589–602. <https://doi.org/10.1007/s11119-019-09684-4>
- Zhong, X., Wang, J., Shi, X., Bai, M., Yuan, C., Cai, C., Wang, N., Zhu, X., Kuang, H., Wang, X., Su, J., He, X., Liu, X., Yang, W., Yang, C., Kong, F., Wang, E., Guan, Y., 2024. Genetically optimizing soybean nodulation improves yield and protein content. *Nat. Plants* 10, 736–742. <https://doi.org/10.1038/s41477-024-01696-x>

Appendix A - Chapter 2 Supplementary Material

Appendix Table A.1. Effect of planter type, seeding rate, and their interaction for each of the four US site years. References: Spacing between plants standard deviation (Spacing sd), percentage of perfectly spaced plants (Perfect), percentage of plants misplaced by 66% (Mis 66), percentage of plants misplaced by 33% (Mis 33), percentage of double plants (Double), percentage of short skips plants (Short-skip), percentage of long skip plants (Long-skip), percentage of double skips plants (Double-skip), percentage of greater than double skip plants (> Double-skip), achieved versus targeted evenness index mean and standard deviation (ATEI and ATEI sd, respectively), and evenness index standard deviation (EI sd).

		Yield (Mg ha ⁻¹)	Spacing sd (cm)	Perfect (%)	Mis 66 (%)	Mis 33 (%)	Double (%)	Short skip (%)	Long skip (%)	Double skip (%)	> Double skip (%)	ATEI	ATEI sd	EI	EI sd
USIrr19															
Planter¹	ME	3.01	5.8	14.2	2.9 a ²	11.3	2.5 b	10.4 a	10.4	10.4	37.9 a	2.1 a	1.0	0.6 a	0.2 b
	EE	2.88	5.8	10.0	11.2 a	10.8	10.8 a	17.9 a	7.9	6.7	24.6 a	1.5 b	1.0	0.5 a	0.3 a
	<i>p-value</i>	0.69	0.975	0.465	0.002	0.879	<0.001	0.009	0.411	0.192	0.004	<0.001	0.825	0.001	<0.001
Seeding rate (k seeds ha⁻¹) (SR)	161	3.04	5.0	6.67 b	10.0 a	11.7	7.5	13.8	6.3	9.6	34.6	1.9	1.23 a	0.5 a	0.3 a
	321	2.84	6.6	17.5 a	4.2 a	10.4	5.8	14.6	12.1	7.5	27.9	1.7	0.80 a	0.6 a	0.3 a
	<i>p-value</i>	0.548	0.124	<0.001	0.029	0.649	0.362	0.772	0.055	0.468	0.149	0.05	0.023	0.007	0.016
P x SR		0.726	0.567	0.357	1.000	0.448	0.362	0.246	1.000	0.664	0.857	0.342	0.566	0.164	0.012
USDry19															
P	ME	2.66	6.6	10.4 a	7.50 a	13.3	12.2 a	9.6 a	7.5	6.3	33.2 a	1.9 a	1.3	0.5 b	0.3 a
	EE	2.45	5.0	27.2 a	2.08 a	20.1	2.5 b	18.9 a	8.8	6.7	13.8 b	1.5 a	1.0	0.7 a	0.2 b
	<i>p-value</i>	0.441	0.226	<0.001	0.006	0.108	<0.001	0.003	0.655	0.882	<0.001	0.006	0.334	<0.001	<0.001
SR	161	2.91 a	6.7	15.8	4.6	17.9	10.4 a	13.3	7.1	9.2 a	21.7	1.9 a	1.6 a	0.5 b	0.3 a
	321	2.21 a	5.0	21.8	5.0	15.5	4.2 b	15.1	9.2	3.8 a	25.4	1.5 a	0.6 b	0.6 a	0.2 a
	<i>p-value</i>	0.01	0.215	0.168	0.831	0.565	0.005	0.461	0.414	0.034	0.313	0.02	0.002	0.001	0.01
P x SR		0.523	0.841	0.006	0.019	0.565	0.042	<0.001	0.141	0.244	0.865	0.378	0.647	0.167	0.161
USIrr20															
P	ME	4.18	3.9	10.8 a	13.3 a	11.3	6.2 a	13.8	7.5	9.2	27.9 a	1.6	0.7	0.5 b	0.3 a

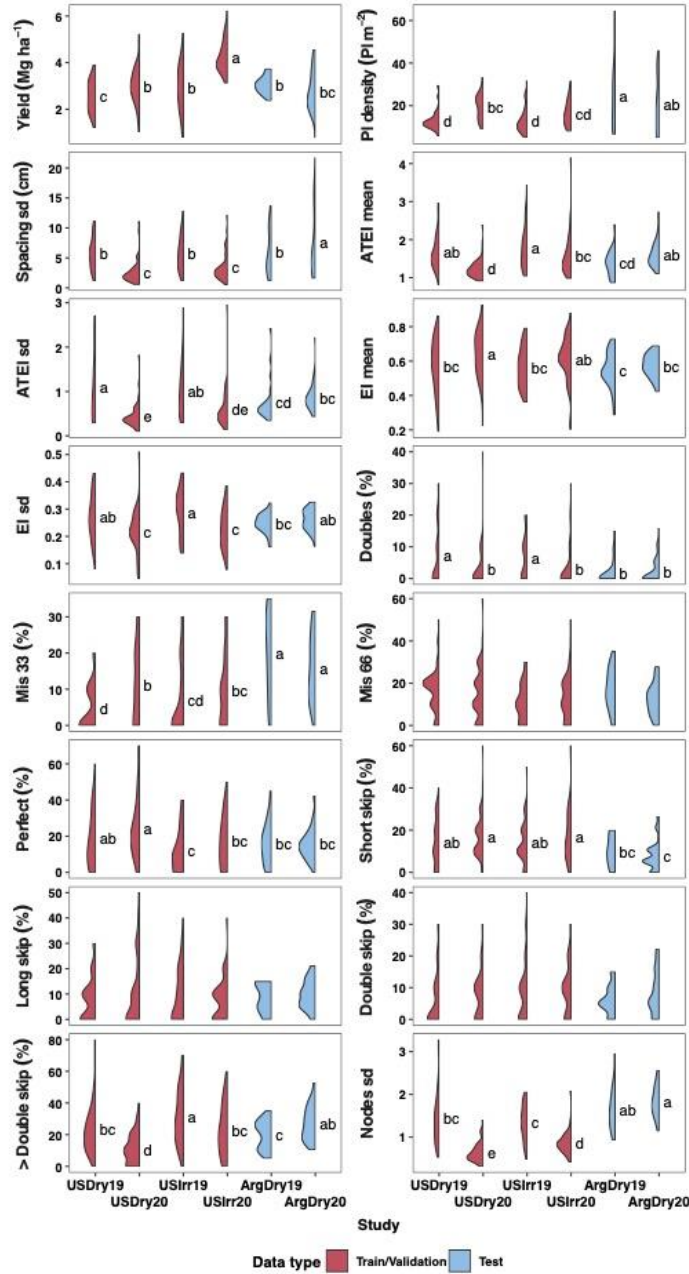
	EE	4.33	3.0	23.7 a	5.8 a	15.4	0.0 a	20.8	7.5	9.2	17.5 a	1.5	0.6	0.7 a	0.2 b
	<i>p-value</i>	0.441	0.187	0.001	0.013	0.235	<0.001	0.258	1.000	1.000	0.025	0.743	0.363	<0.001	<0.001
SR	161	4.51	4.4	20.8	7.5	15.0	2.5	18.3	5.8 ab	10.0	20.0	1.5	0.5	0.6	0.2
	215	4.39	3.7	18.3	10.0	7.5	5.0	20.8	8.3 ab	9.2	20.8	1.6	0.6	0.6	0.2
	269	3.96	2.9	13.3	9.2	14.2	0.8	12.5	12.5 a	11.7	25.8	1.6	0.6	0.6	0.2
	321	4.16	2.8	16.7	11.7	16.7	4.2	17.5	3.3 b	5.8	24.2	1.6	0.7	0.6	0.3
	<i>p-value</i>	0.178	0.255	0.496	0.806	0.268	0.356	0.58	0.036	0.313	0.79	0.971	0.865	0.458	0.063
P x SR		0.424	0.455	0.306	0.534	0.43	0.356	0.838	0.731	0.037	0.896	0.616	0.484	0.286	0.031
USDry20															
P	ME	2.90	2.8	32.9 a	5.0 a	16.3	0.4 a	16.7	12.1	5.8	10.8	1.3 a	0.5	0.7 a	0.19 a
	EE	3.09	2.5	15.4 a	17.5 b	14.2	5.4 a	17.9	8.8	9.6	11.3	1.2 a	0.4	0.5 b	0.25 a
	<i>p-value</i>	0.575	0.611	0.002	<0.001	0.574	0.008	0.724	0.300	0.083	0.905	0.030	0.535	<0.001	0.001
SR	161	2.79	4.1 a	31.7 a	10.0	10.0	3.3	15.8	6.7	8.3	14.2	1.3	0.5	0.6	0.2
	215	3.17	2.7 ab	29.2 ab	11.7	14.2	0.8	14.2	16.7	6.7	6.7	1.2	0.4	0.7	0.2
	269	2.99	1.9 b	18.3 b	13.3	19.2	3.3	15.8	12.5	8.3	9.2	1.2	0.4	0.6	0.2
	321	3.03	1.9 b	17.5 b	10.0	17.5	4.2	23.3	5.8	7.5	14.2	1.3	0.5	0.6	0.2
	<i>p-value</i>	0.707	0.009	0.024	0.749	0.310	0.623	0.257	0.056	0.939	0.134	0.440	0.819	0.266	0.227
P x SR		0.85	0.820	0.122	0.276	0.519	0.581	0.611	0.003	0.013	0.363	0.493	0.743	0.749	0.381

¹Planter type: Max Emerge® (ME) and Exact Emerge® (EE)

²Letter and ANOVA threshold. Different letters mean significant differences among treatments.

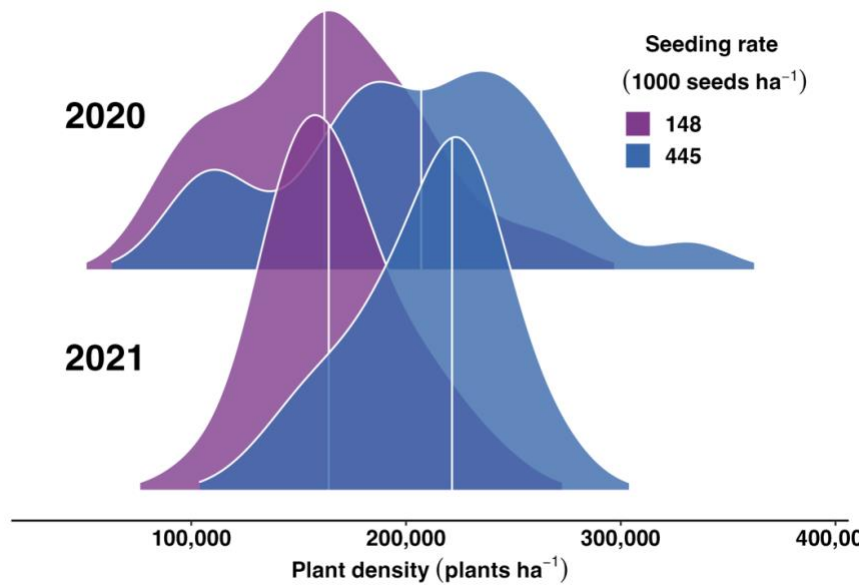
Appendix Table A.2. ANOVA table of ATEI sd (achieved versus targeted evenness index standard deviation) and seeding rate (SR) level effects on soybean yield (Mg ha⁻¹). Low, medium, and high SR levels based on SR lower than 200 seeds m⁻², between 200 and 300 seeds m⁻², and greater than 300 seeds m⁻², respectively.

	Intercept	Slope
SR level	(SR level)	(ATEI sd)
Low	3.91	-1.26
Medium	3.86	-0.732
High	4.01	-1.022
Pr(>F)	ATEI sd	<0.01
	SR level	0.83
	ATEI sd xSR level	0.26

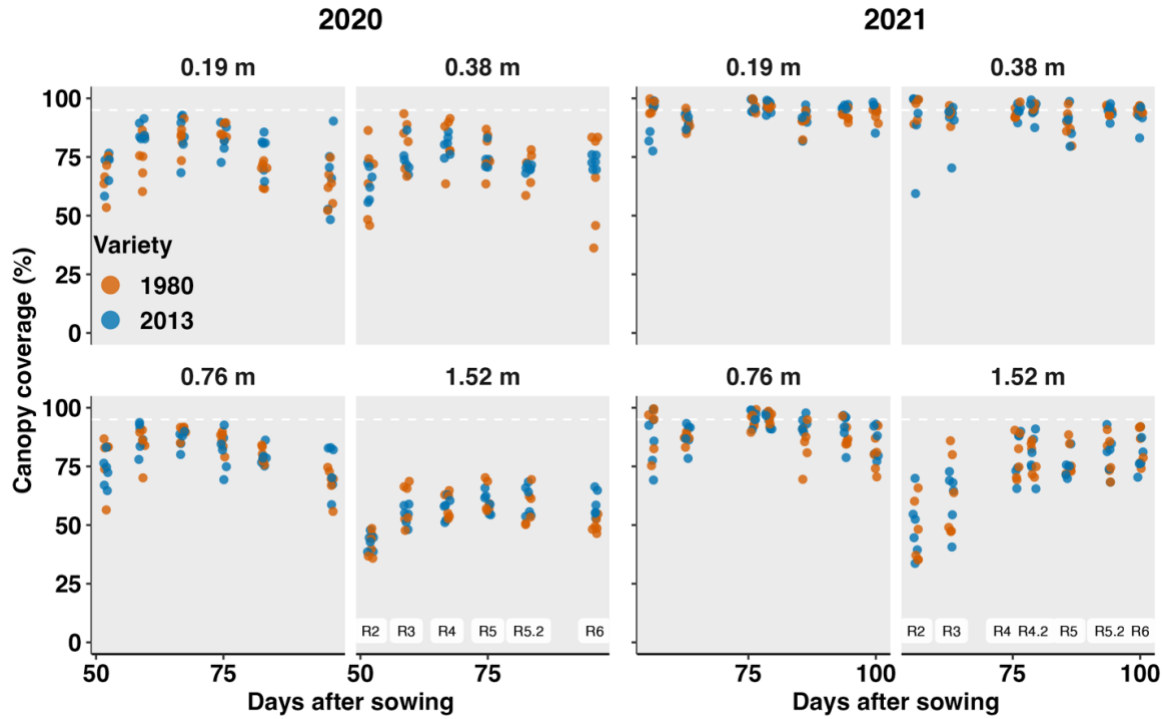


Appendix Figure A.1. Effect of site-year on yield, plant density, and all community-scale metrics of spatial uniformity (ATEI mean, ATEI sd, EI mean, EI sd, Doubles, Mis 33, Mis 66, Perfect, Short-skip, Long-skip, Double-skip and >Double-skip). Different colors denote site-year used as training (red) and testing (blue) data sets. Fisher's LSD pair-wise comparisons was performed on significant variables. Means followed by the same letter are not significantly different at $P = 0.05$. References: ATEI sd = achieved versus targeted evenness index standard deviation, EI sd = evenness index standard deviation, Double skip = percentage of double skips plants, Mis 66 = percentage of plants misplaced by 66%, U.S.: United States, Arg: Argentina, Dry: dryland, Irr: Irrigated, 19 and 20: 2019 and 2020 calendar years.

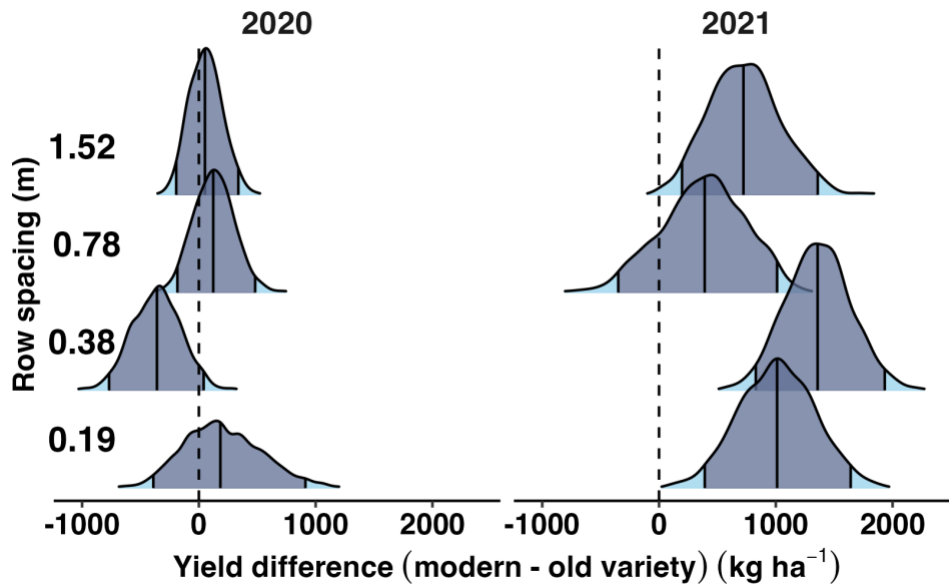
Appendix B - Chapter 3 Supplementary Material



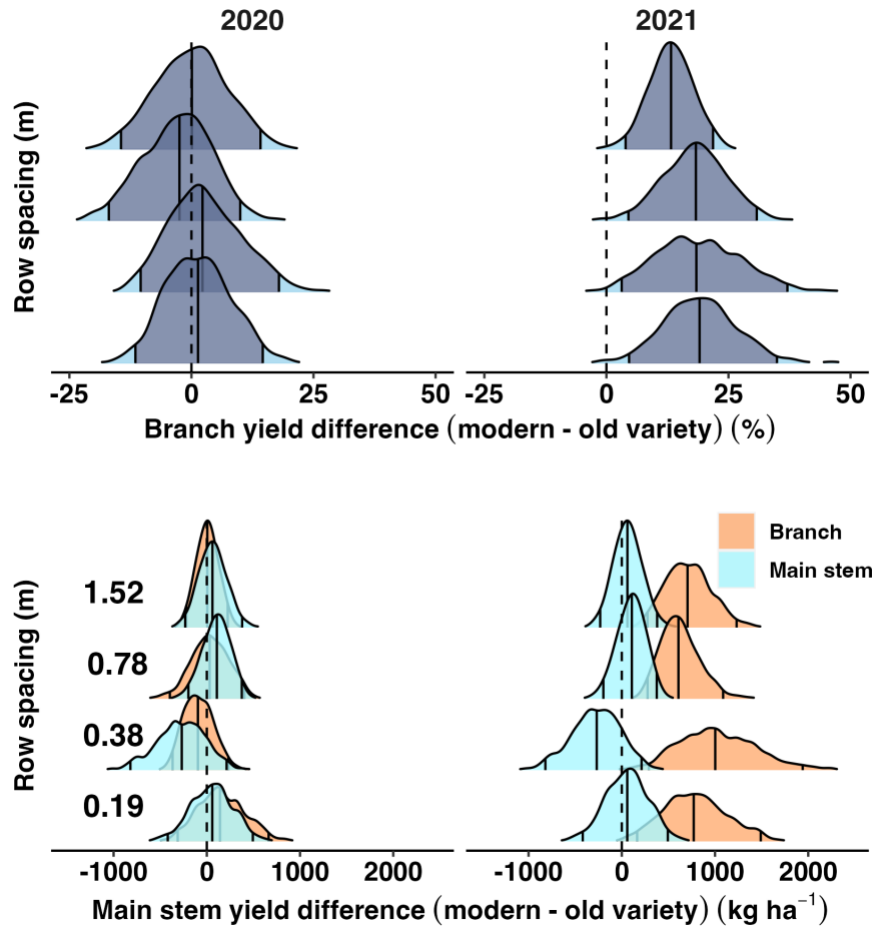
Appendix Figure B.1. Density plot of obtained plant density (plants ha⁻¹) for the evaluated seeding rates (148 and 445 thousand seeds ha⁻¹) in 2020 and 2021. White lines denote the median of plant density.



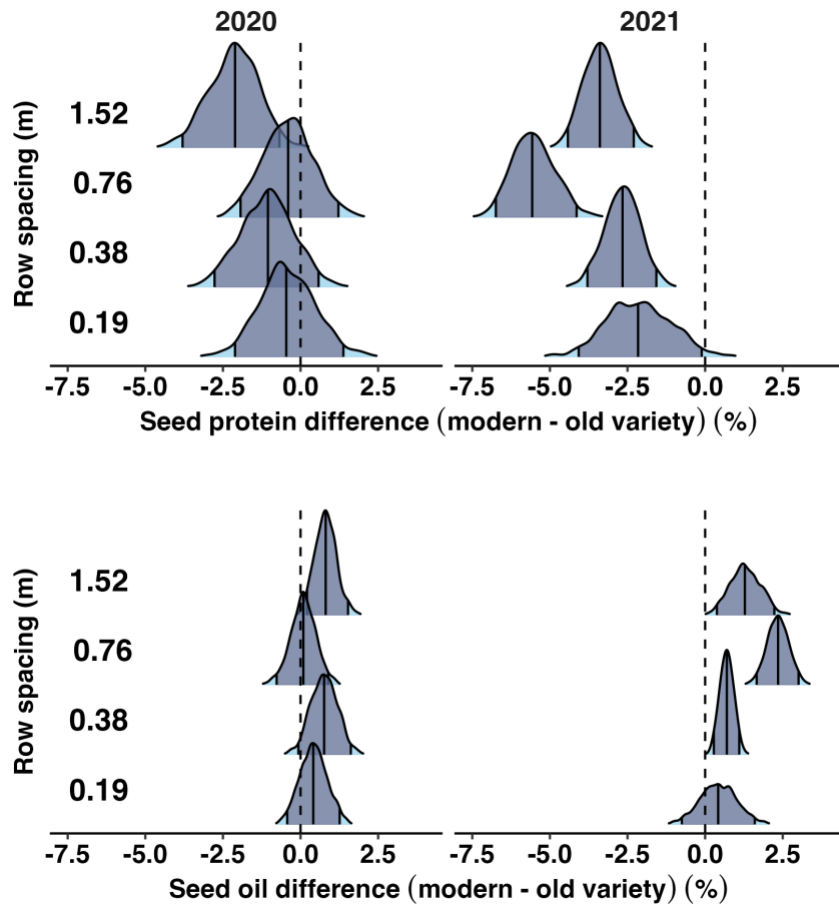
Appendix Figure B.2. Canopy coverage (%) of soybean between 50 and 100 days after sowing (R2 and R6 phenological stages, respectively) of the old (1980) and modern (2013) genotype at row spacings of 0.19, 0.38, 0.76, and 1.52 m, in 2020 and 2021. Dashed lines represent 95% of light interception.



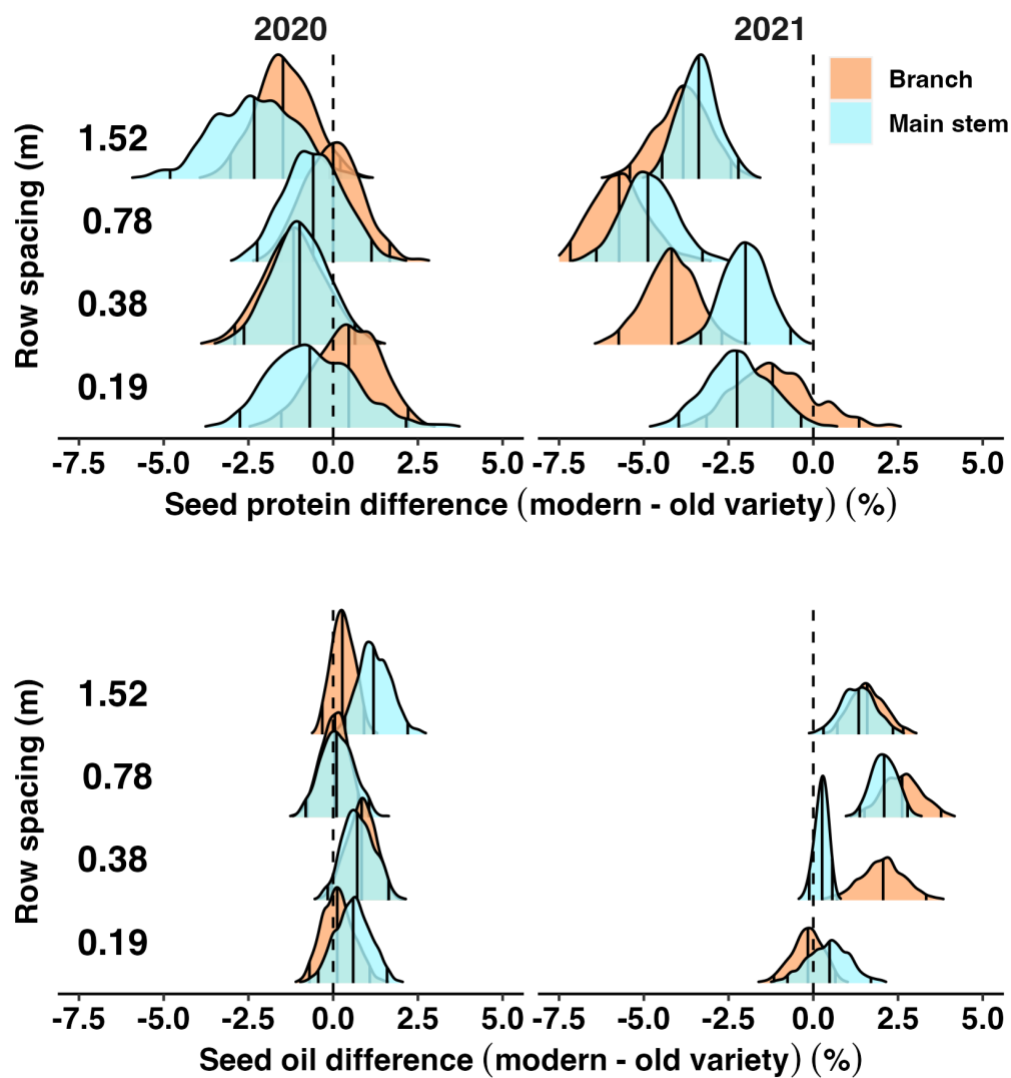
Appendix Figure B.3. Density distribution of the difference between the modern and old genotypes for yield (kg ha⁻¹) in 2020 and 2021. Lines denote the 95 % quantiles, and the mean.



Appendix Figure B.4. Density distribution of the difference between the modern and old genotype for yield in branches (% , top panel) and main stems (kg ha^{-1} , bottom panel) in 2020 and 2021. Lines denote the 95 % quantiles, and the mean.

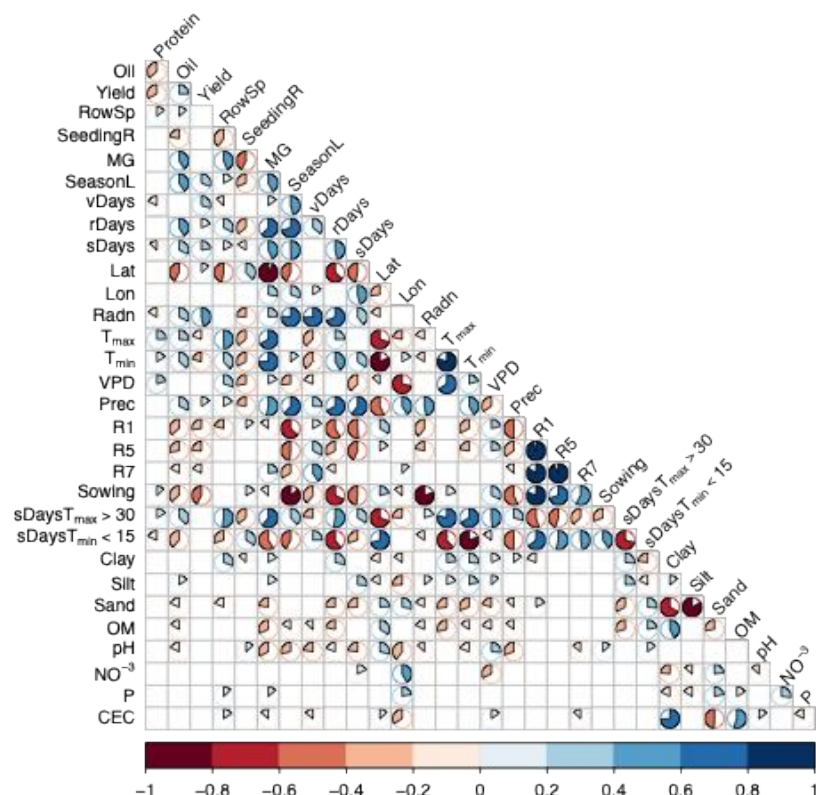


Appendix Figure B.5. Density distribution of the difference between the modern and old genotype for the seed protein (top panel) and oil (bottom panel) concentration (%) in 2020 and 2021. Lines denote the 95 % quantile, and the mean.

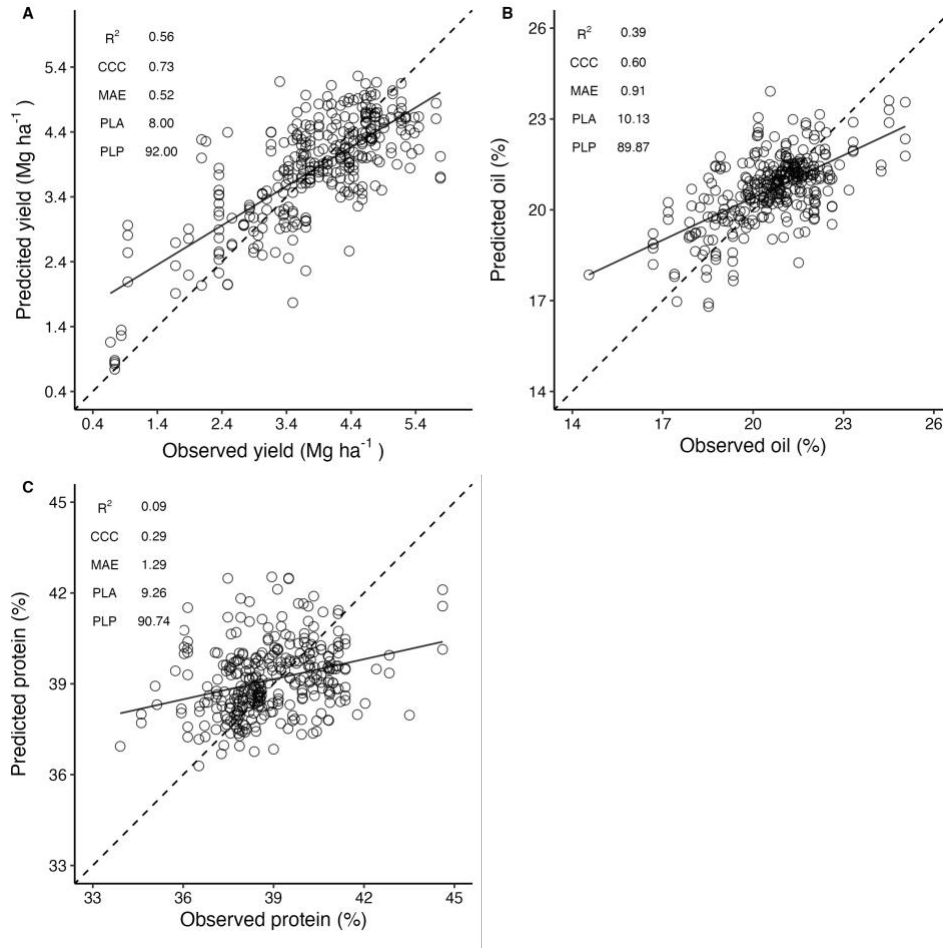


Appendix Figure B.6. Density distribution of the difference between the modern and old genotype for the seed protein (top panel) and oil (bottom panel) concentration (%) in branches and main stems in 2020 and 2021. Lines denote the 95 % quantile, and the mean.

Appendix C - Chapter 4 Supplementary Material

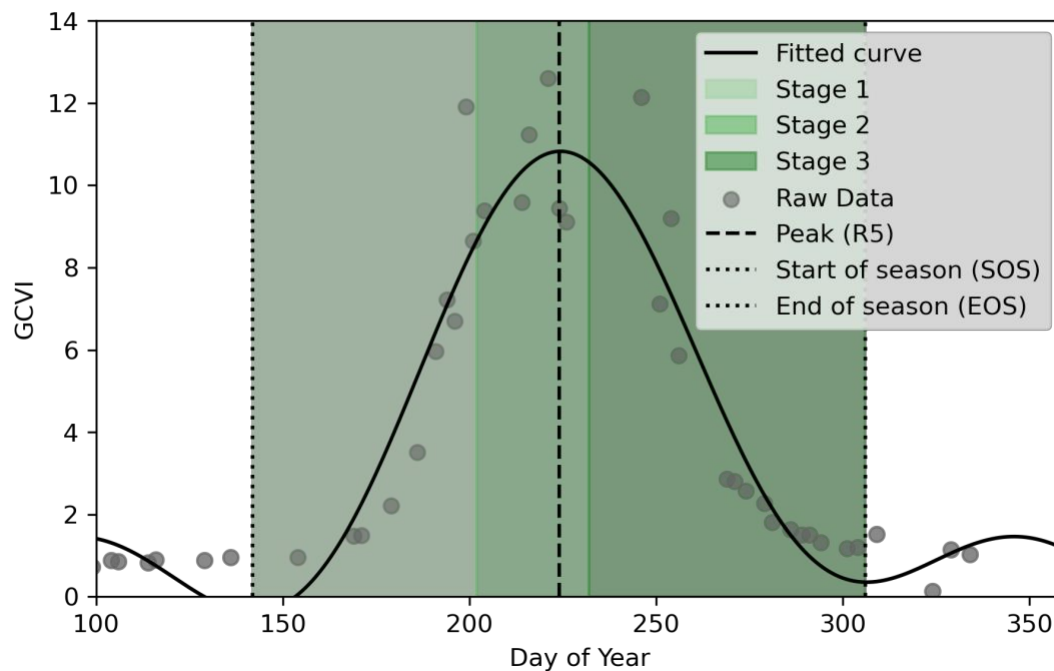


Appendix Figure C.1. Correlation matrix of environmental descriptors. Pearson's correlation is displayed with the pie charts, indicating the direction and magnitude of each correlation. Only variables significantly correlated are displayed ($p < 0.05$). Seed protein and oil (%), yield (kg ha^{-1}), row spacing (RowSp, m), seeding rate (SeedingR, seeds m^{-2}), maturity group (MG), Season length from planting (SeasonL), Latitude (Lat), Longitude (Lon), radiation (Radn, MJ m^{-2}), maximum and minimum temperature (T_{max} and T_{min} , respectively, $^{\circ}\text{C}$), vapor pressure deficit (VPD, kPa), precipitation (Prec, mm), beginning of flowering (R1), beginning of seed (R5), physiological maturity (R7), sowing date (Sowing), number of days during seed fill with maximum temperature $> 30^{\circ}\text{C}$ (Days $T_{\text{max}} > 30$), number of days during seed fill with minimum temperature $< 15^{\circ}\text{C}$ (Days $T_{\text{min}} < 15$), clay (%), silt (%), sand (%), soil organic matter (OM, %), pH, soil NO_3^- (ppm), soil P (ppm), and cation exchange capacity (CEC, $\text{meq } 100 \text{ g}^{-1} \text{ soil}$).

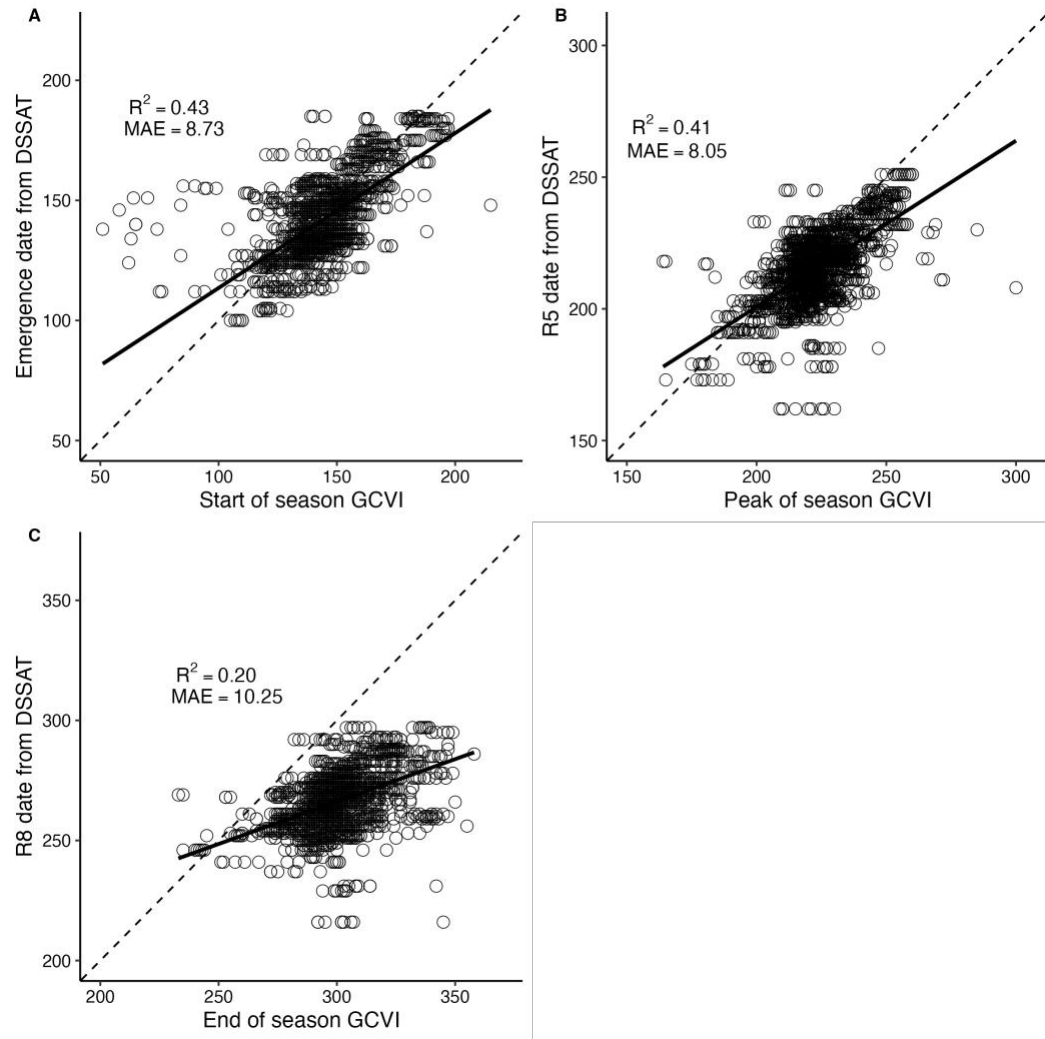


Appendix Figure C.2. Comparison of observed versus predicted yield (Mg ha^{-1} , A), oil concentration (% B), and protein concentration (% C) of soybean. The observed dots represent the median value for protein and oil, and the reported yield from a farmer field of the U.S. that was not used during model training, the dashed line shows the 1:1 relationship, and the solid line denotes the overall relationship between the observed and predicted data. Coefficient of determination (R^2), concordance correlation coefficient (CCC), mean absolute error (MAE), percentage lack of accuracy (PLA), and percentage lack of precision (PLP) are shown.

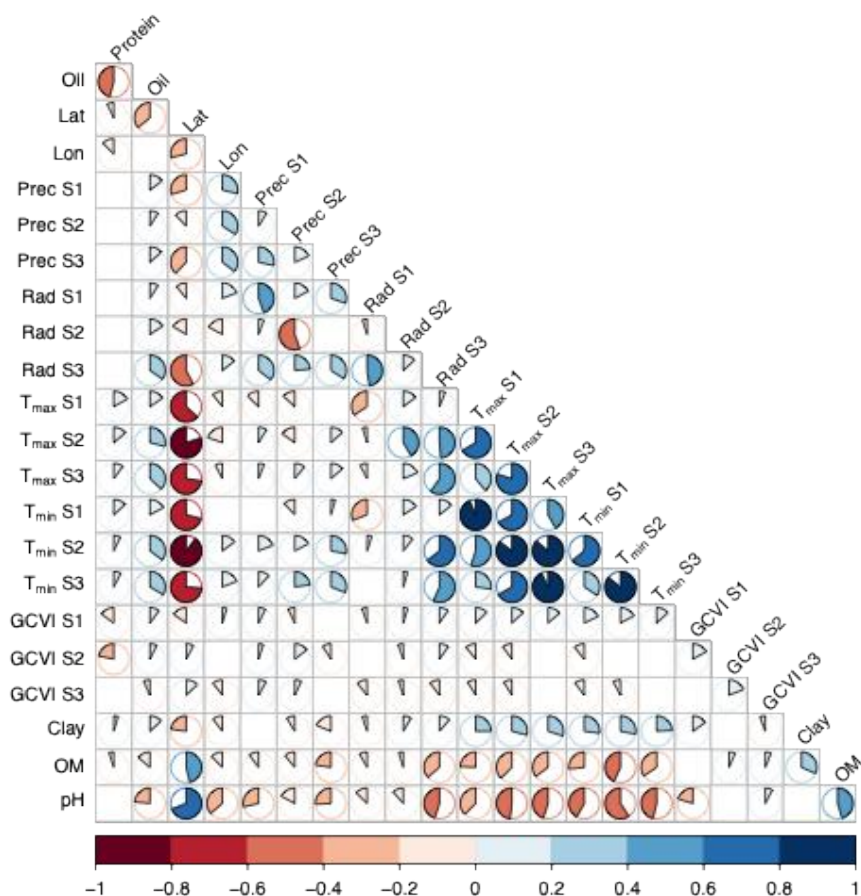
Appendix D - Chapter 5 Supplementary Material



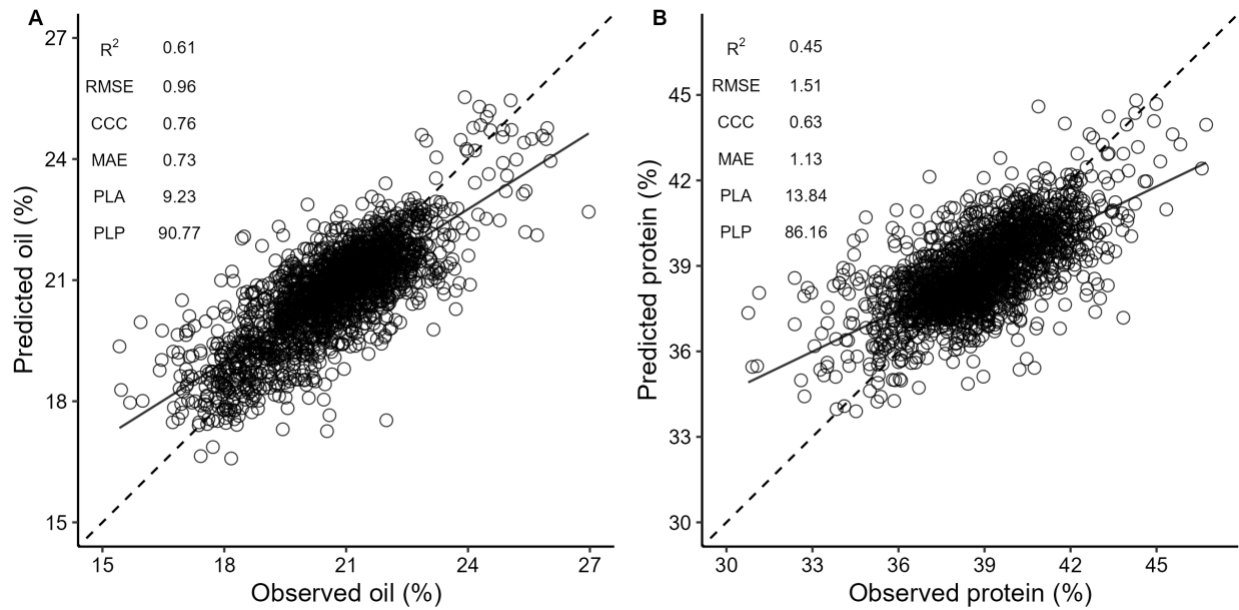
Appendix Figure D.1. Example harmonic fits between day of the year and the Green Chlorophyll Vegetation Index (GCVI). Dots represent cloud-free observations from Sentinel-2.



Appendix Figure D.2. Agreement between the start (A), peak (B), and end (C) of the season obtained with GCVI versus the emergence, beginning of seed formation (R5), and harvest maturity (R8), respectively, obtained from DSSAT. The coefficient of determination (R^2) and mean absolute error (MAE) are shown.



Appendix Figure D.3. Correlation matrix of environmental descriptors. Pearson's correlation is displayed with the pie charts, indicating the direction and magnitude of each correlation. Only variables significantly correlated are displayed ($p < 0.05$). Soybean seed protein and oil (%), Latitude (Lat), Longitude (Lon), precipitation (Prec, mm), radiation (Rad, MJ m^{-2}), maximum and minimum temperature (T_{max} and T_{min} , respectively, $^{\circ}\text{C}$), green chlorophyll vegetation index (GCVI), clay (%), soil organic matter (OM, %), pH.



Appendix Figure D.4. Comparison of observed versus predicted oil (% , A) and protein concentration (% , B) of soybean on all held-out datasets from the cross validation with best combination of hyperparameteres. The dots represent observed values from a farmer field, the dashed line shows the 1:1 relationship, and the solid line denotes the overall relationship between the observed and predicted data. Coefficient of determination (R^2), root mean square error (RMSE), concordance correlation coefficient (CCC), mean absolute error (MAE), percentage lack of accuracy (PLA), and percentage lack of precision (PLP) are shown.