A hierarchical Bayesian approach to assess the impact of environmental factors on soybean yield and yield components

Luthfan Nur Habibi[1*], *Tsutomu* Matsui[2], and *Takashi* S.T. Tanaka[3]

[1]The United Graduate School of Agricultural Science, Gifu University, Gifu 5011193, Japan
[2]Faculty of Biological Sciences, Gifu University, Gifu 5011193, Japan
[3]Department of Agroecology, Faculty of Technical Sciences, Aarhus University, Flakkebjerg, Slagelse 4200, Denmark

Abstract. Dividing soybean (*Glycine max (*L.) Merr) yield into several yield components, including the seeds per area and seed weight, offers better identification of the driver of yield variation, especially that is affected by environmental factors. The objectives of this study are to understand the relationship among yield, yield components, and environmental factors using a hierarchical Bayesian model, and to determine the potential limiting factors for soybean yield production. A hierarchical Bayesian approach offers a natural mechanism of the eco-biological system through a multi-level model. Precipitation data was used to represent the environmental factors during the key stages of soybean development. Yield in seven soybean environments, defined as the combination of location and year, was surveyed from 2018 to 2023. The results indicated that soybean yield varied between environments. Seed numbers per area was the main driver of the soybean yield. Moreover, precipitation during the early reproductive stages, where the seed was being developed, also significantly affected the final yield. Seed weights also contributed to the increase in soybean yield, even though the environmental factors during the seed-filling stage were not substantial. In summary, this study provides evidence of environmental conditions as a potential limiting factor of soybean yield.

1 Introduction

Soybean (*Glycine max* (L.) Merr.) is an important grain crop and one of the primary sources of protein and oil for the human diet. Soybean yield is categorized as grain-type yield and harvested during the maturity stages when all the moisture has evaporated. Soybean yield can be divided into several parts of yield components, written as the following equation:

$$\text{Yield} = \text{plants per area} \times \text{pods per plant} \times \text{seeds per pod} \times \text{seed weight}. \quad (1)$$

Dividing yield into yield components can be used to investigate factors affecting the final yield, as each yield component reflects environmental impacts in specific growth stages.

* Corresponding author: noerhabibii@gmail.com

However, breaking down the yield into a long list of components also creates higher complexity of the analysis and interpretation; therefore, Egli [1] proposed a simplified version of the soybean yield components, which is written as follows:

$$\text{Yield} = \text{seeds per area} \times \text{seed weight}, \quad (2)$$

which combined population-sensitive components, that is plants per area, pods per plant, and seeds per pod, into a single component, seeds per area.

Yield components are affected by environmental factors such as soil and weather conditions and can impact the final yield production. A prior study by Habibi et al. [2] revealed that a decrease in plant population affected soybean yields due to soil conditions and precipitation during the emergence stage. Precipitation is also assumed to affect yield components, especially during early reproduction stages when flowers, pods, and seeds are initiated, and during seed filling stages. Therefore, it is necessary to consider the relationship among yield, yield components, and environmental factors corresponding to the yield components.

The relationship among yield, yield components, and environmental factors is often analysed using frequentist frameworks including linear regression analysis [3]. However, the plant-environmental system is so complex that a linear relationship may not adequately explore the relationship between variables. A hierarchical Bayesian approach offers a method that could provide a natural mechanism of the eco-biological system through a multi-level model [4]. Our study aimed to understand the relationship among yield, yield components, and environmental factors using a hierarchical Bayesian model, and to determine the potential limiting factors for soybean yield production.

2 Materials and methods

2.1 Data collections

Soybean yield data were collected from eight farmer's fields in Kaizu City, Gifu Prefecture, Japan, between 2018 and 2023. Those eight fields were classified into seven soybean environments, defined as the combination of location and year, as two fields in 2019 were taken in an adjacent location. The fields followed a crop rotation system that included paddy rice, winter wheat, and soybean, with a two-year cultivation cycle. Soybean was grown from July to November each year, with sowing dates varying by field depending on soil moisture, although they generally occurred around the third week of July. All fields were planted with the soybean variety 'Fukuyutaka', commonly grown in western Japan. Plant samples were taken from 1-m² row areas during the maturity stage. In total, 513 plant samples were collected.

After harvest, yield components such as seed numbers and individual seed weight were measured. Soybean seeds were separated from the plant, and their weight was determined after being oven-dried at 70°C. Yield was calculated at 15% moisture content, ranging from 4.43 to 417.85 g/m^2.

Weather data were obtained from the Agro-Meteorological Grid Square Data provided by the National Agriculture and Food Research Organization (NARO), which covers observed and forecasted weather data across Japan since 1980. This study focused on precipitation as the primary weather factor influencing soybean growth. The impact of weather might vary depending on the growth stage, so the data were divided into two periods: the early reproductive stages (R2 to R5) and the seed-filling stages (R5 to three weeks post R5).

2.2 Hierarchical Bayesian model

Hierarchical Bayesian model was used to estimate the yield variability in response to the yield components including the seeds per area and seed weight, as explained by Egli [1]. Moreover, yield also expected to be influenced by different field environments, calculated as the random effects of the model. Specifically, the hierarchical Bayesian model was presented in the following equations:

First level model:

$$Yield_{i,j} = a_{0,j} + b_{1,j} * seeds\ per\ area_i + b_{2,j} * seed\ weight_i + e_{i,j} \quad (3)$$

Second level models:

$$a_{0,j} = a_0 + u_{0,j} \quad (4)$$
$$b_{1,j} = b_{10} + b_{11} * Precipitation_{1,j} + u_{1,j} \quad (5)$$
$$b_{2,j} = b_{20} + b_{21} * Precipitation_{2,j} + u_{2,j} \quad (6)$$

Error terms:

$$e_{i,j} \sim N(0, \sigma_e^2) \quad (7)$$
$$\begin{pmatrix} u_{0,j} \\ u_{1,j} \\ u_{2,j} \end{pmatrix} \sim MVN(0, \Sigma); \quad \Sigma = \begin{pmatrix} \sigma^2_{u_{0,j}} & \sigma_{u_{0,j},u_{1,j}} & \sigma_{u_{0j},u_{2,j}} \\ \sigma_{u_{0j},u_{1,j}} & \sigma^2_{u_{1,j}} & \sigma_{u_{1,j},u_{2,j}} \\ \sigma_{u_{0,j},u_{2,j}} & \sigma_{u_{1,j},u_{2,j}} & \sigma^2_{u_{2,j}} \end{pmatrix} \quad (8)$$

where $Yield_{i,j}$ is the observed yield (g/m^2) in the corresponding *j* environment, *seed number per area*$_i$ is the number of seed in a one square meter plot (seed/m^2), and *seed size*$_i$ is the average weight of the soybean seed (g/seed). Variable $a_{0,j}$ represent the variable intercept of the first level model, while $b_{1,j}$ and $b_{2,j}$ correspond to the variable slopes of each yield components in response to the environments as the random effects, which then explained by the second level model through Equation X, Y, and Z, respectively.

The second level model parameters include a_0, b_{10}, and b_{20} representing intercept and slopes in a typical environment setup, respectively. The second level model also contain parameters corresponding to the environmental factor, including $Precipitation_{1,j}$ and $Precipitation_{2,j}$, that are the cumulative precipitation during the early reproductive stages and seed filling stages, respectively, in each *j* environments.

The first level model residual error is assumed to be $e_{i,j} = (0, \sigma_e^2)$. The random effects of second level models ($u_{0,j}$, $u_{1,j}$, and $u_{2,j}$) are assumed to be correlated in a multivariate normal distribution with zero mean and $\sum$ covariance matrix, as written in Equation 8.

In the estimation of hierarchical Bayesian model, Markov chain Monte Carlo chains was used to calculate all parameters and saving 6,000 iterations for posterior inference. All variables except yield were standardized prior to the analysis. We specified weakly informative priors for all parameters and monitored the R-hat values to ensure chain convergence. The analysis was conducted using the 'rstan' package in R [5].

3 Results

Figure 1 illustrates the posterior of predicted soybean yield across all environments, calculated from the hierarchical Bayesian model. Specifically, in mean or aggregated condition (Fig 1a and b), the soybean yield was 171.28 g/m^2, which can be interpreted as the yield when the seeds per area and seed weight are at average values across all environments (524 seeds/m^2 and 31.94 g/seed, respectively). Figure 1a further shows that the yield in each environment (i.e., 1 to 7) had quite similar trends when the seeds per area and seed weight

were at average levels. However, Fig. 1b demonstrates that the estimated yield in each environment was largely affected by differences in the seeds per area and seed weight in each corresponding environment. The average number of seeds in each environment varied, with the lowest in Environment 2 (280 seeds/m^2) and the highest in Environment 6 (866 seeds/m^2), while the seed weight was quite similar across the environments (26.82 – 34.36 g/seed, full data not shown).

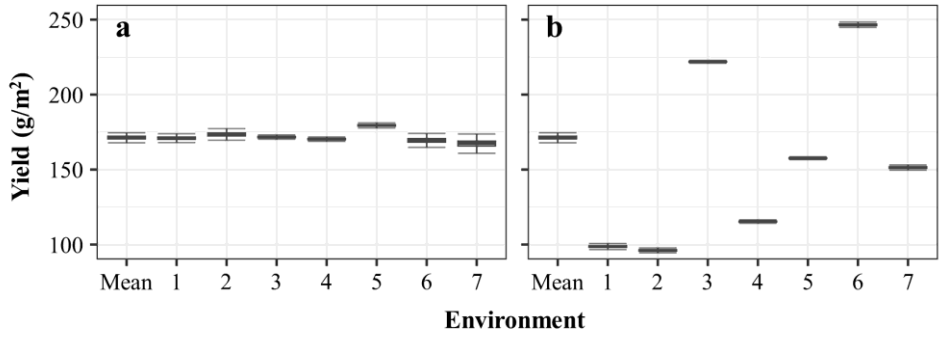

Fig. 1. Predicted soybean yield across environments based on the hierarchical Bayesian model, considering: a) yield in all environments when the seeds per area and seed weight are at typical levels, and b) yield in all environments when seeds per area and seed weight are in each corresponding environment levels. Boxplots indicating mean and 50% posterior highest density intervals (HDI) and the whiskers represent 95% posterior HDI.

Table 1 presents the posterior inference of each estimated parameter of the hierarchical Bayesian model. The intercept (a_0) has a small variance with a 95% highest density interval (HDI) ranging from 168.65 to 173.76 g/m^2, as depicted in Fig. 1a. The slope b_{10} has a high value with a mean of 90.35, while the slope b_{20} has a mean value of 12.28. The hierarchical Bayesian model was computed using standardized variables, so the magnitude of the slopes reflects the importance of the factors on the yield. Consequently, the results indicate that the number of seeds per square meter has a significantly higher influence on the yield.

Furthermore, Table 1 also illustrates the impact of precipitation during different growth stages, represented by the b_{11} and b_{21} parameters. The b_{11} parameter corresponds to the cumulative precipitation during the R2 to R5 growth stages, which is hypothesized to influence the seeds per area. The model estimates that b_{11} has a mean value of 5.35 with a 95% highest density interval (HDI) ranging from 3.21 to 8.15. All positive values in the 95%HDI range of b_{11} parameters indicate that cumulative precipitation had strong probability to aarea, thereby increasing the potential for yield production. On the other hand, the b_{21} parameter has a mean value of 0.65, but the 95% HDI ranges from -2.98 to 4.28, indicating that the effect of precipitation on the seed-filling stage had less probability to improve yield through increasing seed weight. $\sigma^2_{u_{2,j}}$ has the greatest magnitude, indicating that the potential error caused by random effects of the environments on seed weight can lead to the highest uncertainty in yield potential.

Table 1. Posterior summary (mean and 95% highest posterior density interval [HDI]) on all estimated parameters for the hierarchical Bayesian model.

Parameters	Description	Mean	95% HDI*	
First level model				
a_0	Intercept at a typical condition	171.28	168.65	173.76
b_{10}	Slope of seeds per area at a typical condition	90.35	87.77	92.83
b_{20}	Slope of seed weight at a typical condition	12.28	8.91	15.87
Second level model				
b_{11}	Change in b_{10} per unit increase of cumulative precipitation in early reproductive stage	5.35	3.21	8.15
b_{21}	Change in b_{20} per unit increase of cumulative precipitation in seed-filling stage	0.65	-2.98	4.28
Error terms				
σ_e^2	Variance of the model residual errors ($e_{i,j}$)	4.19	3.93	4.46
$\sigma_{u_{0,j}}^2$	Variance of the environmental effect on intercept ($u_{0,j}$)	2.77	1.88	4.11
$\sigma_{u_{1,j}}^2$	Variance of the environmental effect on b_{10} slope ($u_{1,j}$)	2.89	1.65	4.32
$\sigma_{u_{2,j}}^2$	Variance of the environmental effect on b_{20} slope ($u_{2,j}$)	4.74	3.26	6.39

HDI: Highest posterior density interval

4 Discussion

In this study, the hierarchical Bayesian model was utilized to understand the factors influencing soybean yield from the yield components and the environmental variables. The hierarchical Bayesian model could provide a flexible framework to account for uncertainties at multiple levels, incorporating both fixed effects, including yield, yield components, and precipitation variables, and also from the random effects of the environment. This allows for the estimation of the relationships among soybean yield, yield components, and environmental factors.

The study results indicate that yield components, particularly seed number per area, have a significant impact on soybean yield production. This finding is in accordance with previous literature [6,7] that soybean yield was determined around the early reproduction stages when the final seed number was fixed. Moreover, Xu et al [8] also adding that seed number per areas becoming the limiting factors of yield improvement in medium- to high-yield soybean fields. This study also provide evidence that environmental factors could greatly affect the seeds per area, which ultimately impacts the yield. The posterior estimates of b_{11} parameter (Table 1) underscored the critical role of precipitation during the early reproductive stages in influencing seeds per area. Sufficient water availability during late vegetative stages and early reproduction stages could assist maximum biomass accumulation [6] to support the initiation of pods and seeds. Previous study from MacMillan and Gulden [9] also found that yield components per unit area, especially pod and seed numbers, were strongly reduced due to dry condition. Therefore, providing water irrigation during early reproductive growing stages according to precipitation condition to ensure sufficient water availability may become a potential mitigation to avoid reduction of yield production due to low seeds per area, as also suggested by Ashley and Ethridge [10].

Seed weights also had a positive impact on yield production but with a smaller degree compared to the number of seed per square area, meaning that the decreasing of seed weight could potentially also decrease the yield potential. Xu et al [8] provide the evidence that seed weight become the major limiting factor for low-yielding soybean fields, however, in our study, the effect of seed weight was less than seed number per areas. Previous literature [9] found that seed weight has a small correlation with yield, as there are compensatory effects between yield components, especially seed weight and seed numbers. This is also proved by our study, where Environment 6 had the highest number of seed numbers, but the average seed weights are considered one of the lowest of all environments. Moreover, our study results also can be interpreted that soybean yield is mainly considered with sink-limited crops, meaning that the assimilation from photosynthesis during seed filling stages exceeds the demand from the established seed. This is consistent with the previous studies reporting that soybean is more considered as sink-limited crop than source-limited one [7,11]. The impact of environmental factors, including cumulative precipitation in the seed-filling stages, could affect water availability during pod and seed filling, which is known to reduce soybean seed weight if the water is insufficient [12].

5 Conclusion

In conclusion, our study demonstrated that hierarchical Bayesian model is suitable for understanding the complex relationship between soybean yield, yield components, and environmental factors. The seed number per areas become the main drivers of soybean yield, which could potentially become limiting factor to the yield improvement, with precipitation during early reproductive stages significantly affecting seed number. Meanwhile, seed weight also contributed to yield, but its impact was much smaller compared to seed number. Therefore, ensuring sufficient water availability during early reproduction stages through irrigation to maximizing seeds per area may become one of the potential ways to improve yield production.

References

1. D.B. Egli, Seed biology and yield of grain crops, (CABI Publisher, UK, 2017), ISBN 9781780647708.
2. L.N. Habibi, T. Matsui, T.S.T. Tanaka, Assessing the impact of soil clod and seeding rate on soybean seedling establishment and yield using a Bayesian approach, In Proceedings of the #OFE2023; 2024 (In press), December 5-7, 2023.
3. M.C.F. Wei, J.P. Molin, Soybean yield estimation and its components: A linear regression approach, Agriculture **10**, 1–13 (2020), doi:10.3390/agriculture10080348.
4. P. Poudel, N.M. Bello, D.A. Marburger, B.F. Carver, Y. Liang, P.D Alderman, Ecophysiological modeling of yield and yield components in winter wheat using hierarchical Bayesian analysis. Crop Science. **62**, 358–373 (2022), doi:10.1002/csc2.20652.
5. Stan Development Team, Stan Modeling Language Users Guide and Reference Manual (Version 2.35), 2024, Available online: https://mc-stan.org/docs/2_35/stan-users-guide-2_35.pdf
6. J.L. Rotundo, L. Borrás, J. De Bruin, P. Pedersen, Physiological strategies for seed number determination in soybean: Biomass accumulation, partitioning and seed set efficiency, Field Crop Research **135**, 58–66 (2012), doi:10.1016/j.fcr.2012.06.012.

7. L. Borrás, G.A. Slafer, M.E. Otegui, Seed dry weight response to source-sink manipulations in wheat, maize and soybean: A quantitative reappraisal, Field Crop Research **86**, 131–146 (2004), doi:10.1016/j.fcr.2003.08.002.
8. C. Xu, Y. He, S. Sun, W. Song, T. Wu, T. Han, C. Wu, Analysis of soybean yield formation differences across different production regions in China, Agronomy Journal **112**, 4195–4206 (2020), doi: 10.1002/agj2.20373.
9. K.P. MacMillan, R.H. Gulden, Effect of seeding date, environment and cultivar on soybean seed yield, yield components, and seed quality in the Northern Great Plains, Agronomy Journal **112**, 1666–1678 (2020), doi:10.1002/agj2.20185.
10. D.A. Ashley, W.J. Ethridge, Irrigation Effects on Vegetative and Reproductive Development of Three Soybean Cultivars, Agronomy Journal 70, 467-471 (1978)
11. D.B. Egli, Crop growth rate and the establishment of sink size: a comparison of maize and soybean, Journal of Crop Improvement **33**, 346–362 (2019), doi:10.1080/15427528.2019.1597797.
12. J.E. Board, C.S. Kahlon, Soybean Yield Formation: What controls it and how it can be improved. Soybean Physiology and Biochemistry, (INTECH Open Access Publisher, Rijeka, Croatia, 2011)