

Genomic prediction of green fraction dynamics in soybean using UAV observations

Yusuke Toda¹, Goshi Sasaki¹, Yoshihiro Omori¹, Yuji Yamasaki², Hirokazu Takahashi³,
Hideki Takanashi¹, Mai Tsuda⁴, Hiromi Kajiya-Kanegae⁵, Raul Lopez-Lozano⁶, Hisashi
Tsujiimoto², Akito Kaga⁷, Mikio Nakazono³, Toru Fujiwara¹, Frederic Baret⁶, Hiroyoshi Iwata¹

¹Graduate School of Agricultural and Life Sciences, The University of Tokyo, Tokyo, Japan

²Arid Land Research Center, Tottori University, Tottori, Japan

³Graduate School of Bioagricultural Sciences, Nagoya University, Nagoya, Japan

⁴Tsukuba-Plant Innovation Research Center (T-PIRC), University of Tsukuba, Tsukuba, Japan

⁵Research Center for Agricultural Information Technology, National Agriculture and Food Research
Organization (NARO), Tokyo, Japan

⁶Joint Research Unit of Mediterranean Environment and Modelling of Agroecosystems, National
Research Institute for Agriculture, Food and Environment (INRAe), Avignon, France

⁷Institute of Crop Science, NARO, Tsukuba, Japan

* Correspondence:

Hiroyoshi Iwata

aiwata@mail.ecc.u-tokyo.ac.jp

Keywords: soybean¹, unmanned aerial vehicle², remote sensing³, drought⁴, green fraction⁵,
dynamics model⁶, genomic prediction⁷, genomic selection⁸.

Abstract

With the widespread use of high-throughput phenotyping systems, growth process data are expected to become more easily available. By applying genomic prediction to growth data, it will be possible to predict the growth of untested genotypes. Predicting the growth process will be useful for crop breeding, as variability in the growth process has a significant impact on the management of plant cultivation. However, the integration of growth modeling and genomic prediction has yet to be studied in depth. In this study, we implemented new prediction models to propose a novel growth prediction scheme. Phenotype data of 198 soybean germplasm genotypes were acquired for three years in experimental fields in Tottori, Japan. The longitudinal changes in the green fractions were measured using UAV remote sensing. Then, a dynamic model was fitted to the green fraction to extract the dynamic characteristics of the green fraction as five parameters. Using the estimated growth parameters, we developed models for genomic prediction of the growth process and tested whether the inclusion of the dynamic model contributed to better prediction of growth. Our proposed models consist of two steps: first, predicting the parameters of the dynamics model with genomic prediction, and then substituting the predicted values for the parameters of the dynamics model. By evaluating the heritability of the growth parameters, the dynamic model was able to effectively extract genetic diversity in the growth characteristics of the green fraction. In addition, the proposed prediction model showed higher prediction accuracy than conventional genomic prediction models, especially when the future growth of the test population is a prediction target given the observed values in the first half of growth as training data. This indicates that our model was able to successfully combine information from the

early growth period with phenotypic data from the training population for prediction. This prediction method could be applied to selection at an early growth stage in crop breeding, and could reduce the cost and time of field trials.

1 Introduction

Genetic mechanisms of growth processes have become a crucial topic in plant breeding. The genetic dissection of the formation process of target traits will provide profound insights into its mechanism, which will lead to efficient genetic improvement. This understanding is important in genomic selection (GS), where breeders skip field trials and select promising candidates based on the predicted breeding value provided by genomic prediction (GP) (Meuwissen et al., 2001). Most GP studies on crops have focused on traits at harvest, such as yield and quality (Krishnappa et al., 2021). If GP can predict phenotypic changes during the growth process, breeders can accurately determine the behavior of the genotypes obtained by GS and select the appropriate candidates. Further, growth prediction from the early period is likely to reduce the cost of field trials by shortening the period of observation.

However, the measurements required to derive growth traits are cost and labor intensive, representing a major bottleneck for genetic dissection, which requires the characterization of many genotypes. Due to the rapid development of sensing technologies in recent years, high-throughput phenotyping has become available for plant breeding, and the measurement of growth traits is becoming more practical. An accurate and detailed acquisition of growth processes through high-throughput measurements is expected to lead to improved genetic gains in plant breeding (Furbank and Tester, 2011; Cabrera-Bosquet et al., 2012; Araus and Cairns, 2014). For example, an automated phenotyping platform for the monitoring of three-dimensional plant growth in a greenhouse has enabled the genetic dissection of growth processes using a dynamic model (Campbell et al., 2018). In a field experiment, high-throughput phenotyping using unmanned aerial vehicles (UAVs) (Yang et al., 2017) and tractors (White et al., 2012) was used to measure plant growth. Among growth traits, the leaf area index (LAI) is often investigated because it is accessible from high-throughput phenotyping (Verger et al., 2014; Liu et al., 2017), while governing the main growth processes and being sensitive to the environment. Until recently, however, these techniques were mainly used for crop management, and their applications to genetic dissection remain limited (Blancon et al., 2019).

Several methods have been proposed for the analysis of plant growth. One commonly used method involves fitting a growth model, such as Gompertz (Winsor, 1932) and logistic (Nelder, 1961), to the data and using the model parameters to quantify the dynamic pattern. This method can be applied to various types of dynamic measurements. Several methods of quantitative genetics, such as quantitative trait loci analysis (Ma et al., 2002; Wu et al., 2002) and genome-wide association studies (Das et al., 2011; Crispim et al., 2015), have been applied to discover possible associations with growth model parameters. Growth models have also been used as a flexible tool to analyze various factors, such as the effect of selection in breeding (Piles et al., 2003) and the relationship among traits (Onogi et al., 2019). However, its application to GP has not been discussed in previous studies.

In this study, a method integrating a model of growth dynamics and GP was proposed and applied to investigate the growth of soybeans. We focused on the green fraction (GF) to model its dynamics. GF is defined as the fraction of green pixels in an image taken from the nadir. This trait is a proxy for LAI and can be easily measured from UAV observations. The GF dynamics of soybean germplasm accessions were described using the parameters of a dedicated model. Genetic variations in GF dynamics were quantified by decomposing the model parameters into genetic and residual effects using

mixed models. Finally, the GP model is applied to predict the parameters of the GF dynamics model under a range of scenarios to illustrate the potential of the proposed method.

2 Material and methods

2.1 Field trials

Soybean accessions registered in the National Agriculture and Food Research Organization Genebank (https://www.gene.affrc.go.jp/index_en.php) were used. From 2017 to 2019, the field trial was conducted in an experimental field with sandy soil at the Arid Land Research Center, Tottori University (35°32' N lat, 134°12' E long, 14 m above sea level). A total of 198 accessions from GenBank between 2018 and 2019 were used, with 186 out of 198 accessions used in 2017. Each plot consisted of four plants. The distances between two rows, two plots, and two individuals were 50 cm, 80 cm, and 20 cm, respectively (Figure 1d). Sowing was performed at the beginning of July, followed by thinning after two weeks. Fertilizer (15, 6.0, 20, 11, and 7.0 g m⁻² of N, P, K, Mg, and Ca, respectively) was applied to the field before sowing.

Two watering treatment levels, control (C) and drought (D), were used to evaluate the genetic variations in the responses to water stress. White mulching sheets (Tyvek, Dupond, US) were laid to prevent rainwater infiltration (Figure 1), and pipes were installed under the sheets to irrigate the field. Irrigation was applied for daily for 5 h (7:00–9:00, 12:00–14:00, 16:00–17:00), starting the day after the thinning in treatment C, while no water was brought in treatment D. In the following, an abbreviation for denoting a specific combination of the level of the treatment and the year of the experiment is used; for example, treatment C in 2017 is abbreviated to “2017-C.”

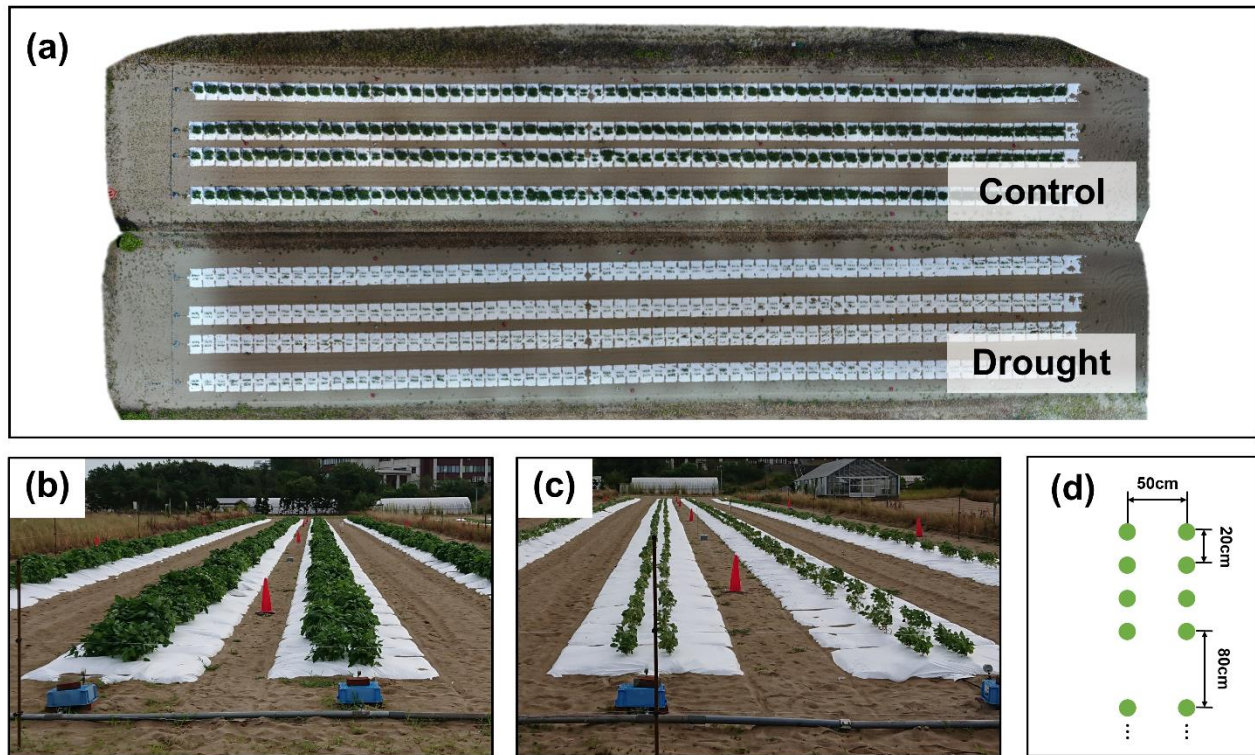


Figure 1. The field experiment. (a) An ortho-mosaic image of the field, obtained on August 25, 2018. The ortho-mosaic images were created for each treatment. (b, c) ground level images of treatments C and D. (d) Planting pattern of plots made of 2 rows of 4 plants (green dots) and separated by 80 cm.

2.2 Remote sensing and image analysis

UAV flights started after thinning and were performed 16–35 times during the cultivation period. A consumer drone (DJI Phantom 4 Advanced, China) was used for image collection. Images consisted of RGB layers and 3648×4864 pixels, captured with an automated focus and white balance. The UAV flew 12–14 m above the ground and captured images every two seconds with an autofocus function. A single UAV flight took approximately 15 min and collected 500 to 600 images, which was repeated twice to cover the entire field.

Ortho-mosaic images were constructed using Pix4Dmapper (Pix4D, Switzerland). The images of individual plots were then segmented from the ortho-mosaic image based on the geolocation of their corners. NDVI-based thresholds were applied to segment the canopy and ground regions from the images of the individual plots. Finally, the GF of each plot was estimated as the ratio of the green pixels to the total number of pixels in the plot. The image analysis process was implemented in Python 2.7 (<https://www.python.org>). For data in 2019, a similar procedure was used by Hiphen Inc. (<https://www.hiphen-plant.com>).

2.3 GF dynamics modeling

The GF derived from the UAV on each day, GF_d , was first converted into the corresponding leaf area index (LAI_d), following (Soltani and Sinclair, 2012) the exponential model:

$$GF_d = 1 - \exp(-k LAI_d) \quad (1)$$

where $k = 0.5$, and is the extinction coefficient commonly used for soybean. The model proposed by (Koetz et al., 2005) to describe the dynamics of LAI was fitted to the time series of GF to estimate the growth pattern of each plot with five parameters (Figure 2):

$$LAI_d = LAI_{amp} \left\{ \frac{1}{1 + \exp(-r_g(T_d - T_g))} - \exp(r_s(T_d - T_s)) \right\} \quad (2)$$

The first term in the parenthesis represents logistic growth, and the second term represents exponential senescence, where T_d is a growth degree day on day d . The growth degree day is a typical development scale corresponding to the cumulative daily mean temperature above the base temperature set to 8°C (Soltani and Sinclair, 2012). LAI_{amp} is the maximum value of the LAI reached, r_g is the maximum LAI growth rate, r_s is the senescence rate, T_g is the growth degree day when the LAI growth rate is maximum, and T_s is the growth degree day when LAI becomes zero. The five parameters, LAI_{amp} , r_g , r_s , T_g , and T_s , were estimated for each plot. However, fitting the dynamics model to each plot was difficult because the GF in the drought treatment (D) was so small that the noise disturbed the search for the best parameter sets. Therefore, the estimation was conducted in two steps:

(1) *First parameter estimates*. In this step, all of the parameters except LAI_{amp} were assumed to be genotype dependent, that is, the same for treatments C and D. The optimal values of T_g and T_s were found with a grid search in the range $300 < T_g < 1200$ and $1400 < T_s < 3000$ on seven points evenly distributed in the range. At the same time, the optimal values of the other parameters were estimated using the Nelder-Mead method. The cost function to be minimized in the Nelder-Mead method was computed as follows:

$$\sum_d \sum_i \frac{(y_{i,d} - \hat{y}_{i,d})^2}{\bar{y}_d} \quad (3),$$

where $y_{i,d}$ is the GF of plot i on day d , $\hat{y}_{i,d}$ is the estimated value of GF with the dynamics model, and $\bar{y}_d$ is the mean value of the GF on day d . The normalization by $\bar{y}_d$ in Equation 3 accounts for the measurement noise, which is roughly proportional to the mean value.

(2) *Fine tuning the parameter estimates.* The parameter estimation was conducted independently for each plot. The optimal values obtained in the previous step for T_g and T_s were the center of the grid search with a narrower range ($T_g - 200 < T_g < T_g + 200$ and $T_s - 400 < T_s < T_s + 400$). The other parameters were estimated using the Nelder-Mead method, using the estimated values in Step 1 as the initial values.

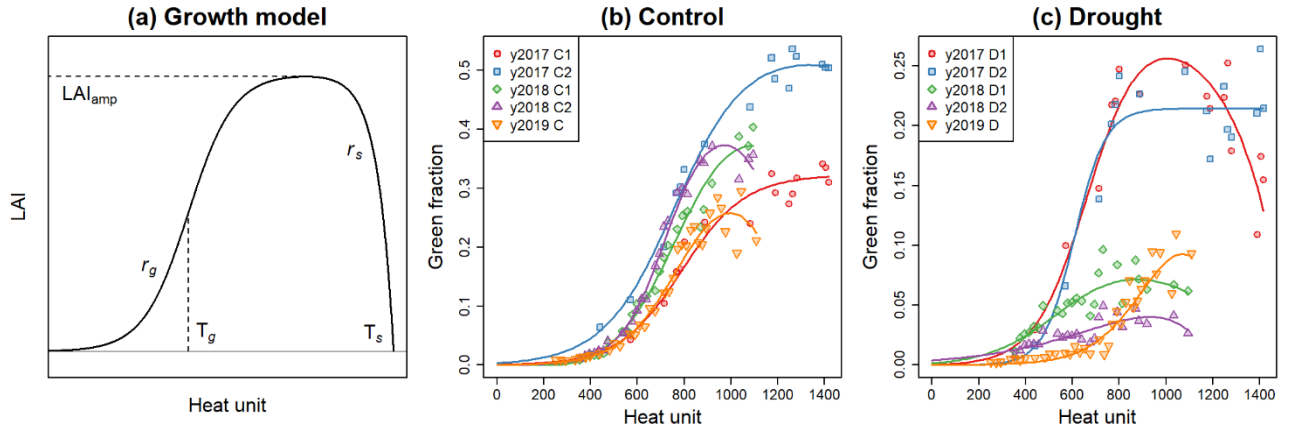


Figure 2. (a) An example of LAI dynamics described with the model (Equation 2). The parameters of the LAI dynamics model are displayed to visualize their roles. (b, c) GF dynamics of the genotype “Enrei”, with the model fitted to the data. Each plot is displayed separately for the control (b) and drought (c) treatments.

2.4 Estimation of genotypic values

Genotypic values of the GF and LAI dynamic parameters were estimated for use in the GP. The following mixed model was fitted for each combination of a trait (GF or LAI dynamics parameter) and a treatment (C or D):

$$\mathbf{y} = \mu \mathbf{1} + \mathbf{L}\boldsymbol{\beta} + \mathbf{W}\mathbf{s} + \mathbf{e} \quad (4)$$

where $\mathbf{y}$ is a vector of the phenotypic values; μ is the mean; $\boldsymbol{\beta}$ is a vector of block effects representing differences between replications; $\mathbf{s}$ is a vector of genotypic values that follows $N(\mathbf{s} | \mathbf{0}, \sigma_s^2 \mathbf{I})$; σ_s^2 is the genotypic variance; $\mathbf{e}$ is a vector of residuals that follows $N(\mathbf{e} | \mathbf{0}, \sigma_e^2 \mathbf{I})$; σ_e^2 is the residual variance; $\mathbf{1}$ is a vector in which all the elements are one; $\mathbf{I}$ is an identity matrix; and $\mathbf{L}$ and $\mathbf{W}$ are design matrices. The genotypic value ($\mathbf{g}$) was calculated as follows:

$$\mathbf{g} = \mu \mathbf{1} + \mathbf{s} \quad (5)$$

The R package lme4 (ver. 1.1-20) was used to solve Equation 4. For the GF, the genotypic value estimation was applied separately for each flight date.

2.5 Genomic relationship matrix and genetic analysis

The whole-genome sequencing data of all 198 accessions were available and used to estimate the genomic relationship matrix (Kajiya-Kanegae et al., 2021). Only the biallelic sites in all accessions with a minor allele frequency (MAF) ≥ 0.025 , missing rate < 0.05 , and linkage disequilibrium < 0.95 were extracted, and the imputation of missing genotypes was applied. Genotyping data identified 425,858 SNPs. Genotypes for individual alleles were represented as -1 (homozygous for the reference allele), 1 (homozygous for the alternative allele), or 0 (heterozygous for the reference and alternative alleles). The genomic relationship matrix $\mathbf{G}$ was estimated as $\mathbf{G} = \mathbf{X}\mathbf{X}^T / c$, where $\mathbf{X}$ is an $n \times m$ marker genotype matrix (n and m are the numbers of lines and markers, respectively), and c is the normalization constant (Endelman and Jannink, 2012). Genetic heritability was estimated for all traits using the genomic best linear unbiased prediction (G-BLUP) model:

$$\mathbf{g} = m\mathbf{1} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon} \quad (6)$$

where $\mathbf{g}$ is a vector of genotypic values estimated using Equations 4 and 5, m is the mean, $\mathbf{u}$ is a vector of random genetic effects that follows $N(\mathbf{u} \mid \mathbf{0}, \sigma_u^2 \mathbf{G})$, $\boldsymbol{\varepsilon}$ is a vector of residuals that follows $N(\boldsymbol{\varepsilon} \mid \mathbf{0}, \sigma_\varepsilon^2 \mathbf{I})$, σ_u^2 and σ_ε^2 are the genetic and residual variances, respectively, and $\mathbf{Z}$ is a design matrix. The R package rrBLUP (ver. 4.6) (Endelman, 2011) was used to solve Equation 6. After solving the mixed model, the genomic heritability was estimated as $h^2 = \sigma_u^2 / (\sigma_u^2 + \sigma_\varepsilon^2)$.

2.6 Prediction of GF dynamics

We investigated three cross-validation schemes for the four different prediction models. The cross-validation schemes and prediction models are detailed as follows: The correlation coefficient between the genotypic values ($\mathbf{g}$) and their predicted values ($\mathbf{u}$) of the GF was used to evaluate the prediction accuracy.

2.6.1 Cross-validation schemes

Cross-validation was repeated three times for the combination of a cross-validation scheme and a prediction model.

(1) *Cross-validation of genotypes (CV1)*. CV1 corresponded to the prediction of LAI dynamics for untested genotypes. Data from a subset of genotypes in any treatment or year were excluded from the training data (Figure 3a). The prediction model built using the training dataset was evaluated for the left-out genotypes. Ten-fold cross-validation was used to randomly select left-out genotypes.

(2) *Cross-validation over combination of genotype and environment (CV2)*. The combination of a treatment and a year was considered as an environment: there were a total of six environments (two treatments $\times$ three years). Here, the predicted LAI dynamics were evaluated for genotypes and environments left out from the training dataset. A ten-fold cross-validation was used for combinations of genotypes and environments (Figure 3a).

(3) *Cross-validation with a focus on the late growth period (CV3)*. The growth cycle was split into early and late growth, with an equal number of observations for the two periods. CV3 was similar to CV2, but the LAI dynamics were evaluated over the late period, while the training was achieved over the early growth period (Figure 3a).

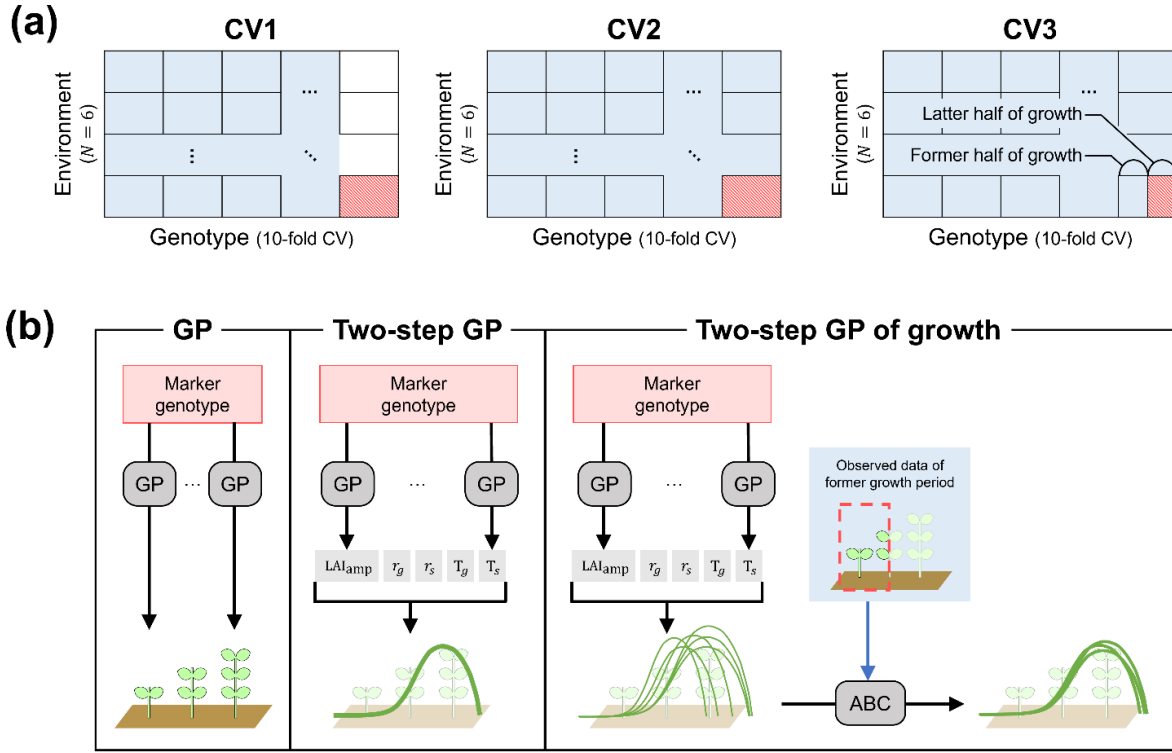


Figure 3. Cross-validation schemes and prediction models. (a) Cross-validation schemes (CV1, CV2, and CV3). Training and test data in cross-validation are expressed as blue and red cells, respectively. (b) Structures of prediction models (GP, TGP, and TGP of growth). Structures of the other prediction models with multivariate GP (MGP, TMGP, TMGPG) can be understood by replacing GPs in the figure to MGPs.

2.6.2 Prediction models in CV1 and CV2

In CV1 and CV2, four prediction models were compared: genomic prediction (GP), two-step GP (TGP), multivariate GP (MGP), and two-step multivariate GP (TMGP).

GP is the most standard model, expressed as shown in Equation 6, and applied to the GF on each day in the training data (Figure 3b). Then, the random genetic values $\mathbf{g}$ of the left-out genotypes were used as the predicted values.

The TGP consisted of two steps (Figure 3b). First, the same model as GP was applied to the LAI dynamics model parameters. Then, the GFs of the left-out genotypes on each day were calculated using the predicted parameters (Equations 1 and 2).

MGP is an extension of the GP, which simultaneously predicts several traits (Calus and Veerkamp, 2011; Jia and Jannink, 2012). The model is expected to enhance the accuracy of genomic prediction via genetic correlations among traits. This model can be expressed as follows:

$$\begin{pmatrix} \mathbf{g}_1 \\ \vdots \\ \mathbf{g}_J \end{pmatrix} = \begin{pmatrix} m_1 \mathbf{1} \\ \vdots \\ m_J \mathbf{1} \end{pmatrix} + \begin{pmatrix} \mathbf{Z}_1 & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{Z}_J \end{pmatrix} \begin{pmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_J \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_J \end{pmatrix} \quad (7)$$

where J is the number of variates in the model, $\mathbf{g}_j$, $\mathbf{u}_j$, and $\boldsymbol{\varepsilon}_j$ are vectors of genotypic values, random genetic effects, and residuals of variate j , respectively, and m_j is the mean of variate j . Assumptions for the random effects were included, in which $\mathbf{u}_{\text{all}} = (\mathbf{u}_1^T, \dots, \mathbf{u}_J^T)^T$ follows $N(\mathbf{u}_{\text{all}} | \mathbf{0}, \mathbf{K} \otimes \mathbf{G})$ and $\boldsymbol{\varepsilon}_{\text{all}} = (\boldsymbol{\varepsilon}_1^T, \dots, \boldsymbol{\varepsilon}_J^T)^T$ follows $N(\boldsymbol{\varepsilon}_{\text{all}} | \mathbf{0}, \mathbf{R} \otimes \mathbf{I})$. Here, $\mathbf{K}$ is a genomic variance-covariance matrix between the variates, and $\mathbf{R}$ is the residual variance-covariance matrix between the variates. The R package MTM (ver. 1.0.0) was used to solve Equation 7 based on the Markov chain Monte Carlo (MCMC) method.

TMGP consisted of two steps, that is, MGP of the LAI dynamics model parameters and the calculation of the GF using the predicted parameters.

MGP and TMGP were expected to improve the prediction accuracy compared to GP by exploiting phenotypic data from environments included only in the training dataset. However, since the GF was measured repeatedly in each environment, it was difficult to include all the phenotype data (152 measurements in total by adding up observation dates in all the environments). Thus, a strategy was applied where the training of prediction models was repeated for each observation date, and ten additional variates were selected from the whole data to support the prediction every time. In other words, the eleven variates included each time consisted of one target variate and ten supporting variates. The criterion for selecting supporting variates is based on heritability and correlation with the target variate. These two factors are essential for improving the prediction accuracy in MGP (Calus and Veerkamp, 2011). Top-10 observations of the following criterion were selected as supporting variables:

$$s(h^2) + s(|r|) \quad (8)$$

where $s(\cdot)$ is a scaling function that makes the mean and variance of an input vector zero and one, respectively; h^2 is the heritability; and r is the correlation coefficient with a target variate.

2.6.3 The prediction models in CV3

As in the other cross-validations, the performances of the four prediction models were compared in CV3. GP was the same as in CV2 because it only uses the data of the measurement day to be predicted for training. MGP was modified to better exploit the first half of the growth period used to train the model. Seven out of the ten supporting variates were selected using the selection criterion from Equation 8, the remaining three variates corresponded to the GF values for the latest three flights of the first growth period. The other two models with two-step structures, TGP and TMGP, were also modified to better exploit the training data for predicting the GF dynamics during the late growth period. They were called TGPG (TGP for growth) and TMGPG (TMGP for growth), respectively.

The TGPG included three steps (Figure 3b). The first two steps were the same as those of the TGP, where the LAI dynamics model parameters were predicted without using the data from the first half of the growth period. However, for TGPG, the distributions of the MCMC values of the LAI dynamics model parameters were used instead of the average value of the samples used in the TGP. As a result, 60,000 samples of the predicted GF dynamics were obtained for each genotype corresponding to the prior distribution when no GF measurements on the genotype were available. Then, the GF data from the first half of the growth period were exploited using the approximate Bayesian computation (ABC) method, and the 60 GF dynamics that minimize the Euclidian distance between the predicted GF dynamics and the actual GF observations were selected. Lastly, the mean values of the 60 samples were used as the predicted values. The modifications on TMGP to obtain TMGPG were the same as those applied to TGP to obtain TGPG.

3 Results

3.1 Dynamics of GF derived from UAV

The UAV observations transformed into GF values show typical dynamics of the several genotypes and environments investigated (Figure 4), showing large variations in the growth patterns. It is worth noting that the period covered by the flights was longer in 2017, with up to 80 days compared to 2018 and 2019, where the flights were stopped after 60 days. For each plot, the GF dynamics were relatively smooth, indicating a good temporal consistency of the GF values derived from the UAV observations. The ranking between genotypes is also generally consistent across growth development, which would indicate good chances to predict the late period from observations covering the early growth period. Drought treatment (D) always showed lower GF values than the control © treatment. However, the water stress experienced by treatment D varied across years, with 2018 being the most severe, and 2017 the mildest. The control treatment also showed differences between the three years: 2017 showed the best growth conditions, while 2019 showed the worst ones.

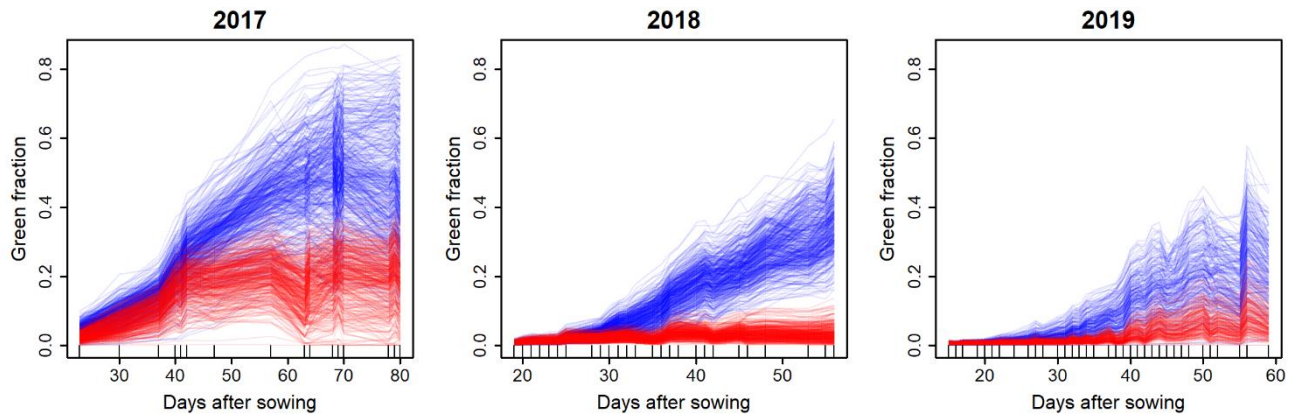


Figure 4. Dynamics of the GF as observed from the UAV. Treatments C and D are shown with blue and red lines, respectively, each line corresponding to a plot. The number of days after sowing is used as the x-axis.

The genomic heritability estimated for each year and treatment by fitting a mixed model to GF for each observation day showed a decrease until 40–50 days after sowing, and then increased with time (Figure 5a, 5b). However, some differences between years were observed, with a higher heritability in the early stages of 2018. The yearly patterns were also similar between the control (C) and drought (D) treatments, while the heritability in treatment C was systematically higher than that in treatment D, except in 2017 for the late UAV flights.

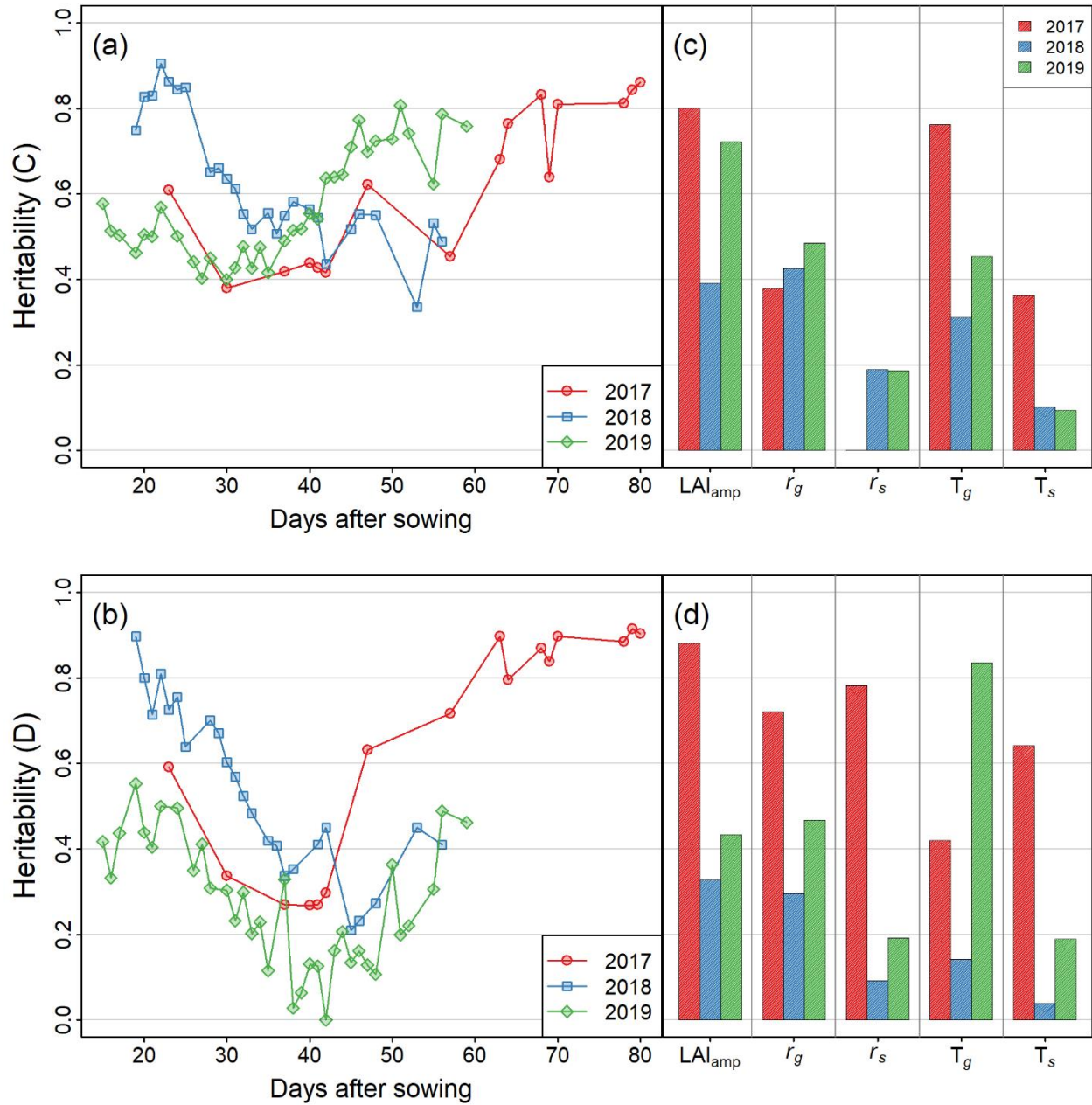


Figure 5. (a, b) Heritability of the GF. Results of treatment C and D are shown in panels (a) and (b), respectively. Red, blue, and green lines indicate the values in 2017, 2018, and 2019, respectively. (c, d) Heritability of the LAI dynamics model parameters. Results of treatment C and D are shown in panels (c) and (d), respectively.

3.2 Growth parameter estimation

The dynamics model fitted a wide range of GF growth patterns in both treatments and all years (Figure 2). The root mean squared errors (RMSE) of growth model fitting of GF on 25 days after sowing were 0.0060 and 0.0057 in treatment C and D, respectively. RMSE reached 0.022 and 0.016 in two treatments on 50 days after sowing, because growth of canopy increased the measurement noise of GF.

The distribution of the estimated growth parameters varied among years and treatments (Figure 6). Fundamentally, the parameters related to period of growth (r_g , T_g , and LAI_{amp}) showed a tendency wherein the values of the parameters became smaller when the plants were subjected to drought stress. For the parameters related to period of senescence (r_s , T_s), the results were not reliable due to the lack of observation of senescence, except in 2017.

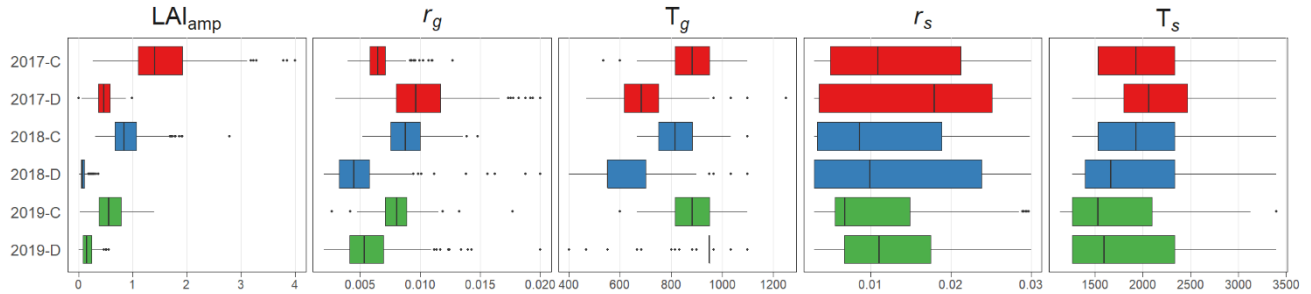


Figure 6. Boxplot of estimated growth parameters.

The genomic heritability of the growth parameters (Figure 5c, 5d) varied among treatments and years. The heritabilities of LAI_{amp} , T_g , and r_g were relatively high, reaching 0.8 in the highest cases. The other two parameters, r_s and T_s , showed lower heritability, ranging between 0.1 and 0.5. These parameters characterize the late development of the canopy that was not well covered by UAV flights, except in 2017. This explains why heritability was highest in 2017. Particularly, heritability was high in 2017-D, with all values exceeding 0.4. On the other hand, heritability was lowest in 2018 when summer heat stress was severe. All the parameters showed heritability lower than 0.4 in 2018-D. The heritability of the parameters of the model is generally lower than that of the GF for each UAV flight (Figure 5), except for LAI_{amp} , which generally shows a higher heritability, except in 2018.

3.3 Prediction of growth patterns

In CV1, the prediction accuracy of TGP and MGP was similar to that of GP (Figure 7). A significant improvement in prediction accuracy in TGP and MGP was observed in 2019-D, where the accuracy of GP was very low in the latter half of the growth period (Figure S1). The accuracy of MGP was slightly higher than that of TGP. The accuracy of TMGP differed among environments; it was lower than that of GP when predicting the GF in 2018, while it was higher than the accuracy of MGP when predicting the GF in the latter half of the growth period in 2019.



Figure 7. Comparison of accuracy of the prediction models. The correlation coefficients between genotypic values and their predicted values are plotted. The results of CV1, CV2, and CV3 are shown in three panels.

The predicted values of GP and TGP in CV2 were equal to those in CV1 because they did not utilize data in environments other than their targets. Thus, for CV2, the focus will be on the accuracy of the MGP and TMGP. The accuracy of MGP was higher in CV2 than in CV1 and was significantly higher than that of TGP (Figure 7, S2). The accuracy of TMGP was lower in CV2 than in CV1 in 2018, while it was higher in CV2 than in CV1 in the other years. Comparing MGP and TMGP, the accuracy of MGP was higher in 2018 and the former half of the growth period, while that of TMGP was higher in other environments.

In CV3, the prediction accuracy of TGPG and TMGPG was higher than that of the other models in all the environments (Figure 7) over the entire growth period (Figure S3). The correlation coefficients between the predicted values of TMGPG and the genotypic values were higher than 0.6 in most cases. The accuracy of MGP was higher than that of GP, but lower than that of TGPG and TMGPG. As in CV2, the predicted values with GP and TGP in CV3 were the same as in CV1.

4 Discussion

4.1 UAV-RS as a tool to evaluate growth patterns

This study showed that the UAV measurements of GF could be used to assess the genetic diversity of soybean growth patterns. A U-shaped longitudinal pattern was observed in the heritability of GF in all environments (Figure 5a, 5b). The U-shaped heritability patterns can be explained in three steps. In the early stage of growth, the GF seemed to be determined by factors regarding initial growth speed, such as radiation use efficiency, which results in high heritability of the GF. At approximately 25 days after sowing, several additional factors, such as growth phenology and plant structure, related to phenological development started to affect the GF, which decreased the heritability. Then, the saturation of the GF occurred around 45–60 days after sowing. During this period, confounding factors related to the differences in phenological development weakened, leading to increased heritability.

When we evaluate crop dynamics using remote sensing techniques, the LAI is often used as the main trait of interest. To apply remote sensing techniques to breeding populations, differences in plant styles should be carefully considered when deriving LAI from the images (Blancon et al., 2019). In this study, we focused on GF rather than LAI as the target trait to model its dynamics, due to the lack of data on plant style. Applying our method to LAI would offer the advantage of being closer to crop growth mechanisms, such as photosynthesis. Nevertheless, this study has shown that GF is a useful trait to describe the growth of soybean germplasm. In genomic prediction, the inclusion of canopy area, which is proportional to GF, has been reported to improve the prediction of biomass in soybean (Toda et al., 2021). Thus, GF can be considered a useful index of genetic variation in plant growth.

4.2 Fitting a dynamics models

The dynamics model was flexible enough to represent various growth patterns in different environments (Figure 2). By fitting the model, the GF time series is represented by five parameters. The distribution of growth parameters reflected the effects of drought stress on growth (Figure 6). The decrease in r_g and T_g in treatment D indicates that the speed and duration of growth were strongly suppressed under drought stress. As a result, the maximum value of the plant canopy was significantly reduced, which was expressed in the reduction of LAI_{amp} .

Genetic analysis of the parameters showed that the heritability of LAI_{amp} was high (Figure 5c, 5d). This result is related to the high heritability of GF in the later stages of growth. T_g , which describes the stage-shift timing of the GF, also showed high heritability in 2017-C and 2019-D.

For the senescence stage, the heritability was low for T_s and r_s , except for 2017-D. Due to the long cultivation period in 2017 and early senescence in treatment D, model fitting of the senescence part was successful in that environment. The heritability of r_s was higher than T_s , which means that the change in GF in the senescence stage was mainly determined by its speed, r_s , rather than the timing of senescence, T_s . The senescence pattern could be evaluated more precisely in 2017 and 2018 by extending the observation period.

Several useful results were obtained by applying the dynamics model, and some problems were found to be improved. Because of the high heritability of GF in the early stages, the growth speed, r_g , was expected to be mainly determined by genetic factors. However, the heritability of r_g was low, except in 2017-D. The use of other dynamic models, such as the Gompertz (Winsor, 1932) curve, may improve the goodness of fit of a growth curve to the GF in the early growth stages. It was reported that the dynamics model that considers leaf appearance could explain the dynamics of green LAI (GLAI) (Blancon et al., 2019). Such structural models may also be candidates for alternative dynamic models.

Another possible improvement of the model is the inclusion of other environmental factors, such as soil moisture and drought stress. In the dynamics model, the effect of temperature on growth stages was considered. However, the inclusion of other factors may allow for improved fitting and simultaneous parameter estimation of multiple environments. For example, the low heritability in 2018 of the growth parameters was due to severe heat stress in the summer of 2018, which made the growth slower than usual years. As a result, the sigmoid pattern in growth was truncated at the end of the cultivation period (Figure 4). Other environmental factors will allow simultaneous parameter estimation in other environments, leading to stability in the estimated parameters.

4.3 Prediction of growth curves

In CV1, the accuracy was close between GP and TGP (Figure 7). This result suggests that the dynamics model used in TGP could extract sufficient genetic variations from phenotypic variations in the GF dynamics pattern to achieve the same predictive accuracy as the GP.

Models with multivariate GP yielded better accuracy than those with univariate GP; the accuracies of MGP and TMGP were higher than those of GP and TGP, respectively. High correlations among variates, a typical property of dynamic data, suggest that multivariate GP improves the prediction accuracy because MTG and TMGP can leverage the among-characteristics correlation. In the following, we focus on the comparison between MGP and TMGP.

In 2018, the accuracy of TMGP was lower than that of MGP in CV1 and CV2 (Figure 7) because of the low heritability of the LAI dynamics model parameters. However, the accuracy of TMGP was higher than that of MGP in 2019 for CV2. TMGP was better than MGP because of the higher heritability of growth parameters than GF in 2018. The extraction of genetic variance in growth patterns in 2019 was successful as LAI_{amp} in 2019-C and T_g in 2019-D, leading to improved prediction accuracy.

In CV3, the prediction accuracies of TGP and TMGP outperformed the other models (Figure 7). The higher prediction accuracy compared to MGP indicates that the former growth period's data could be effectively included in the model by specifying the growth curve's shape through the dynamics model. In most cases, the correlation coefficients between the predicted values of TMGP and the genotypic values exceeded 0.6, indicating that TMGP is robust to changes in the environment. The similar prediction accuracy of MGP and TMGP in 2017-D may be due to the lack of change in the GF in the second half of this environment's growth period. This approach to future prediction through dynamic models has potential applications for selection in early growth stages in crop breeding.

In this study, the dynamics model and GP/MGP were used separately in TGP/TMGP, but they could be integrated into one hierarchical model. Several reports have shown the effectiveness of hierarchical models in the analysis of dynamic traits (Onogi et al., 2019), quantitative trait loci analysis (Ma et al., 2002), and genome-wide association studies (Das et al., 2011; Crispim et al., 2015). The joint analysis is expected to make parameter estimation more robust. In this study, although two steps are required to estimate the growth parameters, joint estimation may simplify the estimation process further.

A random regression model is also known as a regression method of dynamic data with a mixed model structure, which was used in the GP of dynamic traits (Sun et al., 2017; Campbell et al., 2018). Although random regression cannot incorporate the growth curve structure like the dynamics model in this study, its formulation is simple. Although random regression is assumed to perform better than MGP in the prediction scheme of CV3, it is not as accurate as TMGPG.

4.4 Growth analysis on remote sensing data for plant breeding

Applying the dynamics model to crops monitored with UAVs allows us to capture the genetic variation in growth patterns. The combination of the dynamics model and genetic analysis was shown to be an efficient framework for analyzing our field experiments. It was able to predict future GF dynamics from observations covering only the early growth stages. This could contribute to reducing the cost of the field trials. This study suggests that data monitoring the experiment with UAVs and analyzing them using dynamics models and mixed models will benefit crop breeding.

In the near future, UAV-RS is expected to play an active role in plant breeding and provide growth trajectory data from multiple breeding programs. It will be possible for breeders and researchers to focus on new genotypes to select and develop new varieties suitable for the target environment. The integrated use of dynamic models and GP will be a useful method to effectively link growth process data with marker genotype data to improve genetic gain for genomic selection.

Funding

This study was supported by JST CREST (<https://www.jst.go.jp/kisoken/crest/en/index.html>) (grant no. JPMJCR16O2). The funder had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments

We are grateful to the technical staff of the Arid Land Research Center, Tottori University, and Izumi Higashida for the management of the field experiments. We would like to thank Yuji Sawada for being involved in the field experiments, Kosuke Hamazaki for curating whole-genome marker data, and Sawako Maruyama for curating the phenotype data.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

MT, NM, HT, KA, and HI were involved in funding acquisition. AK prepared 198 accessions of soybean genetic resources. YT, GS, YO, YY, HirT, HidT, MT, HT, AK, MN, TF, and HI conducted the field experiments. YT, GS, and HI conducted UAV remote sensing. HK prepared the genome-wide marker genotype data. YT, RL, and FB developed a dynamic model of the GF. YT analyzed the data and wrote the manuscript. HI supervised the study and edited the manuscript. All authors have reviewed and approved the final manuscript.

References

- Araus, J., and Cairns, J. E. (2014). Field high-throughput phenotyping: the new crop breeding frontier. *Trends in Plant Science* 19, 52–61. doi:10.1016/j.tplants.2013.09.008.
- Blanc, J., Dutartre, D., Tixier, M.-H., Weiss, M., Comar, A., Praud, S., et al. (2019). A High-Throughput Model-Assisted Method for Phenotyping Maize Green Leaf Area Index Dynamics Using Unmanned Aerial Vehicle Imagery. *Frontiers in Plant Science* 10, 685. doi:10.3389/fpls.2019.00685.
- Cabrera - Bosquet, L., Crossa, J., Zitzewitz, J., Serret, M., and Araus, J. (2012). High - throughput Phenotyping and Genomic Selection: The Frontiers of Crop Breeding Converge. *Journal of Integrative Plant Biology* 54. doi:10.1111/j.1744-7909.2012.01116.x.
- Calus, M. P., and Veerkamp, R. F. (2011). Accuracy of multi-trait genomic selection using different methods. *Genetics Selection Evolution* 43, 26. doi:10.1186/1297-9686-43-26.
- Campbell, M., Walia, H., and Morota, G. (2018). Utilizing random regression models for genomic prediction of a longitudinal trait derived from high - throughput phenotyping. *Plant Direct* 2, e00080. doi:10.1002/pld3.80.
- Crispim, A. C., Kelly, M. J., Guimarães, S. E. F., Silva, F. F. e, Fortes, M. R. S., Wenceslau, R. R., et al. (2015). Multi-Trait GWAS and New Candidate Genes Annotation for Growth Curve Parameters in Brahman Cattle. *PLoS One* 10, e0139906. doi:10.1371/journal.pone.0139906.
- Das, K., Li, J., Wang, Z., Tong, C., Fu, G., Li, Y., et al. (2011). A dynamic model for genome-wide association studies. *Human Genetics* 129, 629–639. doi:10.1007/s00439-011-0960-6.
- Endelman, J. B. (2011). Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome Journal* 4, 250. doi:10.3835/plantgenome2011.08.0024.
- Furbank, R. T., and Tester, M. (2011). Phenomics – technologies to relieve the phenotyping bottleneck. *Trends in Plant Science* 16, 635–644. doi:10.1016/j.tplants.2011.09.005.
- Jia, Y., and Jannink, J.-L. (2012). Multiple-Trait Genomic Selection Methods Increase Genetic Value Prediction Accuracy. *Genetics* 192, 1513–1522. doi:10.1534/genetics.112.144246.
- Kajiya-Kanegae, H., Nagasaki, H., Kaga, A., Hirano, K., Ogiso-Tanaka, E., Matsuoka, M., et al. (2021). Whole-genome sequence diversity and association analysis of 198 soybean accessions in mini-core collections. *DNA Research* 28, dsaa032. doi:10.1093/dnares/dsaa032.
- Koetz, B., Baret, F., Poilvé, H., and Hill, J. (2005). Use of coupled canopy structure dynamic and radiative transfer models to estimate biophysical canopy characteristics. *Remote Sensing of Environment* 95, 115–124. doi:10.1016/j.rse.2004.11.017.
- Krishnappa, G., Savadi, S., Tyagi, B. S., Singh, S. K., Masthigowda, M. H., Kumar, S., et al. (2021). Integrated genomic selection for rapid improvement of crops. *Genomics* 113, 1070–1086. doi:10.1016/j.ygeno.2021.02.007.
- Liu, S., Baret, F., Abichou, M., Boudon, F., Thomas, S., Zhao, K., et al. (2017). Estimating wheat green area index from ground-based LiDAR measurement using a 3D canopy structure model. *Agricultural and Forest Meteorology* 247, 12–20. doi:10.1016/j.agrformet.2017.07.007.
- Ma, C.-X., Casella, G., and Wu, R. (2002). Functional mapping of quantitative trait loci underlying the character process: a theoretical framework. *Genetics* 161, 1751–62.
- Meuwissen, T. H., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–29.
- Nelder, J. A. (1961). The Fitting of a Generalization of the Logistic Curve. *Biometrics* 17, 89. doi:10.2307/2527498.
- Onogi, A., Ogino, A., Sato, A., Kurogi, K., Yasumori, T., and Togashi, K. (2019). Development of a structural growth curve model that considers the causal effect of initial phenotypes. *Genetics Selection Evolution* 51, 19. doi:10.1186/s12711-019-0461-y.

- Piles, M., Gianola, D., Varona, L., and Blasco, A. (2003). Bayesian inference about parameters of a longitudinal trajectory when selection operates on a correlated trait. *Journal of Animal Science* 81, 2714–2724. doi:10.2527/2003.81112714x.
- Soltani, A., and Sinclair, T. (2012). *Modeling Physiology of Crop Development, Growth and Yield*. Wallingford: CABI.
- Sun, J., Rutkoski, J. E., Poland, J. A., Crossa, J., Jannink, J.-L., and Sorrells, M. E. (2017). Multitrait, Random Regression, or Simple Repeatability Model in High-Throughput Phenotyping Data Improve Genomic Prediction for Wheat Grain Yield. *The Plant Genome* 10(2), . doi:10.3835/plantgenome2016.11.0111.
- Toda, Y., Kaga, A., Kajiya-Kanegae, H., Hattori, T., Yamaoka, S., Okamoto, M., et al. (2021). Genomic prediction modeling of soybean biomass using UAV-based remote sensing and longitudinal model parameters. *The Plant Genome*, e20157. doi:10.1002/tpg2.20157.
- Verger, A., Vigneau, N., Chéron, C., Gilliot, J.-M., Comar, A., and Baret, F. (2014). Green area index from an unmanned aerial system over wheat and rapeseed crops. *Remote Sensing of Environment* 152, 654–664. doi:10.1016/j.rse.2014.06.006.
- White, J. W., Andrade-Sanchez, P., Gore, M. A., Bronson, K. F., Coffelt, T. A., Conley, M. M., et al. (2012). Field-based phenomics for plant genetics research. *Field Crops Research* 133, 101–112. doi:10.1016/j.fcr.2012.04.003.
- Winsor, C. P. (1932). The Gompertz Curve as a Growth Curve. *Proceedings of the National Academy of Sciences of the United States of America* 18, 1–8. doi:10.1073/pnas.18.1.1.
- Wu, W., Zhou, Y., Li, W., Mao, D., and Chen, Q. (2002). Mapping of quantitative trait loci based on growth models. *TAG Theoretical and Applied Genetics* 105, 1043–1049. doi:10.1007/s00122-002-1052-8.
- Yang, G., Liu, J., Zhao, C., Li, Z., Huang, Y., Yu, H., et al. (2017). Unmanned Aerial Vehicle Remote Sensing for Field-Based Crop Phenotyping: Current Status and Perspectives. *Frontiers in Plant Science* 8, 1111. doi:10.3389/fpls.2017.01111.