博士論文

Time-series measurement of crop growth process and modeling of genetic and environmental effects

(作物成長過程の時系列計測とその遺伝的・環境効果のモデル化)

Yusuke Toda

戸田悠介

Index

1 Introduction 3

2 Summary of prior studies about the modeling of growth process in crop breeding 5

2.1 General methods of quantitative genetic analysis 5

2.2 Modeling methods of the growth process 7

2.3 Use of crop growth models 9

3 Predicting biomass of rice with intermediate traits: modeling method combining crop growth models and genomic prediction models 11

3.1 Introduction 11

3.2 Materials and methods 13

3.3 Results 23

3.4 Discussion 31

4 Genomic prediction modeling using longitudinal model parameters and its application to soybean biomass and UAV-based remotely sensed data 35

4.1 Introduction 35

4.2 Materials and Methods 36

4.3 Results 48

4.4 Discussion 55

5 Longitudinal growth analysis of soybean using UAV-based remote sensing and its application on genomic prediction 60

5.1 Introduction 60

5.2 Material and methods 61

5.3 Results 71

5.4 Discussion 84

6 Prediction of soybean growth curves by modeling genetic and environmental effects on daily growth 87

6.1 Introduction 87

6.2 Material and methods 88

6.3 Results 97

6.4 Discussion 125

7 Discussion 127

7.1 Data acquisition 127

7.2 Modeling methods 128

8 Acknowledgments 130

9 Abstract 132

10 References 136

1 Introduction

The mission of crop breeding is to improve agricultural products' quality and quantity by creating new genotypes. To date, many theories and methods have been studied in order to achieve better improvement in breeding procedures. In recent years, the development of technologies for acquiring genomic data has increased the importance of statistical models that relate genomic data (i.e., whole-genome genotype data) to phenotypic data of target traits. Quantitative genetic models, such as genome-wide association study (GWAS) and genomic prediction (GP), have been applied to crop improvement by enabling the discovery of genetic variants and the prediction of genetic ability. Quantitative genetic models will continue to play an essential role in plant breeding. The use of genomic data and quantitative genetic models is expected to increase the efficiency and speed of plant breeding.

To increase the potential of genomics-assisted plant breeding, it is necessary to consider the impact of non-genetic factors on the traits of interest. Environmental effects and their interactions (genotype-by-environment interactions, G × E) have considerable effects on crop production (Kang, 2001). Several analytical methods for environmental effects and G × E (Burgueño et al., 2012; Jarquín et al., 2014) and the integrated use of quantitative genetic and plant physiological models (Heslot et al., 2014; Technow et al., 2015) were proposed. Crop growth models (CGM) are systematic models of plant growth under environmental changes in which accumulated physiological knowledge is aggregated, providing the ability to simulate plant growth under a given environment. The integration of CGMs and GP is expected to be a promising approach to systematically consider both genetic and environmental effects (Ramirez-Villegas et al., 2015).

In CGMs, crop traits at harvest are described as a chronological accumulation of growth. In plant physiology, many models of the relationship between plant growth and environmental factors have been discovered and accumulated in the main and sub-models of CGMs. However, genetic effects and the effect of G × E on plant growth have not been fully considered in CGMs because they require time-series data observed during the plant growth process of many genotypes. Since many genotypes are evaluated simultaneously in a breeding program, the collection of time-series data requires a great deal of effort.

In recent years, rapid advances in sensing technologies have made it possible to measure plant phenotypes in a high-throughput and time-series manner. A variety of sensing machinery, such as unmanned aerial vehicles (UAVs), tractors, and hand-held devices, can not be used to acquire phenotypic data of various characteristics for a large number of plants in the field. Also, several automated sensing facilities, called high-throughput phenotyping platforms, have been developed to track plant growth accurately under controlled environments. Thus, it is expected that the application of the high-throughput phenotyping will provide time-series data of a large number of genotypes, which will enable the modeling and analysis of G × E patterns present in plant growth processes.

This dissertation consists of four topics to develop new models for the relationships among crop growth process data, genomic data, and environmental data.

As the first step, CGM parameters of various rice genotypes were estimated using accurate growth process data measured manually. Next, the application of GP to the estimated CGM parameters was tested on whether it would improve the prediction accuracy of biomass at harvest. Here, an integrated CGM and GP model including growth process data was developed in Chapter 3. In this study, phenotype data of recombinant inbred lines of rice were used. The time-series data of leaf number and tiller number were parameterized with other growth-related traits in CGM, and the estimated parameters were predicted in GP using the genotypic data. The predicted biomass was calculated using the parameters predicted by CGM. The results showed that the proposed model improves the accuracy of biomass prediction.

In the previous section, the inclusion of manually measured growth process data improved the GP accuracy of biomass. In Chapter 4, the same approach was examined with the growth process measurements using UAV remote sensing (UAV-RS). UAV-RS was used as growth process data to measure the canopy area and height of soybean core collections. Simple growth models were used to summarize their growth patterns as parameters because UAV-RS was less accurate than manual measurements, and it was difficult to estimate CGM parameters. These parameters were used as secondary traits in GP of biomass to test whether this growth evaluation method would benefit genomic selection. The inclusion of growth process data and estimated parameters was found to improve the biomass prediction accuracy over the standard GP. Growth model fitting was shown to be an effective way to summarize the growth process data.

In Chapter 5, the fitting of the growth model was used to predict the growth process measured with UAV-RS. Time-series changes in the canopy area of soybean core collections were observed more frequently than in Chapter 4 to trace the growth process more accurately. Next, the prediction accuracy of an integrated model of the growth model and GP was tested in several prediction schemes. The proposed model successfully improved the second half of the growth process's prediction accuracy when the first half of the growth process was supplied as training data.

In Chapters 4 and 5, models were developed that connect the marker genotype and growth process, ignoring the environmental factors' influence. In Chapter 6, using the same data as Chapter 5, the environmental effects and the G × E effect on daily growth were estimated. Since the UAV-RS data had significant noise and bias, making it difficult to estimate the subtle environmental response, a noise filtering model was developed and applied to the canopy area and height. Next, statistical and machine-learning models were developed to explain the relationship between the estimated daily growth and environmental factors. Although the environmental response estimation was largely dependent on environmental data characteristics, the G × E related to daily growth was visualized by soil moisture.

2 Summary of prior studies about the modeling of the growth process in crop breeding

2.1 General methods of quantitative genetic analysis

Various quantitative genetics methods have been proposed to model the association between marker genotype data and phenotype data. The objectives of the methods include analysis of the genetic effects on traits, such as heritability estimation, detection of genes and quantitative trait locus (QTL), such as genome-wide association study (GWAS), and prediction of phenotypes also known as genomic prediction (GP). Despite the wide range of objectives, the basis of these methods is linear mixed models. This section briefly explains the basic concepts of these methods.

Before these analyses, it is common to calculate the genotypic values (or breeding values) with mixed models in recent studies. The genotypic value is assumed to reflect the actual genetic ability that each genotype possesses. The phenotypic values of block i, genotype j, and repetition k are decomposed into three terms:

$$y_{ijk} = \mu + \beta_i + s_j + e_{ijk} \quad (2\text{-}1),$$

where y_{ijk} is a phenotypic value, μ is a mean, β_i is a fixed effect, s_j is a random effect of genotype (breeding value), and e_{ijk} is a residual. The breeding values and residuals are assumed to follow the Gaussian distribution, $N(s_j \mid 0, \sigma_g^2)$ and $N(e_{ijk} \mid 0, \sigma_e^2)$. The matrix notation of Eq. 2-1 is

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{L\beta} + \mathbf{Qs} + \mathbf{e} \quad (2\text{-}2),$$

where $\mathbf{1}$ is a vector in which all the elements are one, $\mathbf{y}$, $\boldsymbol{\beta}$, $\mathbf{s}$, and $\mathbf{e}$ are vectors of y_{ijk}, β_i, s_j, and e_{ijk}, respectively, and $\mathbf{L}$ and $\mathbf{Q}$ are design matrices. Here, $\mathbf{s}$ and $\mathbf{e}$ are assumed to follow the multivariate Gaussian distribution, $N(\mathbf{s} \mid \mathbf{0}, \sigma_g^2\mathbf{I})$ and $N(\mathbf{e} \mid \mathbf{0}, \sigma_e^2\mathbf{I})$. The meanings and the number of the terms depend on each analysis (e.g., β_i may be a block effect, a yearly effect, or an environmental effect). The genotypic value ($\mathbf{g}$) is then calculated by adjusting breeding values with the mean,

$$\mathbf{g} = \mu\mathbf{1} + \mathbf{s} \quad (2\text{-}3).$$

A basic model for associating genotypic values with marker genotype data, which is common to genomic heritability estimation and GP, can be described with the simple equation:

$$\mathbf{g} = m\mathbf{1} + \mathbf{Zu} + \boldsymbol{\varepsilon} \quad (2\text{-}4),$$

where m is a mean, $\mathbf{u}$ is a vector of fitted or predicted genotypic values that follows the multivariate Gaussian distribution, $N(\mathbf{u} \mid \mathbf{0}, \sigma_u^2\mathbf{G})$, and $\boldsymbol{\varepsilon}$ is a residual vector that follows $N(\boldsymbol{\varepsilon} \mid \mathbf{0}, \sigma_\varepsilon^2\mathbf{I})$. $\mathbf{G}$ is a matrix determining the covariance structure of the predicted values, calculated with

marker genotype data. One definition of heritability can be written as the ratio of the estimated variances, $\sigma_u^2 / (\sigma_u^2 + \sigma_\varepsilon^2)$, which is called genomic heritability.

GP is a statistical method for predicting genotypic values of selection candidates from whole-genome marker genotypes. Equation 2-4 expresses a prior assumption in GP; the similarity of their whole-genome marker genotypes determines genotypic values. Two genotypic values will be similar if their marker genotypes are similar. If their whole-genome marker genotypes are significantly different, two genotypic values will not correlate or negatively correlate. Therefore, the genotypic values not obtained can be predicted using obtained genotypic values with similar marker genotypes.

The estimation of genotypic values (Eq. 2-2) and prediction of the genotypic values (Eq. 2-4) are sometimes integrated,

$$\mathbf{y} = \mathbf{L\beta} + \mathbf{Zu} + \mathbf{e} \tag{2-5}.$$

Here, the mean value, μ, is included as one of the elements of $\boldsymbol{\beta}$. This integrated approach is often used in growth analysis models to utilize marker genotype data in estimating growth model parameters, described in Section 1.1.2.

The variance-covariance matrix of the predicted genotypic values ($\mathbf{G}$) can be calculated in several ways. The inner product of the marker genotype data is one of the indices to evaluate the genotypic similarity,

$$\mathbf{G} = \mathbf{X}\mathbf{X}^{\mathrm{T}}/c \tag{2-6},$$

where $\mathbf{X}$ is an $N \times M$ marker genotype matrix (N and M are the numbers of genotypes and markers, respectively), and c is the normalization constant (Endelman & Jannink, 2012). Here, the elements of $\mathbf{X}$ were represented as −1, 0, and 1. Instead of the inner product, any kernel functions such as Gaussian kernel can also be applied.

Replacement of random effect ($\mathbf{u}$) of Eq. 2-4 with marker regression will lead to another formulation of GP (Meuwissen et al., 2001),

$$\mathbf{g} = m\mathbf{1} + \mathbf{Z}\left(\sum_{m=1}^{M} w_m \mathbf{x}_m\right) + \boldsymbol{\varepsilon} \tag{2-7},$$

where $\mathbf{x}_m$ is a vector of markers (m^{th} column vector of $\mathbf{X}$), and w_m is an effect of m^{th} marker on the genotypic values. In this case, the sum of the M marker effects is used to predict the unobserved genotypic values. Since the number of markers is usually larger than the number of genotypes ($N < M$), the least sum of squares' criterion cannot estimate the marker effects., Regularization methods (ridge regression, LASSO, and elastic net) and the Bayesian methods (Habier et al., 2011; Gianola, 2013) have been mainly used to circumvent this problem.

In breeding programs and field experiments, multiple traits are often measured from the same genotype. In such cases, a multivariate GP model can be applied to predict multiple traits simultaneously (Calus & Veerkamp, 2011; Jia & Jannink, 2012). If the measured traits have high

genetic correlations among them, the prediction accuracy of traits with low heritability can be improved (Calus & Veerkamp, 2011). The multivariate GP model can be expressed by extending Eq. 2-4 to K traits,

$$\begin{pmatrix} \mathbf{g}_1 \\ \vdots \\ \mathbf{g}_K \end{pmatrix} = \begin{pmatrix} m_1\mathbf{1} \\ \vdots \\ m_K\mathbf{1} \end{pmatrix} + \begin{pmatrix} \mathbf{Z}_1 & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{Z}_K \end{pmatrix} \begin{pmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_K \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_K \end{pmatrix} \quad (2\text{-}8).$$

The predicted breeding values $\mathbf{u}_{all} = (\mathbf{u}_1{}^T, \ldots, \mathbf{u}_K{}^T)^T$ are assumed to follow the multivariate Gaussian distribution $N(\mathbf{u}_{all} \mid \mathbf{0}, \mathbf{K} \otimes \mathbf{G})$, where $\mathbf{K}$ is the genotypic variance-covariance matrix of the traits and $\otimes$ is the Kronecker product. Therefore, data of correlated traits can be used to estimate or predict genotypic values. The residuals $\boldsymbol{\varepsilon}_{all} = (\boldsymbol{\varepsilon}_1{}^T, \ldots, \boldsymbol{\varepsilon}_K{}^T)^T$ are assumed to follow the distribution $N(\boldsymbol{\varepsilon}_{all} \mid \mathbf{0}, \mathbf{R} \otimes \mathbf{I})$, where $\mathbf{R}$ is the residual variance-covariance matrix of the traits.

For GWAS, the effect of the target marker is included in the mixed model (Eq. 2-4),

$$\mathbf{g} = m\mathbf{1} + b_m\mathbf{x}_m + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon} \quad (2\text{-}9),$$

where b_m is the effect of m^{th} marker. In this case, $\mathbf{u}$ is included to eliminate confounding genetic factors such as other QTLs and subpopulation effects. By testing the estimated effect of marker b_m, the probability of the existence of a causal QTL is statistically quantified (Zeng, 1994).

2.2 Modeling methods of the growth process

2.2.1 Drawbacks in the application of standard approaches

The simplest way to analyze time-series data is to apply the standard mixed models to the phenotype data at each time point separately. However, this method ignores covariances among time points, which leads to failure in analyses, such as loss of GP accuracy (Yin et al., 2004; Sun et al., 2017).

Another method is to apply a multivariate mixed model, but this is also problematic. As the number of measurements increases, the variance-covariance matrix's size among traits ($\mathbf{K}$ in Eq. 2-8) becomes too large to be estimated (Pletcher & Geyer, 1999). Thus, it is necessary to introduce models considering the longitudinal data structure.

2.2.2 Use of growth models

A common way to analyze time-series growth data is to fit a growth model. Several growth models have been proposed, including the Gompertz model (Winsor, 1932), the logistic model (Nelder, 1961), and the Richards model (Richards, 1959),

$$y = a\exp\{-b\exp(-kt)\} \quad (2\text{-}10),$$

$$y = a/\{1 + b\exp(-kt)\} \quad (2\text{-}11),$$

$$y = a\{1 + b\exp(-kt)\}^c \quad (2\text{-}12),$$

where t is observed time, and a, b, c, and k are the model parameters (Crispim et al., 2015). The parameters of a growth model are estimated by fitting the model to observed growth data. The estimation can be expressed by the following equation:

$$\mathbf{y}_t = f(t|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K) + \mathbf{e}_t \quad (2\text{-}13),$$

where $\mathbf{y}_t$ is a vector of phenotypic values at time point t, and $f(t \mid \theta_1, \ldots, \theta_K)$ is a growth function such as Eq. 2-10–2-12 with parameters $\theta_1, \ldots, \theta_K$. Here, $f(t \mid \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ means the vector of length L in which the elements are $f(t \mid \theta_{1l}, \ldots, \theta_{Kl})$ ($l = 1, \ldots, L$). The growth trajectory is summarized into the growth model parameters representing growth characteristics such as growth rate and growth period. The estimated model parameters are then analyzed with mixed models; if we apply Eq. 2-5, the parameters are expressed as

$$\boldsymbol{\theta}_k = \mathbf{L}\boldsymbol{\beta}_k + \mathbf{Z}\mathbf{u}_k + \boldsymbol{\varepsilon}_k \quad (k = 1, \ldots, K) \quad (2\text{-}14).$$

This model is used as an analytical tool (Onogi et al., 2016), QTL analysis (Ma et al., 2002; Wu et al., 2002), and GWAS (Crispim et al., 2015). Estimation of growth parameters (Eq. 2-13) and fitting of mixed models to the parameters (Eq. 2-14) can be done either stepwise (Wu et al., 2002; Crispim et al., 2015) or jointly (Ma et al., 2002; Onogi et al., 2016).

Since most growth models have several correlated parameters, it is natural to develop a multivariate mixed model of the parameters (Ma et al., 2002; Onogi et al., 2016),

$$\begin{pmatrix} \boldsymbol{\theta}_1 \\ \vdots \\ \boldsymbol{\theta}_K \end{pmatrix} = \begin{pmatrix} \mathbf{L}_1 & \cdots & \mathbf{O} \\ \vdots & \ddots & \vdots \\ \mathbf{O} & \cdots & \mathbf{L}_K \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_K \end{pmatrix} + \begin{pmatrix} \mathbf{Z}_1 & \cdots & \mathbf{O} \\ \vdots & \ddots & \vdots \\ \mathbf{O} & \cdots & \mathbf{Z}_K \end{pmatrix} \begin{pmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_K \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_K \end{pmatrix} \quad (2\text{-}15).$$

2.2.3 Random regression

Random regression is also one of the common modeling methods for time-series data. Although random regression is not employed as an analytical tool in this dissertation, it is introduced here because it is a powerful tool for dealing with growth processes. The first implementation was explained by introducing a time-series structure for covariances among traits (Kirkpatrick & Heckman, 1989; Kirkpatrick et al., 1990), but here, since these two implementations were equivalent (Meyer & Hill, 1997), time-series functions of fixed or random effects (Jamrozik & Schaeffer, 1997) are introduced.

Random regression can be described as an expansion of the mixed model (Eq. 2-5),

$$\mathbf{y}_t = \mathbf{L}\boldsymbol{\beta}(t) + \mathbf{Z}\mathbf{u}(t) + \mathbf{e}_t \quad (2\text{-}16),$$

where $\boldsymbol{\beta}(t)$ and $\mathbf{u}(t)$ are vectors of fixed and random effects, respectively, determined by the measurement time. Unlike Section 1.2.1, the shape of growth curves is not determined. Instead, the growth trajectory is expressed with the functions which show flexible shapes such as polynomials, the cubic spline, or the Legendre polynomials. An essential characteristic of these

functions is that they can be described with linear expansion (i.e., linear connections of feature vectors),

$$\boldsymbol{\beta}(t) = \boldsymbol{\Phi}_{\beta t}\mathbf{w}_{\beta}, \qquad \mathbf{u}(t) = \boldsymbol{\Phi}_{ut}\mathbf{w}_{u} \tag{2-17},$$

where $\boldsymbol{\Phi}_{\beta t}$ and $\boldsymbol{\Phi}_{ut}$ are $N \times D$ matrices of features of $\boldsymbol{\beta}(t)$ and $\mathbf{u}(t)$, N and D are the numbers of samples and features, and $\mathbf{w}_{\beta}$ and $\mathbf{w}_{\beta}$ are coefficients. Therefore, Eq. 2-16 can be expressed as

$$\mathbf{y}_t = \mathbf{L}\boldsymbol{\Phi}_{\beta t}\mathbf{w}_{\beta} + \mathbf{Z}\boldsymbol{\Phi}_{ut}\mathbf{w}_{u} + \mathbf{e}_t \tag{2-18}.$$

Thus, the random regression can be written with a linear combination of feature vectors making it easy to implement random regression in various software (Meyer, 2002; Meyer, 2007). The use of random regression is widely seen in research on the growth process such as GP (Sun et al., 2017; Campbell et al., 2018) and GWAS (Das, Li, Fu, et al., 2011; Das, Li, Wang, et al., 2011). One of the shortcomings of random regression is that the meanings of the estimated coefficients (**w**) are hardly understood compared to using a growth function (Section 1.2.2).

2.3 Use of crop growth models

Crop growth models (CGMs) have been developed to describe and predict plant growth using environmental data. The construction of CGM began in the 1960s with photosynthesis models and light interception (Wit, 1965; Bouman et al., 1996). Nowadays, many advanced CGMs are available for various purposes, from field management to political decision-making (Soltani & Sinclair, 2012). By utilizing the extensive knowledge of plant physiology, CGMs include various environmental factors on crop growth processes.

Here, the basic components of CGMs are explained. One essential component of CGMs is a model of the growth stage. Its basis is the cumulation of daily growth:

$$\mathrm{DVS}_d = \sum_{d} \mathrm{DVR}_d \tag{2-19},$$

$$\mathrm{DVR}_d = f(\mathrm{T}_d, \mathrm{P}_d) \tag{2-20},$$

where DVS_d is a developmental stage of day d, DVR_d is a developmental rate, T_d is the daily temperature, and P_d is the day length. The developmental stage corresponds to the plant phenology, e.g., DVS = 0, 1, and 2 correspond to emergence, flowering, and maturity, respectively (Horie, 1987).

The developmental stage can be used as an explanatory variable for growth process data. QTL analysis of the rice growth process using the developmental stage has shown that the effect of a gene related to the heading date was distinguished (Yin et al., 1999).

Another essential component of the model is the formalization of plant biomass products (Soltani & Sinclair, 2012). The daily value of dry matter production (DMP_d) can be expressed as

$$\mathrm{DMP}_d = \mathrm{RUE}_d \times \mathrm{PAR}_d \times \mathrm{FINT}_d \quad (2\text{-}21),$$

where RUE_d is radiation use efficiency (g MJ^{-1}), PAR_d is photosynthetically active radiation (MJ), FINT_d is the fraction of the incident radiation intercepted by the leaves. It is known that FINT_d can be estimated as an exponential equation of the leaf area index (LAI) based on the Beer-Lambert Law,

$$\mathrm{FINT}_d = 1 - \exp(-\mathrm{KPAR} \times \mathrm{LAI}_d) \quad (2\text{-}22),$$

where KPAR is the extinction coefficient, which differs for each species. The formation of LAI can be decomposed into detailed factors such as the number of nodes, the growth stage, and leaf shape parameters. Environmental stresses such as drought, heat, nutrition deficit can be included in CGMs by setting models of environmental effects on RUE, LAI, and other detailed components.

In terms of data assimilation, integrative analysis of CGM and growth process data has been considered. In those studies, CGM parameters were adjusted for each field and cultivar by data assimilation using time-series satellite RS data (Jin et al., 2018; Kasampalis et al., 2018). The purpose of these studies was the application to field management, not breeding. However, CGM integration with quantitative genetic models is expected to further improve crop production through breeding (Ramirez-Villegas et al., 2015).

3 Predicting biomass of rice with intermediate traits: modeling method combining crop growth models and genomic prediction models

3.1 Introduction

Genomic selection (GS) is a novel method increasingly being used in plant and animal breeding. In GS, the candidate genotypes are selected using genotypic values predicted with GP (Meuwissen et al., 2001). GP enables the prediction of genotypic values of a target trait without information about its causal genes, even when the target trait is controlled by a number of genes with complex interactions. Recent falls in the cost of genotyping high-density genome-wide markers have inspired the increased use of GP in animal breeding (García-Ruiz et al., 2016) and plant breeding (Asoro et al., 2013; Yabe et al., 2014; Rutkoski et al., 2015). Because phenotypic values predicted by GP can be used as alternatives to phenotypic values observed in field trials, GP can accelerate breeding by skipping field experiments for selection, and thus is expected to increase selection gains per unit time (Heffner et al., 2009).

Because environmental effects, i.e., the main effects of the environment and of the genotype-by-environment interaction (G×E), are generally not trivial in plant breeding, the use of GP models without consideration of these effects can cause difficulties in the application of GP to yield-related traits, which can be strongly influenced by these effects (Kang, 2001). Several methods have been proposed to consider environmental effects, including the modeling of covariance between genotype and environment (Burgueño et al., 2012; Jarquín et al., 2014), consideration of marker-by-environment interactions (Schulz-Streeck et al., 2013), and inclusion of environmental covariates (Pierre et al., 2016). Moreover, a GP model that can take environmental effects into account will benefit the application of GS in plant breeding because it will lead to more accurate predictions of genetic values for yield-related traits under a target environment and thus to a higher genetic gain per cycle (Heffner et al., 2009).

Crop growth models (CGMs) are expected to be an important tool for plant breeding because they incorporate environmental effects into the GP framework (Ramirez-Villegas et al., 2015). For example, a CGM was used to select the environmental covariates which were included in a GP model (Heslot et al., 2014). Also, a method for integrating a CGM and a GP model with approximate Bayesian computation was proposed (Technow et al., 2015), and it was applied the method to maize data (Cooper et al., 2016). However, the models in these studies attained only a small improvement in accuracy when applied to real data. One of the reasons for the small improvement may be the difficulty in parameter estimation of CGMs. The accurate estimation of CGM parameters is difficult when it is only based on observations of a target trait. In other words, the accuracy can be improved when observation of traits related to the target traits is included in the parameter estimation of CGM.

The growth-related traits may be good candidate traits to improve the prediction accuracy of target traits. Several studies have used growth-related traits with multivariate GP models to improve the prediction accuracy of target traits (Rutkoski et al., 2016; Sun et al., 2017), suggesting that the growth-related traits convey precise growth details and provide useful information for target trait prediction. To date, there has been no research that used growth-related traits for CGM and GP integration. In this study, a method to use the phenotypic data of growth-related traits in the integrated models of GP and CGM was proposed. This method has two steps. First, the growth-related traits are treated as "intermediate traits" and are predicted by GP. Second, the target traits are predicted from the predicted values of the "intermediate traits" and environmental data using a CGM. By dividing the model into two steps that correspond to GP and CGM, the "intermediate traits" can be naturally included into the model without complex statistical modeling of the relation between GP and CGM.

To validate this integrated model, rice [*Oryza sativa* (L.)] is a suitable research species because there have been previous studies of the application of GP (Xu et al., 2014; Grenier et al., 2015; Onogi et al., 2015; Spindel et al., 2015; Spindel et al., 2016; Wang et al., 2016) and CGMs (Pinnschmidt et al., 1995; Timsina & Humphreys, 2006; Iizumi et al., 2009), such as SIMRIW (Horie, 1987) and CERES-rice (Ritchie et al., 1987). However, attempts to integrate these methods to predict phenotypic variations in rice have been lacking, with some exceptions (Onogi, Watanabe, et al., 2016). Biomass is also a suitable trait for validation. Biomass is a direct target of breeding for biofuel rice (Oraby et al., 2007; Jahn et al., 2010) and is an important component of grain yield (Zhang et al., 2004; Khush, 2013).

In this study, models were developed to predict the biomass of rice, in which the observed phenotypic data of growth-related traits, whole-genome marker genotype, and environmental data were used. The model comprised two steps wherein the intermediate traits were predicted with GP in the first step and biomass was predicted from the predicted values of the intermediate traits in the second step. In the intermediate traits, the heading date is exceptionally predicted using a development rate (DVR) model based on the data obtained from multi-environmental trials (METs) and the genotypes of heading-date-related markers. Additionally, in the second step, the potential of a "black box"-type machine-learning model was evaluated, in which a detailed model structure was not defined as a priority for substituting the CGM.

These models were validated with a recombinant inbred line (RIL) population of japonica rice for biomass prediction. Phenotype data of 2-year field experiments of the population was provided for the analysis. The experiments were conducted with different timings of sowing (and planting) between both years to evaluate the potential of the models under different environments. The difference in sowing (and planting) dates was about one month, and this caused different phenological developments of the plants between those two years. Finally, the models were evaluated for their accuracy of biomass prediction within the experiments (using the same-year experiment for training and validation) and between the experiments (using one year's experiment for training and the other year's experiment for validation).

3.2 Materials and methods

3.2.1 Phenotype data

The analyses were conducted using phenotype data of 123 RILs derived from a cross between two *japonica* cultivars—Koshihikari and Kinmaze—and both parental lines. The construction of RIL was in the F_8 generation in 2014 and in the F_9 generation in 2015. Because Kinmaze and Koshihikari have different growth patterns and plant structure, these RILs were expected to be suitable for analyzing genetic variations observed in growth differences. In 2014 and 2015, experiments were conducted in an experimental paddy field of the National Agriculture and Food Research Organization, Tsukuba, Ibaraki, Japan (36° 01' N, 140° 06' E, 22m above sea level). Sowing and transplanting were performed in different months between years to produce results under different conditions of day length and temperature; seeds were sowed on 22 April 2014 and 19 May 2015 and transplanted seedlings into the field on 20 May 2014 and 18 June 2015. Because of different cultivation periods during 2014 and 2015, the 2-year experiments were not simply yearly replications but were expected to induce different growth patterns under different environmental conditions. Plants were transplanted 15 cm apart in rows 30 cm apart in plots. Two seedlings were transplanted per hill. The area for each line per replicate was 60 cm × 105 cm (2 rows × 7 hills). Inorganic fertilizer (80–100–100 kg of N-P_2O_5-K_2O ha^{-1}) was applied to the field. Aboveground plant organs were harvested to determine biomass at physiological maturity, which spanned from 29 August to 10 October in 2014 and from 17 September to 5 November in 2015 depending on variation among lines. Dry matter weight above ground was used as biomass.

Leaf age and number of tillers were recorded on each of several dates to evaluate variations in the growth pattern of the RILs (Table 3-1). The leaf age is calculated using the following formula (Zhou et al., 2001):

$$\text{Leaf age} = \text{Number of developed leaves} + \frac{\text{Length of the developing leaf}}{\text{Final length of the developing leaf}}$$

Year	Sowing date	Trait	Dates
2014	4/22	Leaf age	5/19, 6/2, 6/9, 6/16, 6/23, 6/30, 7/7, 7/14, 7/22, 7/28, 8/4
		Number of tillers	6/9, 6/16, 6/23
2015	5/19	Leaf age	6/15, 6/25, 7/2, 7/9, 7/16, 7/23, 7/30, 8/5, 8/10, 8/17, 8/24, 8/31
		Number of tillers	6/25, 7/2, 7/9, 7/16, 7/23

Table 3-1. Dates of observation of leaf age and number of tillers.

Leaf age was used instead of leaf number to treat the development of leaves as continuous values. The maximum tiller number was determined on the basis of measurements of the tiller number observed at three and five different time points in 2014 and 2015, respectively. The measurements were continued until the leaf number on the main culm reached to 11 or more. This was because the preliminary experiments with nine diverse cultivars suggested that the tiller number reached its maximum before 11 leaves were observed.

Length of the fully expanded leaf blades was measured for the 5th leaf, 11th leaf, flag leaf and 2 leaves below the flag leaf. According to the preliminary study, the final length of the leaf blade on the main culm increased almost linearly with the leaf age from 5 to 11. The increment in the final length per leaf age (ΔLL) was derived from the 5th and 11th leaves. Leaf age, number of tillers and leaf blade length on the main culm were recorded for two plants per entry for each replicate. Heading date and biomass were recorded on 6 plants per entry.

The marker genotype data of RILs were also provided. A method similar to (Okada et al., 2017) for the genotyping of RILs was used by extracting DNA from bulked seedlings of each F7 line (corresponding to the F6 generation). Single-nucleotide polymorphism (SNP) markers were used for the linkage map construction, and a total of 703 SNPs were selected from genome-wide SNP data (Nagasaki et al., 2010; Yamamoto et al., 2010). Using R software (R core team, 2017) and the R/qtl package (Broman et al., 2003), SNPs with identical genotypes were deleted. Finally, a total of 315 SNPs were used for the genotyping of RILs (Fig. 3-1).

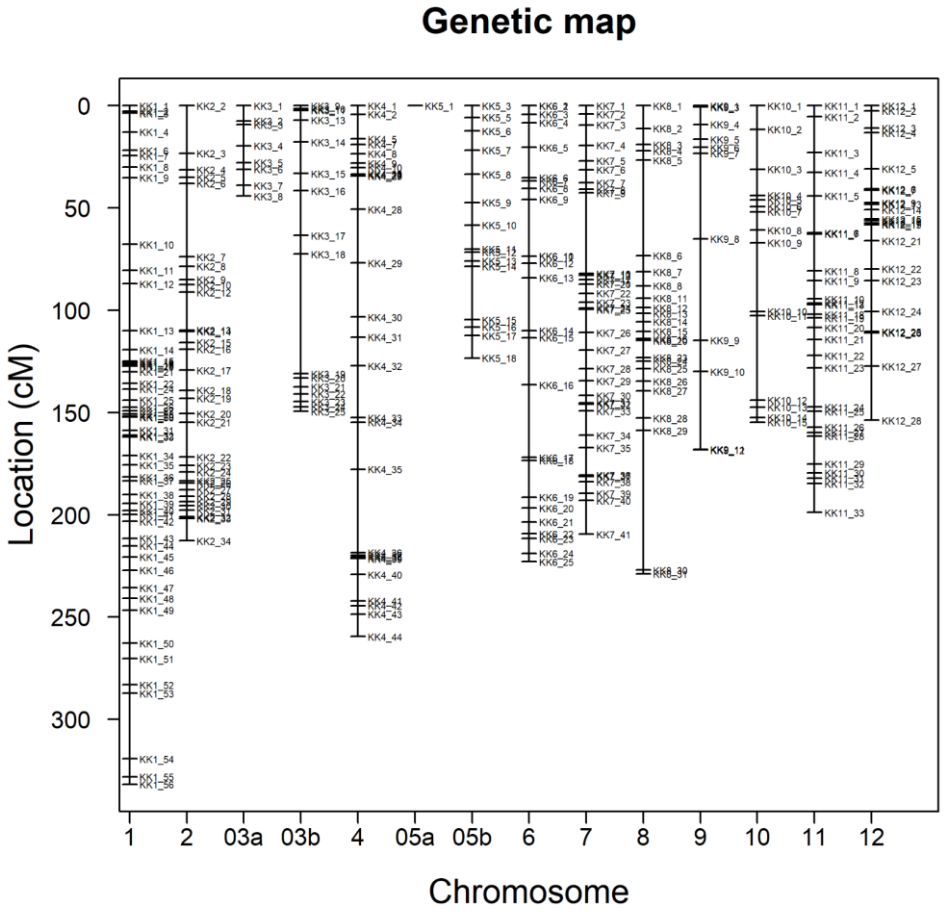

Figure 3-1. Genetic map of SNP markers of RILs.

Air temperature and solar radiation were recorded on-site (available at http://www.naro.affrc.go.jp/org/niaes/aws/). Photosynthetically active radiation (PAR) was estimated from the solar radiation assuming that proportion of PAR to the global solar radiation is 0.5 (Soltani & Sinclair, 2012). Daily means of temperatures are shown in Fig. 3-2.

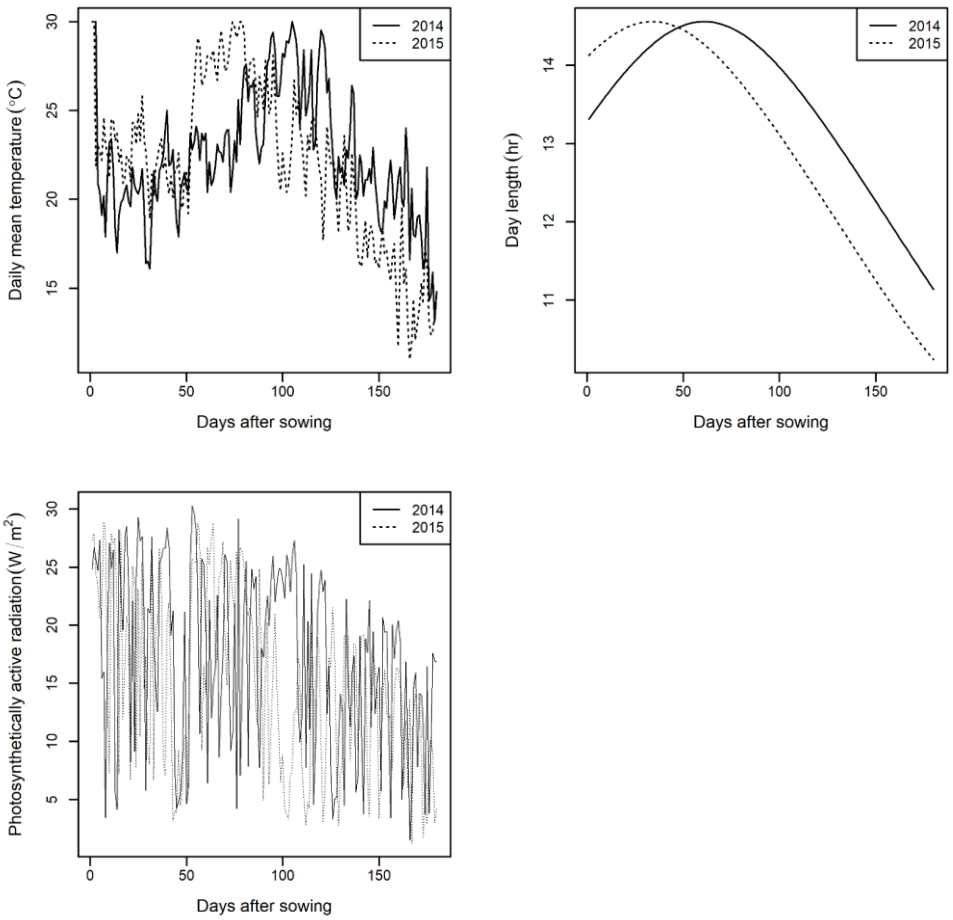

Figure 3-2. Environmental data during growing season. Daily mean temperature, theoretical day length and photosynthetically active radiation (PAR) of Tsukuba under field trial of RILs are shown. Data in both 2014 and 2015 are expressed as solid and dotted lines, respectively.

Because the RIL population was cultivated in only one field, it was difficult to estimate model parameters for heading date in CGM. To obtain the model parameters, heading dates recorded in METs were used that tested 112 cultivars, including Kinmaze and Koshihikari, most of which were developed in Japan. METs were conducted in six locations in several years (33 trials, Table 3-2).

Location	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Daisen, Akita									1	1	1
Tsukuba, Ibaraki					1	1	2	3	3	1	1
Tsukubamirai, Ibaraki	2	2									
Kasai, Hyogo			1	1	1	1	1	1	1	1	1
Fukuyama, Hiroshima			1	1	1	1	2	2	2	2	2
Fukuoka, Fukuoka									1	1	1

Table 3-2. Location, year, and number of replications of field experiments to record heading date.

3.2.2 Genetic analysis of observed traits

All statistical analyses were conducted in R software (R core team, 2017). The arithmetic means of observed values were used as phenotypic values for each RIL in the following analysis. The number of replications for each trait was described in the previous section. Analysis of variance (ANOVA) was conducted to evaluate the significance of genotype and environmental effects and their interaction.

The accuracy of GP of all traits was evaluated with 10-fold cross-validation. For building the GP models, four methods were employed. Two of them were regularized regression: ridge regression (RR) and LASSO, and the other two were Gaussian process regression (Xavier et al., 2016): one based on an additive relationship matrix (GBLUP) and the other on a Gaussian kernel matrix (RKHS) as a representative of covariance matrix. The "glmnet" package (Friedman et al., 2010) was used for RR and LASSO, and the "rrBLUP" package (Endelman, 2011) was used for GBLUP and RKHS. The narrow-sense heritability of each trait was estimated using a mixed model based on an additive relationship matrix in GBLUP.

3.2.3 Growth process analysis

The change in leaf age and the number of tillers during growth was analyzed as a simple function of heat units (accumulated daily mean temperature). The leaf age and the number of tillers on the i^{th} day from sowing (Leaf_i, Till_i, dimensionless) were represented as:

$$\text{Leaf}_i = \min(\Delta\text{Leaf} \times \text{HU}_i, \text{Leaf}_{\text{MAX}}) \tag{2-1}$$

$$\text{Till}_i = \begin{cases} 1 & (\text{HU}_i \leq 800) \\ \min(\Delta\text{Till} \times \text{HU}_i, \text{Till}_{\text{MAX}}) & (\text{HU}_i > 800) \end{cases} \tag{2-2}$$

where HU_i (°C) represents heat unit (daily mean temperature from emergence to the i^{th} date); ΔLeaf ($°C^{-1}$) and ΔTill ($°C^{-1}$) represent the rate of change per HU; and Leaf_{MAX} and Till_{MAX} represent maximum values. Because the growth of each line was observed, ΔLeaf and ΔTill were estimated as slopes of linear regression of phenotypic data during the study period, whereas Leaf_{MAX} and Till_{MAX} were measured at the end of the growth period. Because leaf age and number of tillers are generally not considered linear to HU, it was assumed that its growth was approximated by a combination of linear functions.

Generally, the growth of rice does not proceed when the daily temperature is low. To take this assumption into consideration, the growth models of leaf age and number of tillers were developed based on the heat unit, in which the base temperature of the growth of rice was considered (Σ max(0, daily mean temperature − 8°C)) instead of the simple heat unit. The lower bound of temperature was obtained from the previous knowledge (Soltani & Sinclair, 2012). However, the result did not largely differ or was even more inaccurate in the prediction accuracy than models developed based on the simple heat unit. Thus, only the results based on the simple heat unit are presented in this paper.

3.2.4 Prediction of heading date by DVR model

To predict heading date in a target environment, Yin et al.'s model (Yin et al., 1997) modified by Nakagawa et al. (Nakagawa et al., 2005) was used, which describes daily developmental rate (DVR) as a function of environmental factors (DVR model, hereafter). In the DVR model, daily progress of a developmental stage is expressed as a continuous value representing developmental stage (DVS), ranging from 0 (emergence) to 1 (heading). The DVS at the nth day after emergence is the sum of the daily development rates (DVR_i):

$$\mathrm{DVS}_n = \sum_{i=1}^{n} \mathrm{DVR}_i \tag{2-3}$$

where DVR_i is given by daily mean temperature (T_i, °C) and day length (P_i, h):

$$\mathrm{DVR}_i = \begin{cases} \dfrac{f(T_i)^{\alpha} g(P_i)^{\beta}}{G} & (\text{if } 0.145 + 0.005G \le \mathrm{DVS} \le 0.345 + 0.005G) \\ \dfrac{f(T_i)^{\alpha}}{G} & (\text{if } \mathrm{DVS} < 0.145 + 0.005G, 0.345 + 0.005G < DVS) \end{cases} \tag{2-4}$$

$$f(T_i) = \begin{cases} \dfrac{T_i - T_\mathrm{b}}{T_\mathrm{o} - T_\mathrm{b}} \left(\dfrac{T_\mathrm{c} - T_i}{T_\mathrm{c} - T_\mathrm{o}} \right)^{\frac{T_c - T_o}{T_o - T_b}} & (if\ T_\mathrm{b} \le T_i \le T_\mathrm{c}) \\ 0 & (if\ T_i < T_\mathrm{b}, T_\mathrm{c} < T_i) \end{cases} \tag{2-5}$$

$$g(P_i) = \begin{cases} \dfrac{P_i - P_\mathrm{b}}{P_\mathrm{o} - P_\mathrm{b}} \left(\dfrac{P_\mathrm{c} - P_i}{P_\mathrm{c} - P_\mathrm{o}} \right)^{\frac{P_c - P_o}{P_o - P_b}} & (if\ P_i \ge P_\mathrm{o}) \\ 1 & (if\ P_i < P_\mathrm{o}) \end{cases} \tag{2-6}$$

Six parameters were fixed (T_b = 8°C, T_o = 30°C, T_c = 42°C, P_b = 0h, P_o = 10h, P_c = 24h) among lines as in (Yin et al., 1997). The parameters α, β, G represent sensitivity to temperature, sensitivity to day length, and growth period, respectively, and are assumed to have specific values for each line. These parameter were estimated from the MET data using particle swarm optimization (Eberhart & Kennedy, 1995), which is used to optimize non-linear functions (Experiment A in Fig. 3-3).

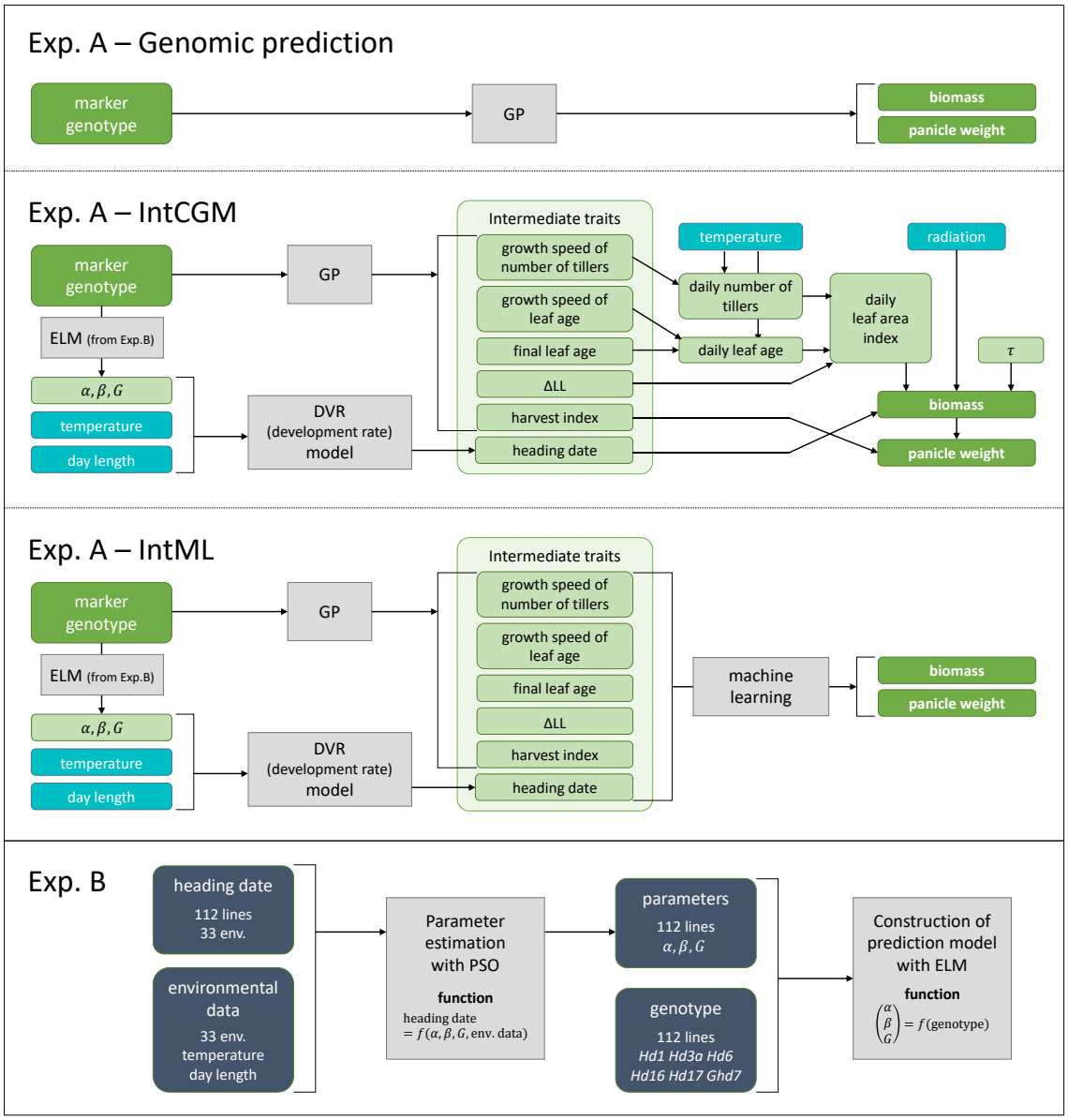

Figure 3-3. Flow chart of model structures. Experiment A: Process for estimating values of three parameters (α, β, G) related to heading date. Multi-environment trial data of heading date of 112 lines were used to model the relationship between parameter values and marker genotypes of heading-date-related genes using Extreme Learning Machine (ELM). Experiment B: Structure of conventional genomic prediction (GP), integrated CGM (IntCGM), and integrated machine-learning (IntML) models.

To calculate the values of α, β, G of the target RILs, models were constructed to predict them from marker genotypes (Experiment A in Fig. 3-3) of six heading-date–related genes (*Hd1*, *Hd3a*, *Hd6*, *Hd16*, *Hd17*, and *Ghd7*) (Yano et al., 2000; Takahashi et al., 2001; Kojima et al., 2002; Xue et al., 2008; Matsubara et al., 2012; Hori et al., 2013) of 112 lines. Extreme Learning Machine (ELM) (Huang et al., 2006), which is a machine learning method based on a neural network with advantages in generalization performance and learning speed, was used to model the relationships between the parameter values and the marker genotypes. After modeling these relationships, the values of α, β, G of the RILs were estimated by using the ELM prediction model (Experiment B in Fig. 3-3). The marker genotypes of the heading-date-related genes of the RILs were assumed to be the same as those of the SNP nearest to the genes, and were used as inputs of the ELM model.

3.2.5 Integrated GP–CGM model

Environmental effects were included in the model of yield-related traits by integrating the GP models and a CGM proposed by (Soltani & Sinclair, 2012), with modifications, to create an integrated CGM (IntCGM).

IntCGM has two steps (Experiment A in Fig. 3-3). First, the GP and DVR models predict "intermediate traits" related to biomass. LASSO was selected as a representative GP model because it showed the highest accuracy among all the GP models in 10 of 14 traits (i.e., six intermediate traits and biomass in two years). Second, the CGM simulates the daily change in biomass from the "intermediate traits".

Total biomass (BM, g m^{-2}) was estimated as the product of the total biomass at the day of termination of seed growth (BM_{TSG}, g m^{-2}) and a technical coefficient τ (dimensionless):

$$\mathrm{BM} = \tau\, \mathrm{BM}_{\mathrm{TSG}} \tag{2-7}$$

where τ represents the influence of factors that are not included in the model (e.g., precipitation, nutrient condition, disease) (Iizumi et al., 2009). The parameter τ was estimated as an average of the ratio of BM_{TSG} and observed BM when the prediction was conducted. The day of termination of seed growth was presumed to be the day when the accumulation of daily mean temperature after heading date reached 630 °C (Soltani & Sinclair, 2012). BM_{TSG} was calculated as the sum of daily increases of biomass:

$$\mathrm{BM}_{\mathrm{TSG}} = \sum_{i=1}^{\mathrm{TSG}} \mathrm{RUE}_i \times \mathrm{FINT}_i \times \mathrm{PAR}_i \tag{2-8}$$

where TSG is the day of termination of seed growth, $FINT_i$ is fraction of PAR intercepted by canopy of i^{th} day (dimensionless), RUE_i is radiation use efficiency (g MJ^{-1}), PAR_i is photosynthetically active radiation (MJ m^{-2}). RUE_i is the product of the maximum RUE (IRUE = 2.2 g MJ^{-1}) and the ratio of actual daily RUE to IRUE ($TRFRUE_i$, dimensionless):

$$\mathrm{RUE}_i = \mathrm{IRUE} \times \mathrm{TRFRUE}_i \tag{2-9}$$

where TRFRUE_i is a function of daily mean temperature (T_i) (Soltani and Sinclair, 2012):

$$\mathrm{TRFRUE}_i = \begin{cases} \dfrac{T_i - 10}{15} & (10 < T_i \leq 25) \\ 1 & (25 < T_i \leq 32) \\ \dfrac{T_i - 42}{10} & (32 < T_i \leq 42) \\ 0 & (otherwise) \end{cases} \tag{2-10}$$

FINT_i is estimated from the leaf area index, LAI_i (dimensionless), and the extinction coefficient ($k = 0.6$):

$$\mathrm{FINT}_i = \exp(1 - k\mathrm{LAI}_i)\,. \tag{2-11}$$

Although IRUE and k are known to have variation among lines and environments (Soltani & Sinclair, 2012), they are assumed to be constant in this study because of the difficulty in the estimation of IRUE and k for each line and environment. LAI_i is expressed as:

$$\mathrm{LAI}_i = \frac{\beta\ \mathrm{Till}_i}{S} \sum_{l=1}^{\mathrm{Leaf}_i} (l \times \Delta\mathrm{LL})^2 \tag{2-12}$$

where ΔLL (m) is the increase of leaf length per unit increase of leaf age, $\beta = 0.003$ is a technical coefficient explaining shape of leaves and S = 225cm^2 is the ground area of one plant. Thus, $l \times \Delta\mathrm{LL}$ represents the length of a leaf in one node, which came out in lth order, and $\sum_{l=1}^{\mathrm{Leaf}_i}(l \times \Delta\mathrm{LL})^2$ is expected to be proportional to leaf area of one tiller.

3.2.6 GP model integrated with machine learning

A model replacing the CGM with a machine learning method was also constructed. This integrated machine-learning model (IntML) has the same two-step structure as IntCGM, but the second step uses machine learning methods. In the second step, machine learning models that use intermediate traits as explanatory variables to predict biomass was built. A multiple regression model was chosen as a linear machine-learning method (IntML1) and the Random forest (Breiman, 2001) model was chosen as a non-linear method (IntML2). The R package "randomForest" (Liaw & Wiener, 2002) was used to build the Random forest prediction models. When building the model, the parameter "mtry" was set as 2 and the other parameters were set as default.

3.2.7 Model validation

The ability of the models to predict biomass was evaluated with 10-fold cross-validation among genotypes. The prediction of tested (i.e., training) and untested (i.e., validation) environments was also attempted. In the prediction of the tested environment, the data from the same year were used as both training and validation data; that is, biomass of a fold in one year was predicted from the data of the remaining folds and environmental data in the same year. This

assumption corresponds to the situation in which we want to predict the biomass of untested lines in tested environments. In the prediction of the untested environment, data from different years were chosen as training and validation data; that is, biomass of a fold in one year was predicted from the data of the remaining folds and environmental data in the other year. This assumption corresponds to the situation in which we want to predict the biomass of untested lines in untested environments.

Three statistics were calculated to measure prediction accuracy. The correlation coefficient of observed versus predicted values (r) is a measure of strength of relative relation between both values. The root mean squared error (RMSE) expresses the discrepancy between predicted and observed values. The regression coefficient of observed versus predicted values (slope) is a measure of shrinkage in the predicted values over the observed values. Observed and predicted values were used as dependent and independent variables, respectively. When predicted values approach observed values, r and slope approach 1 and RMSE decreases. The cross validation was repeated of 100 replicates of each combination of models and prediction schemes to estimate the standard deviation of indices (r and slope) of prediction accuracy. The Steel–Dwass test, a nonparametric multiple comparison test, was performed to examine significant differences in prediction accuracy.

3.3 Results

3.3.1 Growth patterns and correlations among traits

Growth curves and fitted models of leaf age and number of tillers are shown in Fig. 3-4. The results indicated that the models could express the growth of each trait despite their simplicity.

The comparison of phenotypic values between the two years of experiment is shown in Fig. 3-5. Among estimated parameters of the growth models, strong correlations between the years were observed in $Leaf_{MAX}$ and heading date whereas weak correlations were observed in $Till_{MAX}$ (Fig. 3-5). However, the distributions of ΔLeaf and ΔTill differed between the years. The ranges of phenotypic values of the heading date (e.g., minimum values were ca. 90 and 80 days in 2014 and 2015, respectively) and biomass also differed between the years, despite their high correlations. The G×E effect was found to be significant ($p < 0.01$) for all traits using ANOVA. The correlation coefficients between growth-related traits and biomass were higher in 2015 than in 2014.

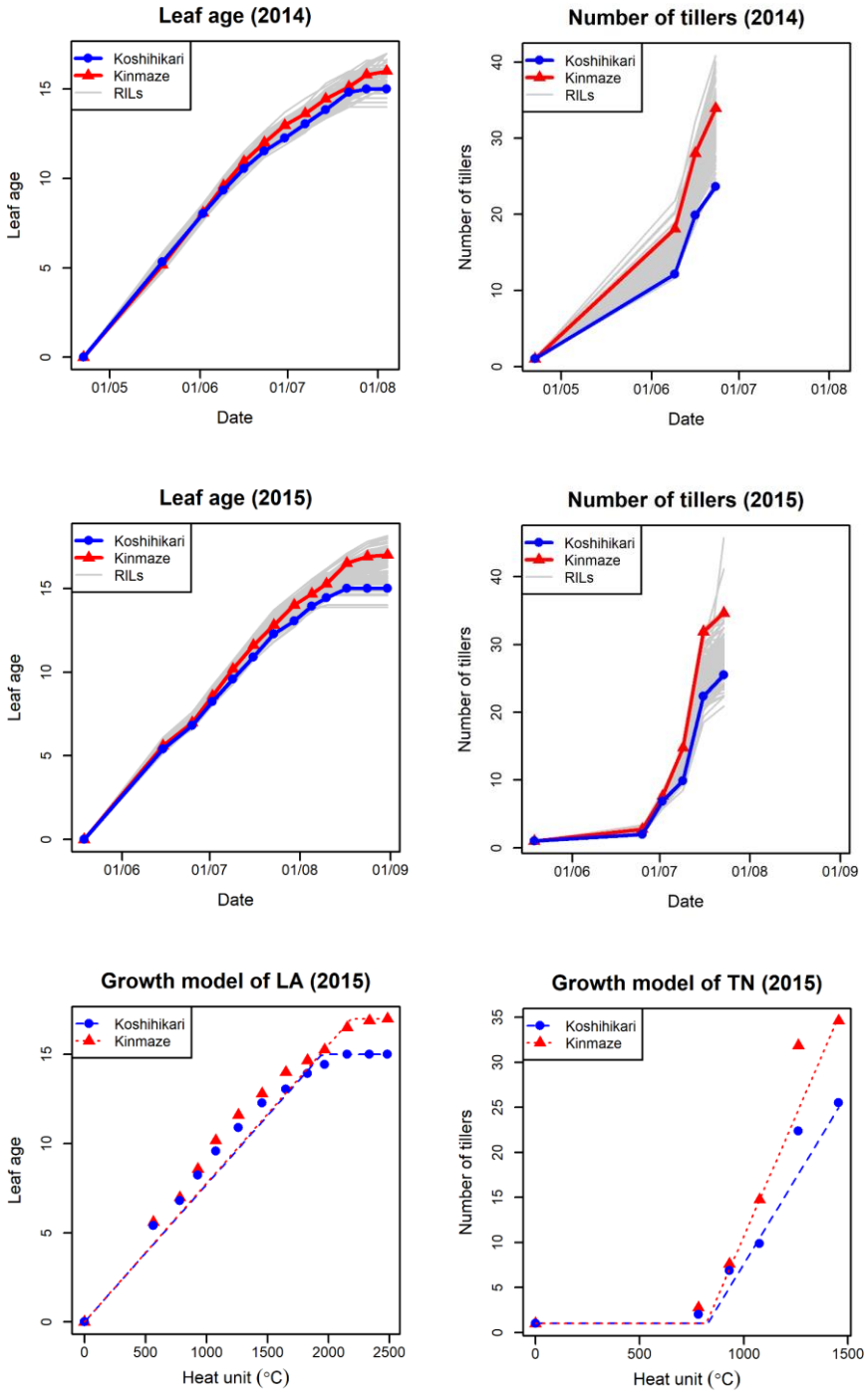

Figure 3-4. Growth curves and growth models of leaf age and number of tillers. The line means of both traits in 2014 and 2015 are plotted in four figures in the upper side. The parents, Koshihikari and Kinmaze, and RILs are expressed as blue, red and gray lines, respectively. The growth models of those traits are shown in two figures in the bottom side. The growth model and the observed values of parents in 2015 are shown. Heat unit is used as horizontal axes.

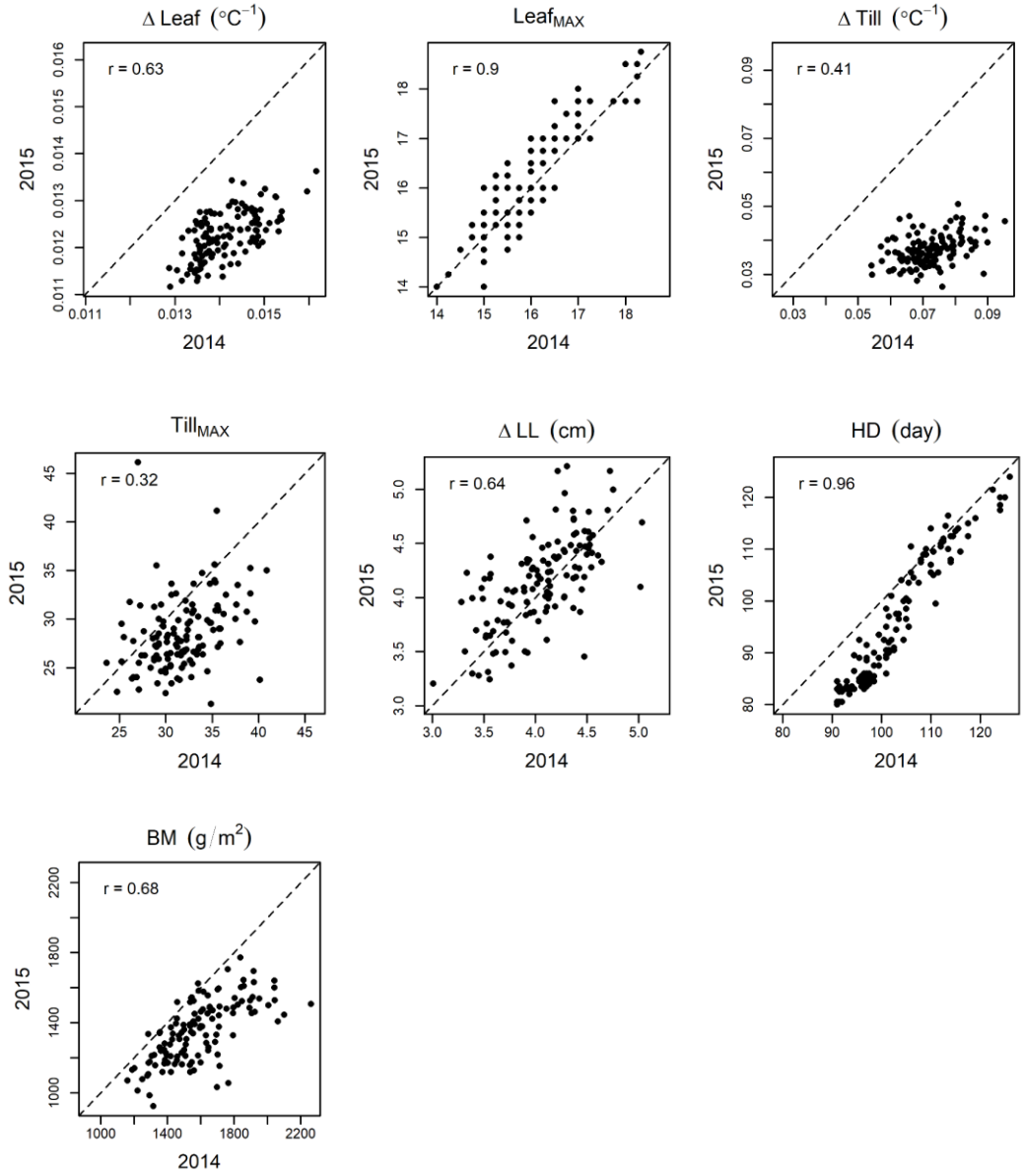

Figure 3-5. Comparison of observed traits between 2014 and 2015. Estimates of correlation coefficients between phenotypes of two years are shown in the top-left of each box. Abbreviations: ΔLeaf, growth rate of leaf age; $Leaf_{MAX}$, final leaf age; ΔTill, growth rate of number of tillers; TillMAX, maximum number of tillers; ΔLL, growth rate of leaf length per leaf age; HD, heading date; HI, harvest index; BM, biomass.

3.3.2 Genomic prediction of growth-related traits

The prediction accuracy of the GP models (Fig. 3-6) in growth-related traits were assessed, which corresponded to the first step of integrated models (IntCGM and IntML, Experiment B in Fig. 3-3). Accuracy was higher in 2015 than in 2014. Traits that showed higher correlation between years in Fig. 3-5 tended to have higher values both in heritability and prediction accuracy. In ΔTill and Tillmax, the accuracy was lower than in biomass. In the following analyses, LASSO was chosen as a representative GP model because it showed the highest accuracy among the models in 10 of 14 traits (six intermediate traits and biomass for two years). Five models were compared for heading date: the DVR model which used weather data and genome-wide marker data as explanatory variables and 4 GP models used only genome-wide marker data. The prediction accuracy was slightly lower in the DVR model than that in GP.

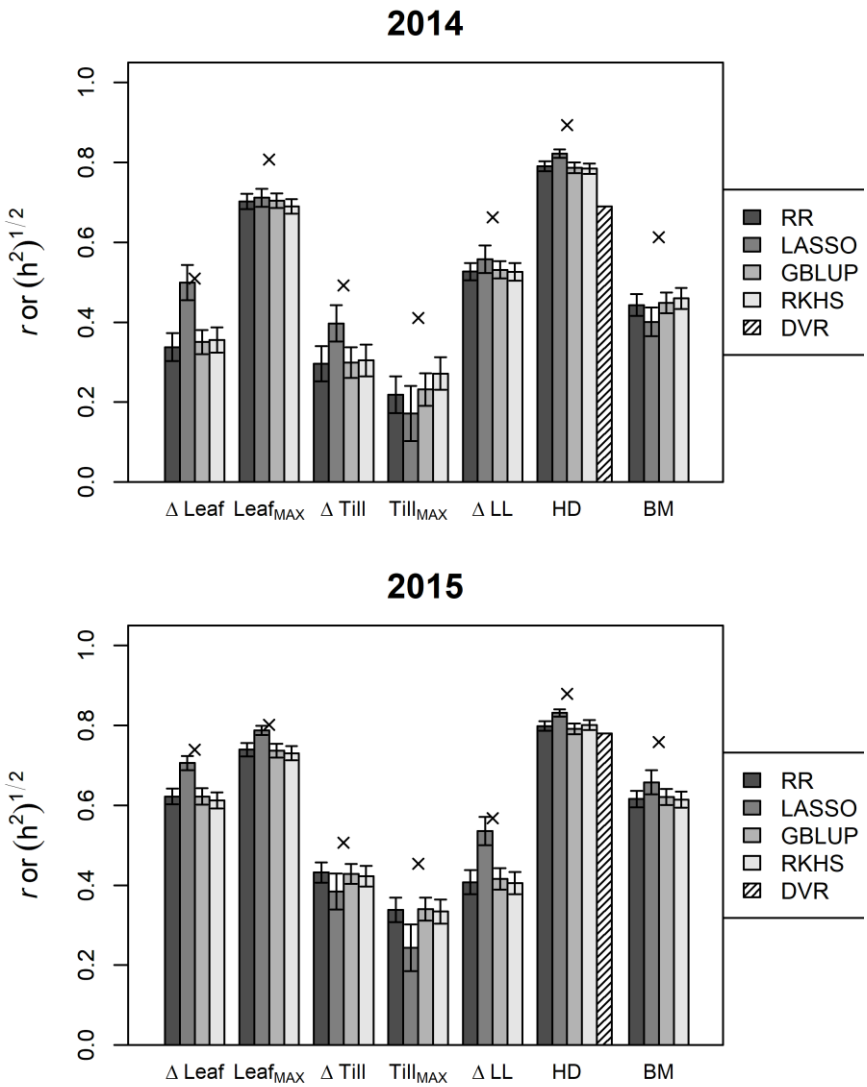

Figure 3-6. Comparison of prediction accuracy of GP and heritability in growth-related traits. Estimated correlation coefficients of observed values and values predicted using the five models for seven growth-related traits are shown as bars. The five models included four methods of whole-genome prediction (for all traits) and a DVR model with marker genotypes of the heading-date–related genes (for heading dates). The square roots of heritability of the seven traits are shown as crosses. Error bars represent ± 1 s.d.

3.3.3 Prediction of biomass

In the tested environment, IntCGM, IntML, or both were more accurate at biomass prediction than GP with LASSO by all three statistics (Fig. 3-7-A), especially when the 2014 dataset was used as validation data: that is, IntCGM and IntML gave higher r values and regression slopes closer to one than GP, and IntML gave lower RMSE than GP. This tendency was supported by the fact that differences between r and slope of the proposed models and those of GP were all statistically significant ($p < 0.01$).

IntCGM, IntML, or both performed better than or the same as GP in the untested environment (Fig. 3-7-B); both models gave significantly higher r and slope than GP except when IntML2 was tested with 2014 dataset as validation. IntCGM had a lower RMSE than that of GP using the 2015 dataset for validation but had a higher RMSE than that of GP using the 2014 dataset for validation.

The prediction of the panicle weight was also attempted with IntCGM, wherein the panicle weight was expressed as the multiplication of biomass and harvest index and the harvest index was predicted using GP. However, the prediction accuracy of IntCGM was worse than GP because the harvest index itself was largely affected by the environment (Fig. 3-8).

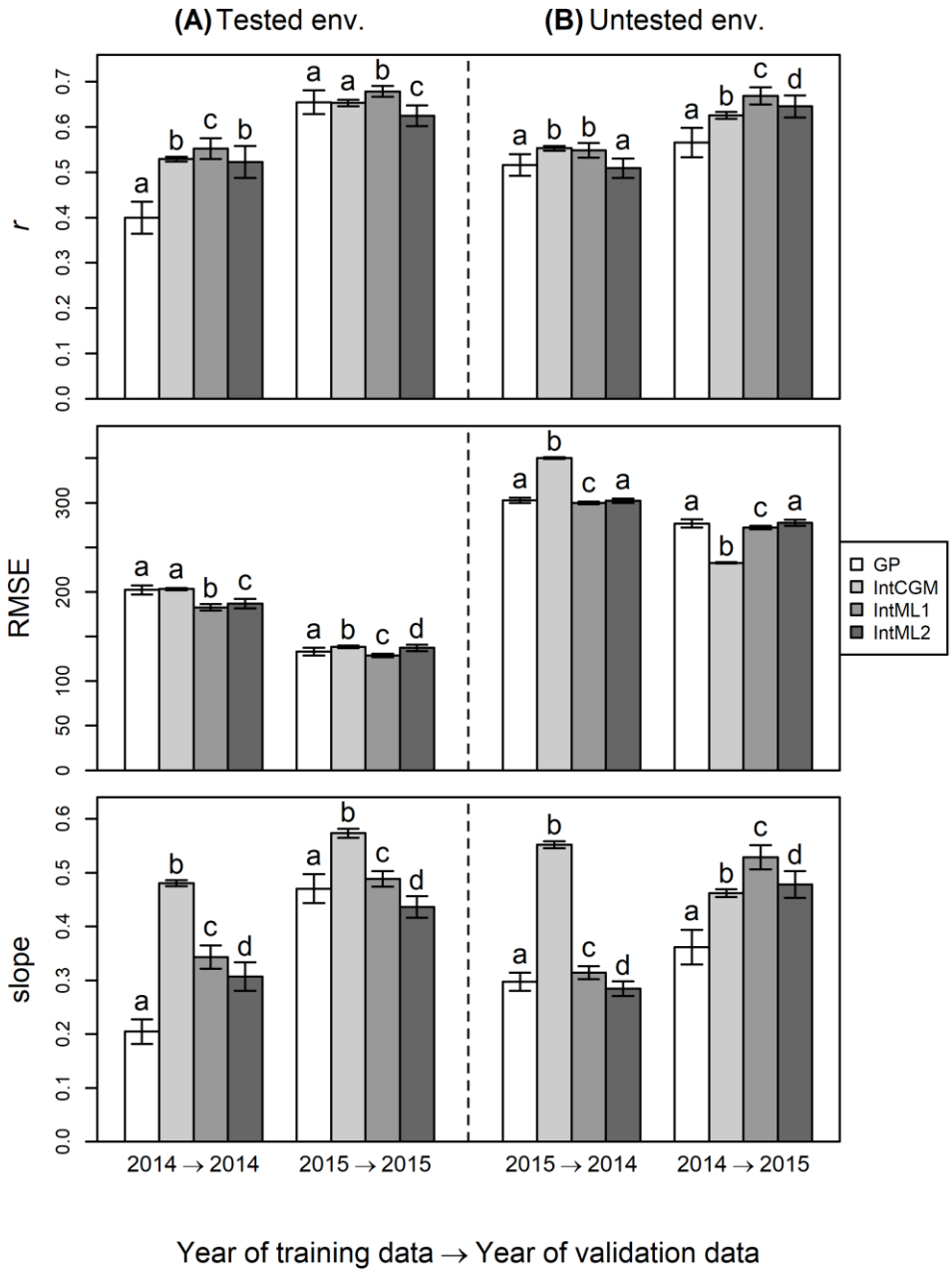

Figure 3-7. Comparison of prediction accuracy of biomass. Result of prediction of tested environment (A) and untested environment (B) are shown. LASSO was chosen as a representative GP model. Three indices are used: Correlation coefficient (*r*), RMSE (root mean squared error), and slope of the regression line for predicted and observed values. Error bars represent ± 1 s.d. Letters above the bars indicate a significant difference as determined by the Steel-Dwass test ($p < 0.01$).

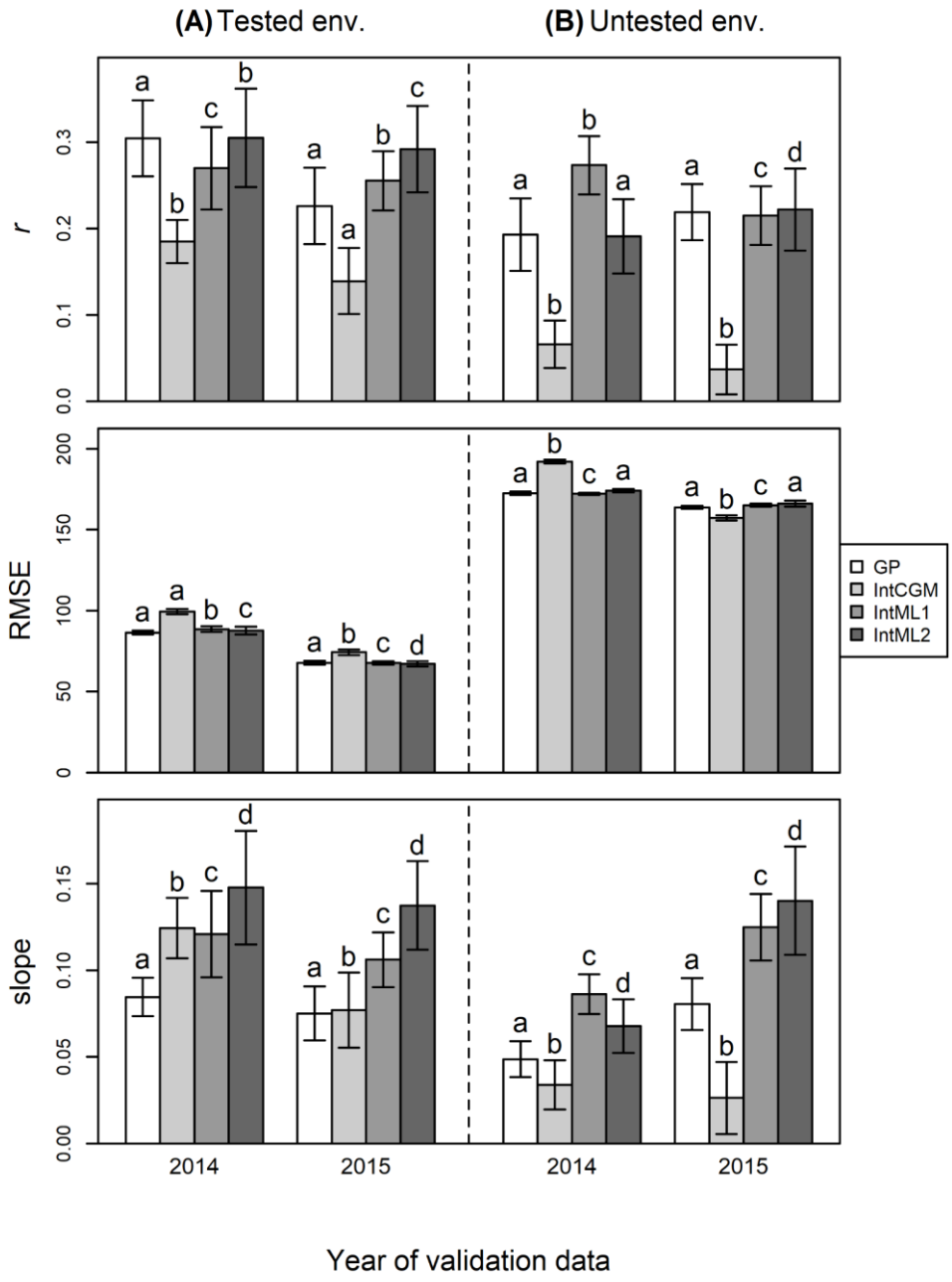

Figure 3-8. Comparison of prediction accuracy of panicle weight. The result of prediction of tested (A) and untested (B) environments are shown. LASSO was chosen as a representative GP model. Three indices are used: Correlation coefficient (*r*), RMSE (root mean squared error), and slope of the regression line for predicted and observed values. Error bars represent ± 1 s.d. Letters above the bars indicate a significant differences determined by the Steel-Dwass test ($p < 0.01$).

3.4 Discussion

3.4.1 Accuracy of prediction of biomass

The *r* in the new models was the same as, or higher than, that of the conventional GP in the prediction of biomass (Fig. 3-7). There was a substantial difference in the *r* of GP between 2014 and 2015 in the prediction of the tested environment, indicating that there was a difficulty in explaining the variation of biomass in 2014 through the direct linear regression of the genotypic markers. In contrast, the integrated models showed the significant increase in *r* compared with that of GP in the 2014 prediction. These results indicate that the use of the intermediate traits was beneficial for improving accuracy of biomass prediction. Heading date prediction, which showed high heritability in both years, mostly contributed to the improved prediction accuracy.

Focusing on the GP trained with biomass of 2014, the accuracy was higher in biomass prediction of 2015 than in that of 2014. This intuitively unexpected result might be owing to two reasons. One is the low heritability of biomass in 2014, which led to lower prediction accuracy in the models (Daetwyler et al., 2008; T. H. E. Meuwissen, 2009). To reduce the influence of the heritability level on the index of the prediction accuracy (i.e., a correlation coefficient between observed and predicted phenotypes), the value of *r* was adjusted by dividing it by the square root of genomic heritability. The adjusted values of *r* became 0.652 and 0.746 for the biomass in 2014 and 2015, respectively, and had smaller differences than the previous r. Another reason for the higher biomass prediction accuracy in 2015 is the GS model built with LASSO. In Fig. 3-6, the biomass prediction accuracy was lower in LASSO than in other models in 2014, whereas the result was the opposite in 2015. Polygenic marker effects seemed more dominant in biomass in 2014 than in 2015 because LASSO is not good at capturing the small effects of a large number of variables. In contrast, the estimation of genomic heritability effectively reflects polygene effects. The differences in the characteristics of each estimation method subsequently caused the difference in the adjusted values of *r* for the biomass in 2014 and 2015.

Although heading date was predicted by ELM and DVR models in the prediction models, the prediction accuracy was worse than that by GP. One possible reason is that the heading date of RILs could not be completely explained by heading-date-related genes, (i.e., *Hd1*, *Hd3a*, *Hd6*, *Hd16*, *Hd17*, and *Ghd7*) considered in ELM and DVR models. However, the DVR model was employed in the prediction models because it can be used to predict the heading date in a new environment.

3.4.2 Comparison with models in other studies

An advantage of the new approach over conventional researches of integrated models of GP and CGM is the inclusion of observed growth data in the model as “intermediate traits”. This enables us to treat parameters in the model as representations of actual crop conditions. Two studies designed to integrate a genomic prediction model with a crop model (Technow et al., 2015; Onogi, Watanabe, et al., 2016) tried to estimate growth parameters by using only phenotypic values of target traits. Technow et al. integrated GP and CGM to predict the yield of maize using

parameter estimation with the approximate Bayesian computation (Technow et al., 2015). Onogi et al. also constructed an integrated model to predict the heading date of rice (Onogi, Watanabe, et al., 2016). However, this approach is difficult to apply to a complex trait, such as yield, and did not improve the prediction accuracy when it was applied to real-yield data (Cooper et al., 2016). It is also difficult to validate the accuracy of the estimated growth parameters. The use of the intermediate traits was beneficial for improving prediction accuracy and for further understanding how the parameters influence the target traits.

A multivariate GP is another approach to predict target traits with intermediate traits (or secondary traits). In this model, the covariance structure among target and intermediate traits is considered to improve prediction accuracy (Calus & Veerkamp, 2011; Jia & Jannink, 2012). For example, there are studies in which longitudinal traits measured by remote sensing were used as intermediate (or secondary) traits and modeled with a multivariate GP model to predict wheat grain yield (Jessica Rutkoski et al., 2016; Sun et al., 2017). Grain yield was predicted for untested environment in which phenotypic data of a target population was not available. The prediction accuracy, however, was not improved with a multivariate GP model (Sun et al., 2017). Compared with multivariate GP model approach, the proposed two-step approach has a good flexibility to model nonlinear relationship among target and intermediate traits through applying a nonlinear model at the second step (e.g. CGM as in IntCGM or Random forest as in IntML2).

Another benefit of IntCGM was that the range of predicted among-lines variation [i.e., the regression coefficient (slope) of observed versus predicted values of IntCGM] was closer to 1 compared with that of GP (Fig. 3-7). This would be important in breeding programs (Moser et al., 2009; Onogi, Watanabe, et al., 2016), although it has not been evaluated in recent studies of the prediction of G×E by GP (Burgueño et al., 2012; Heslot et al., 2014; Jarquín et al., 2014; Cooper et al., 2016). In those studies, the accuracy of prediction models was assessed mainly by correlation between predicted and observed (or estimated) values. Although correlation is a good measure of the ordinal accuracy of the prediction (i.e., the accuracy of predicting the order of genotypic values), it does not necessarily reflect the range of genetic variations (González-Recio et al., 2014). In some cases, the accurate prediction of phenotypic values is important for breeding; for example, we may need to maintain the flowering date within a certain range for ease of field management or limit plant height to prevent lodging. When aiming at the application of GP to actual breeding the accurate prediction of the size of genetic variation in a population is as important as the ordinal relationship among genotypes in the population.

3.4.3 Further improvement of the prediction model

The prediction accuracy of the models was validated using 2-year experiments, which had a 1-month difference in the timing of sowing and planting; one year was used for training, whereas the other year was used as testing the prediction accuracy as previous researches did (Technow et al., 2015; Cooper et al., 2016). Although experiments in 2014 and 2015 were performed in one location, the 2-year experiments were conducted under different environmental conditions (e.g., temperature, day length, and radiation) by employing different cropping seasons. However, other environmental factors, such as soil condition, were fixed in these experiments. To apply the

proposed models to a dataset with multiple locations and years, we should take into account other environmental factors, such as soil condition, water supply, and cultivation management, in the models.

In this study the biomass was selected as the target trait for prediction, but the prediction of yield was more challenging. A possible method of implementing accurate prediction of yield is the use of sophisticated CGMs. The potential of several CGMs, such as APSIM (Holzworth et al., 2014), has been already demonstrated in practical applications. However, certain complexities may create problems. One of the problems is the accumulation of errors: the errors of parameter estimation would be large if the model includes several parameters. Therefore, models must be simplified in ways such as the use of machine learning (IntML) or variable selection. A sensitivity analysis will be effective to select modules of the models in which variables with little influence on target traits will be distinguished.

Another problem is the increased effort required for measuring plant growth if a model requires a large number of growth parameters. Parameter estimation is one effective solution (Iizumi et al., 2009; Technow et al., 2015; Onogi, Watanabe, et al., 2016). Through these methods, the measurement of some growth-related traits can be omitted by estimating them as parameters in a CGM while measuring the remaining traits in the field. The use of high-throughput phenotyping is another way to enable plant growth to be measured in detail. For example, LAI (Córcoles et al., 2013; Duan et al., 2014) and biomass (Montes et al., 2011; Watanabe et al., 2017) can be measured in a non-destructive way by remote sensing with unmanned aerial vehicles. Such techniques would enable us to measure various kinds of growth-related traits continuously during growth. GP and high-throughput phenotyping technologies could revolutionize breeding (Cabrera-Bosquet et al., 2012).

Moreover, the use of a deterministic model in IntCGM may reduce phenotyping costs for the target traits. In IntCGM, the phenotypic values of biomass in the training data were used only for scaling the model's prediction values onto the phenotypic values with τ as the scaling parameter. Using τ, the RMSE of biomass in known environments decreased by 45% and 68% in 2014 and 2015, respectively. However, the scaling procedure (i.e., the training of model with the phenotypic values of biomass) was not necessary with the use of the prediction values for selecting superior genotypes because the correlation between the predicted and genotypic values of biomass did not change with scaling. This is because the CGM used in this study was deterministic and did not include any parameters to be estimated other than τ. This is another great advantage of IntCGM because the model does not require the phenotypic data of biomass, which in turn requires the laborious destructive measurements of plants.

3.4.4 Toward application for breeding

In this study, the proposed method was validated with the dataset of the 2-year experiments, which had a 1-month difference in their timings of sowing and planting to simulate different environmental conditions. Although the validation is insufficient to evaluate the potential of the method, the proposed models may be applicable to multi-location-multi-year dataset because

CGM is expected to describe G×E when it has an appropriate model structure and the necessary environmental factors. Thus, IntCGM may enable accurate prediction of phenotypes in each target environment and accelerate the development of varieties having excellent viability in the target environments.

The proposed models may also help to explain the mechanisms causing G×E effects on yield-related traits because they can predict the effects physiologically through CGMs. The predicted values of growth-related "intermediate traits," as well as of yield-related traits, allow us to understand how environmental factors affect growth and have a large impact on yield. This understanding will be of benefit to the mechanical evaluation of environmental characteristics of locations and the appropriate choice of locations used in METs.

4 Genomic prediction modeling using longitudinal model parameters and its application to soybean biomass and UAV-based remotely sensed data

4.1 Introduction

The phenotype of a plant at harvest is a result of its genotype and growth process. From a plant physiology or crop growth modeling viewpoint, plant phenotype at harvest can be expressed as a result of the translocations and accumulation of carbon production throughout their growth (Soltani & Sinclair, 2012). Therefore, it is expected that the estimation or prediction of plant traits at harvest will improve in accuracy if the growth processes are appropriately accounted for in the models. This hypothesis can also be applied to genomic prediction (GP) models, which predict genotypic values (breeding values) from genome-wide marker data using statistical methods (T. H. Meuwissen et al., 2001). Recently, some extensions of GP models were proposed that incorporated plant growth processes (Technow et al., 2015). These reports suggested that the inclusion of growth processes into the GP models may improve the prediction accuracy and further the applications of GP in plant breeding (Technow et al., 2015).

To model growth processes in GP appropriately, sufficient data about the growth patterns of the target plant population are essential. However, there is a high labor cost to measure plants during growth in a breeding program, if relying on hand measurements and visual judgments. Recent technological developments have yielded an alternative phenotyping procedure, i.e., remote sensing (RS). For plant breeding, RS with tractors (White et al., 2012) or aerial devices such as unmanned aerial vehicles (UAVs) (Yang et al., 2017) were used to evaluate genetic variation in plant growth. It has already been shown that RS could capture the variation in a breeding population (Watanabe et al., 2017). Also, GP models that use a time-series of RS data could improve the prediction accuracies of wheat yields (Rutkoski et al., 2016). Therefore, the accuracy of GP is expected to improve by including longitudinal growth processes of plants as measured by RS.

Several approaches can be applied to include RS data into GP models. A multivariate GP (MGP) model (Calus & Veerkamp, 2011; Jia & Jannink, 2012) has been recently used to incorporate several traits into the GP model (Rutkoski et al., 2016; Crain et al., 2018). By leveraging correlation among traits, an MGP model could improve the prediction accuracies of those traits with low heritability (Jia & Jannink, 2012). However, the accurate estimation of a covariance matrix may be difficult if the number of traits being measured becomes too large. This leads to a dilemma, in which frequent RS will provide detailed data of growth processes of plants to the improvement of GP accuracy, but incorporation of this data enlarges the dimension of the covariance matrix of the MGP model and hence decreases the accuracy due to the estimates of the covariance matrix. In such cases, an appropriate dimension-reduction method, which can be

applied to longitudinal RS data, may improve the accuracy of a MGP model with a low dimensional covariance matrix.

There are several statistical methods of dimension reduction, including a principal component analysis. However, there is no guarantee that the selected direction of data compression reflects the genetic divergence of growth curves. On the other hand, usually some empirical prior knowledge about the patterns of the growth curves can be obtained. The modeling of the variations in the patterns of growth curves based on prior knowledge may enable the effective extraction of growth characteristics as model parameters. Here fitted simple growth models were fitted to RS data and used model parameters as a representation of each growth curve. Several reports have shown that the use of growth models improved the results of QTL analyses (Ma et al., 2002; Wu et al., 2002). The use of growth parameters in the MGP model is also expected to lead to an improvement in prediction accuracies.

Another possibility to improve the prediction accuracy of target traits is the use of a two-step GP (TGP) model. In the TGP model, secondary traits were first predicted using GP, and target traits were identified using predicted values of secondary traits (Chapter 3). Because this method estimates the direct linear relationship between secondary and target traits, instead of the covariance matrix between all traits, the TGP model may be effective even when high-dimensional RS data is included.

Prediction accuracies of several models were compared in this study, in which soybean [*Glycine max* (L.) Merr.] biomass was chosen as the target trait. Canopy height and area were estimated with UAV images, and simple growth models were fitted to growth curves of the canopy height and area to extract growth characteristics. The prediction accuracies of the biomass using GP, MGP, and TGP models were compared, using the observed values or growth parameters of canopy height and area as secondary traits.

4.2 Materials and Methods

4.2.1 Field trial

198 accessions of soybean registered in the National Agriculture and Food Research Organization Genebank (https://www.gene.affrc.go.jp/index_en.php) were employed in this study. In 2016, a field trial of the 198 accessions 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). Distances between each row, plot, and individual were 100 cm, 80 cm, and 20 cm, respectively (Fig. 4-1a). Each plot consisted of five plants. Sowing was performed on 4 July, 2016. Basal fertilizer (7.8, 9.6, 9.6, 11.2, and 40 g/m^2 of N, P, K, Mg, and Ca, respectively) was applied to the field before sowing, and 25% more was applied as additional fertilizer on 30 Aug, 2016.

Plants were watered with sprinklers three times per day for 15 minutes each until 9 Aug, 2016. To evaluate the phenotypic variation between different environmental conditions, plants were

divided into two treatments, i.e., well-watered (WW) and less-watered (LW). Since 10 Aug., 2016, plants in the WW treatment group were watered with sprinklers almost once every two days, whereas plants in the LW treatment were watered almost once every four days. Two replications with randomized plots were taken for each treatment, but the places of cultivars were the same between the corresponding replications of treatments (WW1/LW1, WW2/LW2) (Fig. 4-1b and 4-1c).

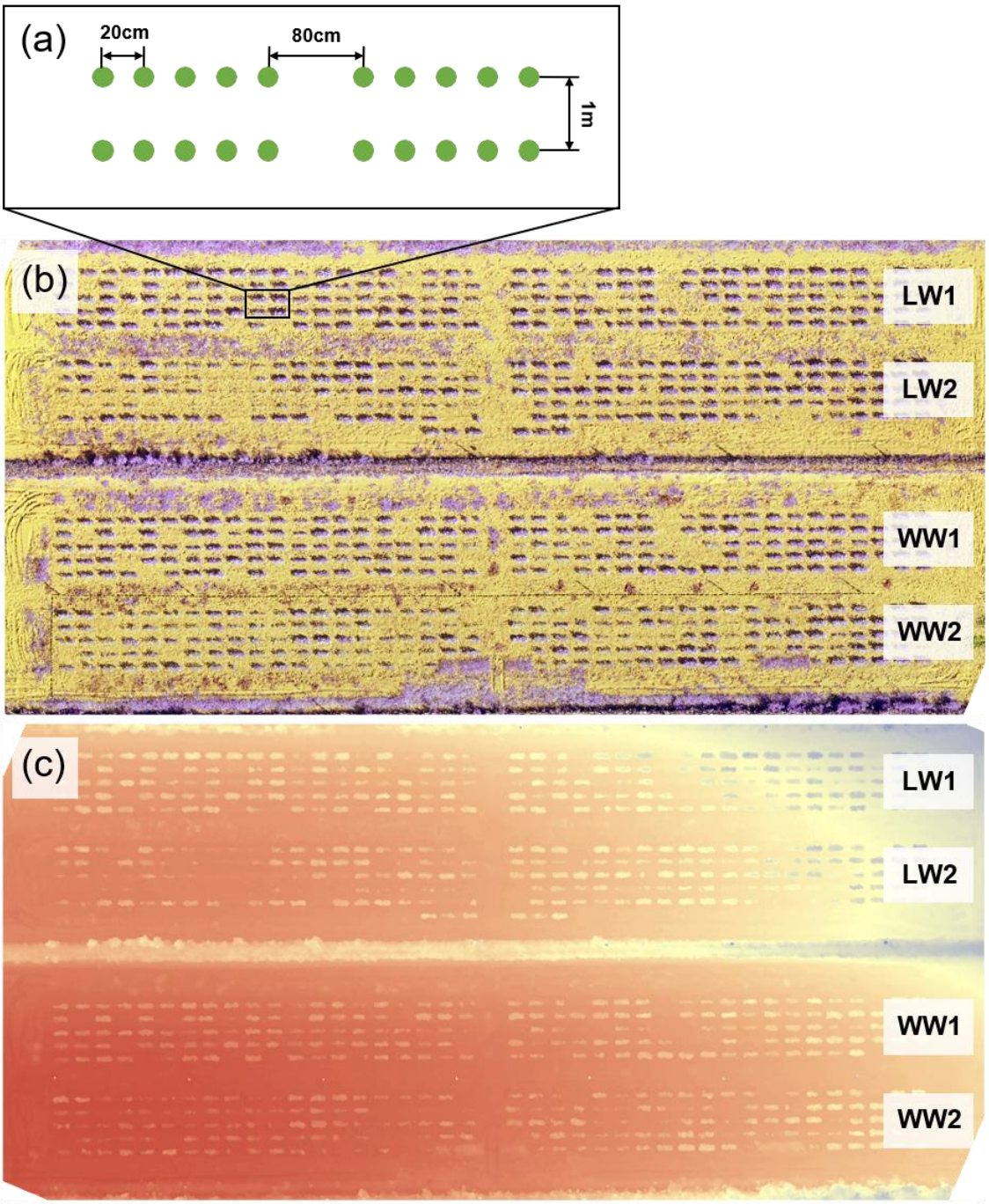

Figure 4-1. The layout of the plots of well-watered (WW) and less-watered (LW) plants and the result of remote sensing. (a) The layout of plots. (b) An ortho-mosaic image (OMI) of the field, obtained on 24 Aug, 2016. The near-infrared (NIR) information was used instead of blue color. (c) Digital surface model (DSM) of the field obtained on 24 Aug, 2016.

4.2.2 Remote sensing and destructive measurements

UAV-RS began two weeks after sowing, and was performed about once every ten days. A Red-Green-NIR (near-infrared) camera (DJI Zenmuse X3 Camera Green-Red-NIR 800–900nm, LDP LLC, USA) mounted on a consumer drone (DJI Inspire 1, Shenzhen, China) was used for the image collection. The UAV flew 12 m above the ground and took images at two second-intervals using an autofocus function. A single UAV flight took about 10 minutes, and collected 500 to 600 images from each treatment region. The height of plants cultivated in the WW1 block was measured manually with rulers on the same day of RS as the ground-truth of the canopy height.

Destructive sampling was conducted from 13 to 15 Sept., 2016, to measure the biomass of the plants (fresh weight of the above-ground part of a plant) to use as a target trait for prediction. Also, images of leaves separated from tillers were taken with a camera (7R, SONY, Japan) with a 35 mm lens (SEL35F28Z, SONY, Japan) to estimate the total leaf area of a plant, and to compare with canopy area measured by UAV-RS. The camera had been remodeled to take a Red, Green, NIR (RGN) image. Images of leaves were taken only for plants cultivated in the WW1 block. The estimated leaf area was compared with the values of the canopy area obtained by RS.

4.2.3 Estimation of traits using RS data

Digital surface models (DSM) and ortho-mosaic images (OMI) of the field on each observation date were constructed (Fig. 4-1b and 4-1c) using the images collected by UAV-RS and Pix4Dmapper software (Pix4D, Switzerland). Constructions of 3D data failed on some dates due to the difficulty in finding matching points of reference between several images. Individual plots were segmented out from the DSM and OMI using geolocation information. Then, regions of the canopies and the ground were segmented out from the OMI of an individual plot via the thresholding of the OMI based on NDVI. Finally, the canopy height and canopy area of each plot were estimated based on the DSM and OMI, respectively. The canopy height was estimated as the difference in mean values of ground elevation and the top elevation of the canopy in the DSM. The canopy area was estimated as the area of the canopy projected in the OMI. Python 2.7 (https://www.python.org) was used for image analysis, and R 3.5.2 (R Core Team, 2020) was used for the other analyses below.

Since canopy height measured with the image analysis had such a large bias that longitudinal growth patterns were hardly observable, it was corrected using the manually-measured plant height of a subset of plots. It was assumed that canopy height measured by UAV was modeled as

$$\mathrm{CH}_{ij} = a_j \mathrm{PH}_{ij} + e_{ij} \quad (4\text{-}1),$$

where CH_{ij} is the canopy height measured by the UAV of the i^{th} plot on the j^{th} observation date, PH_{ij} is plant height measured manually, e_{ij} is the error of estimation of the i^{th} plot on the j^{th} observation date, and a_j is a bias in the estimation of canopy height due to error inherent in the construction of the DSM. The bias a_j was estimated based on the regression, and then obtained

canopy height calibrated by the bias, i.e., CH_{ij}/a_j. The accuracy of calibrated canopy height was evaluated with root mean squared errors (RMSE) between calibrated canopy height and plant height, and RMSE divided by the mean of calibrated canopy height (ratio of RMSE to mean values, RRMSE). The calibrated canopy height was used as the phenotypic value in subsequent analyses. As plant height was not measured manually around 5 or 6 Sept., 2016, the correction of canopy height measured by UAV was not performed for data collected on those dates. Also, canopy height observed on 10 Aug., 2016 was removed in the following analysis due to lack of reliability (Fig. 4-2).

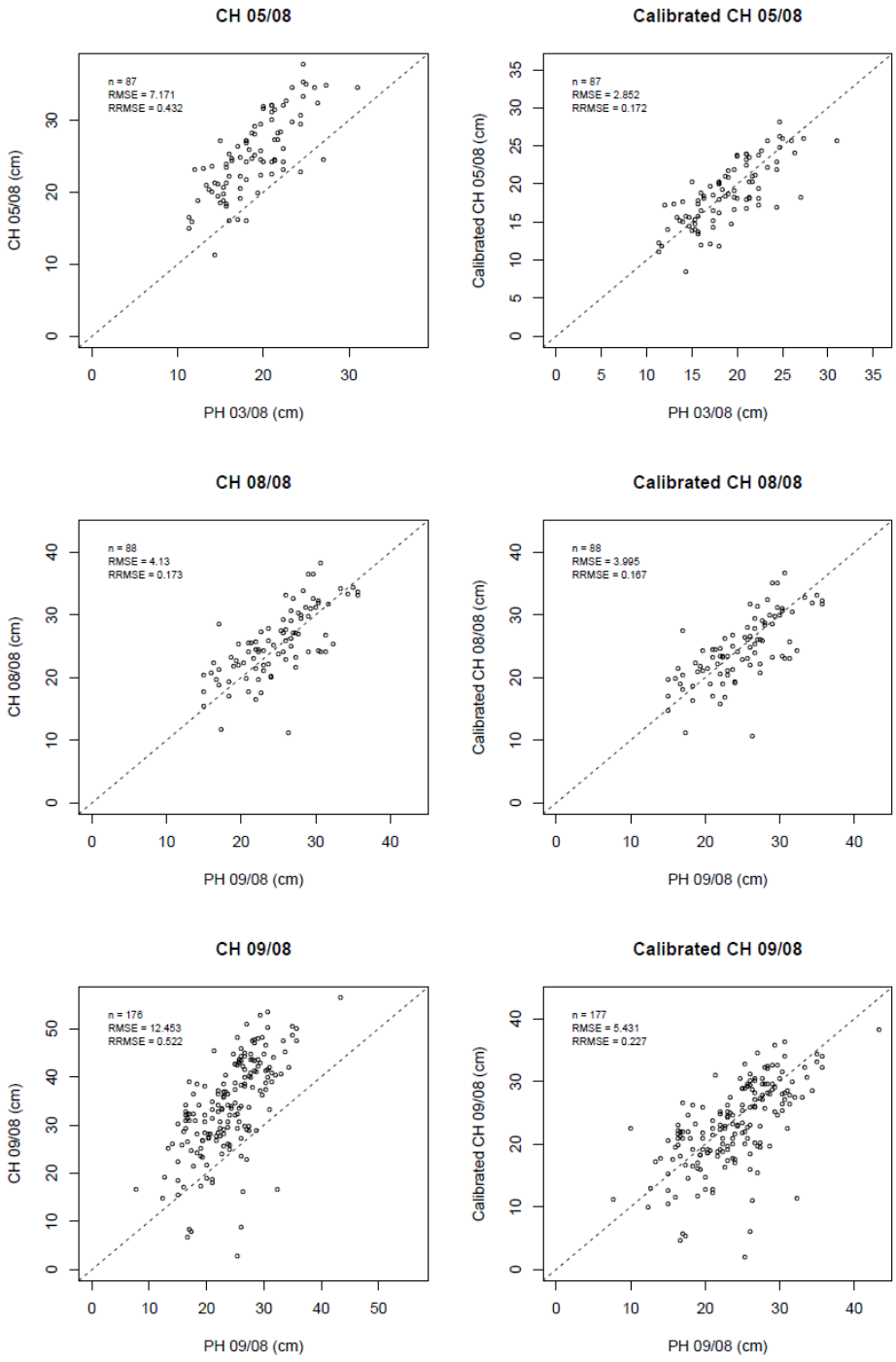

Figure 4-2-1. Comparison of the canopy height, as measured by UAV-RS, and plant height, measured manually. Non-calibrated and calibrated canopy heights are plotted in the left and right panels, respectively. The plant height of the closest date to the UAV-RS was chosen to be plotted. The plant height of the closest day was chosen to plot. The number of data points (n), RMSE (root mean square error), and RRMSE (the ratio of RMSE to the mean) are indicated in the figure.

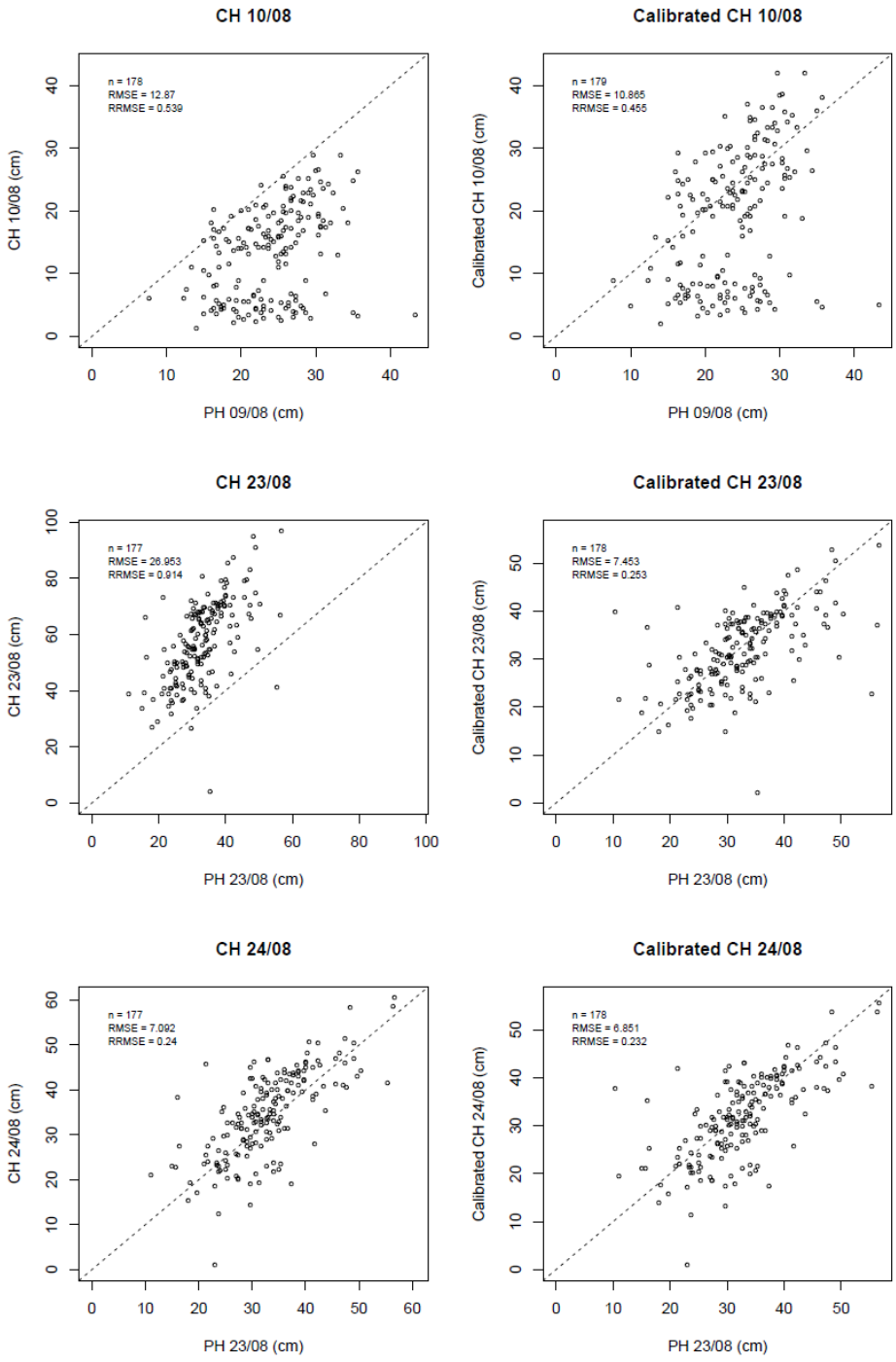

Figure 4-2-2. Comparison of the canopy height, as measured by UAV-RS, and plant height, measured manually. Non-calibrated and calibrated canopy heights are plotted in the left and right panels, respectively. The plant height of the closest date to the UAV-RS was chosen to be plotted. The number of data points (n), RMSE (root mean square error), and RRMSE (the ratio of RMSE to the mean) are indicated in the figure.

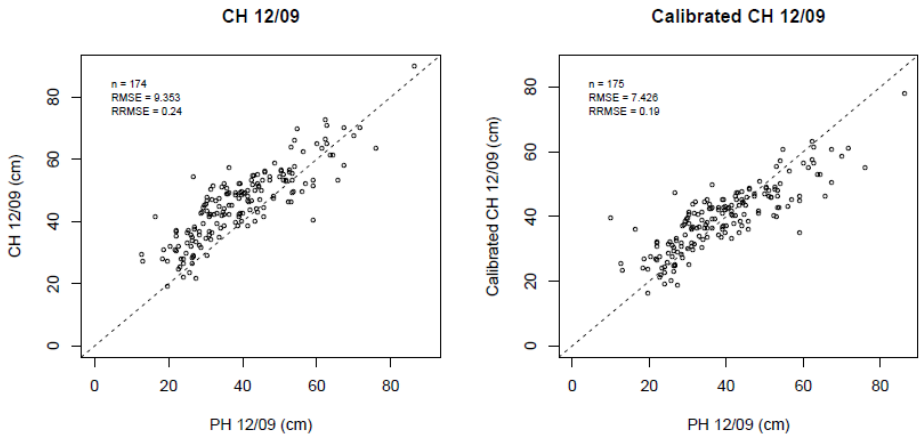

Figure 4-2-3. Comparison of the canopy height, as measured by UAV-RS, and plant height, measured manually. Non-calibrated and calibrated canopy heights are plotted in the left and right panels, respectively. The plant height of the closest date to the UAV-RS was chosen to be plotted. The number of data points (n), RMSE (root mean square error), and RRMSE (the ratio of RMSE to the mean) are indicated in the figure.

4.2.4 Growth model of longitudinal traits

To model growth patterns of canopy height and area, segmented regression lines were fitted to the longitudinal data. The cumulative temperature after the sowing date, or heat unit, was used as a dependent variable. The segmented regression model for the canopy height was

$$y_{ij} = \begin{cases} k_i \mathrm{HU}_j & if\ \mathrm{HU}_j < \mathrm{GTP}_i \\ \mathrm{ymax}_i & if\ \mathrm{HU}_j \geq \mathrm{GTP}_i \end{cases} \quad (4\text{-}2),$$

where y_{ij} is the canopy height of the i^{th} plot on the j^{th} observation date; HU_j is the heat unit on the j^{th} observation date; k_i is the growth speed of the canopy height of the i^{th} plot; GTP_i is the value of the heat unit at the growth termination point of the i^{th} plot; ymax_i is the maximum value of the canopy height of the i^{th} plot. The segmented regression model for the canopy area is

$$y_{ij} = \begin{cases} 0 & if\ \mathrm{HU}_j < \mathrm{GSP}_i \\ k_i(\mathrm{HU}_j - \mathrm{GSP}_i) & if\ \mathrm{GSP}_i \leq \mathrm{HU}_j < \mathrm{GTP}_i \\ \mathrm{ymax}_i & if\ \mathrm{HU}_j \geq \mathrm{GTP}_i \end{cases} \quad (4\text{-}3),$$

where y_{ij} is the canopy area of the i^{th} plot on the j^{th} observation date, GSP_i is the heat unit at the growth start point of the i^{th} plot, and the other variables are the same as those in Eq. 4-2. Fig. 4-3 illustrates the models fitted to the data of the WW treatment group.

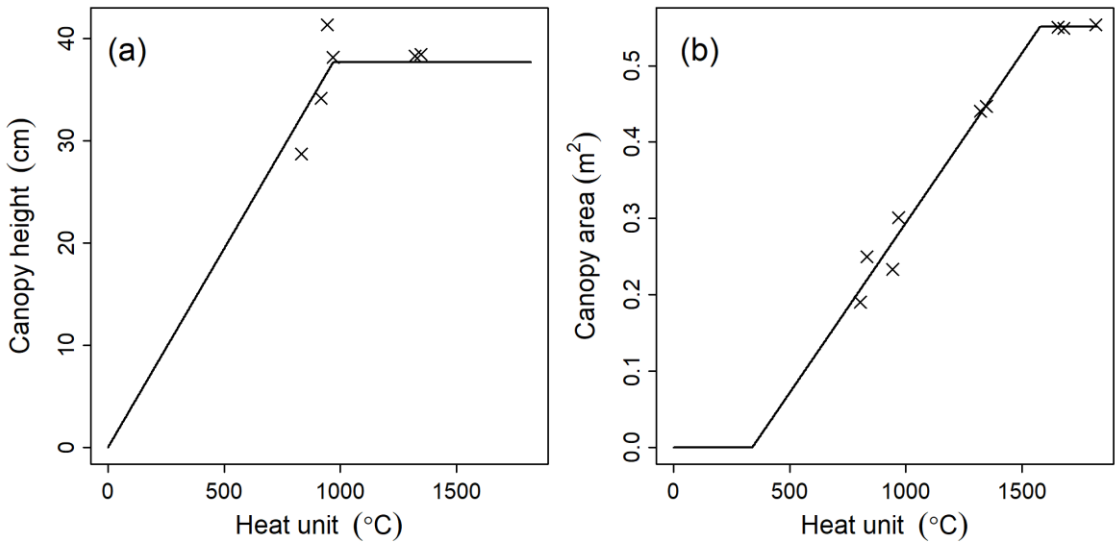

Figure 4-3. Examples of growth models used. The canopy height and area of a plot in the WW2 block were plotted as crosses. The solid lines represent the fitted segmented regression lines of the growth models.

4.2.5 Genotypic values of traits and parameters

Genotypic values of all traits (biomass, canopy height, leaf area, canopy height, and canopy area) and growth parameters were estimated for use in genomic prediction. The following mixed model was fitted for each trait and treatment:

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{L\beta} + \mathbf{Qs} + \mathbf{e} \quad (4\text{-}4),$$

$$\mathbf{g} = \mu\mathbf{1} + \mathbf{s} \quad (4\text{-}5),$$

where $\mathbf{y}$ is a vector of the phenotype, μ is a mean, $\mathbf{1}$ is a vector in which all the elements are one, $\boldsymbol{\beta}$ is a vector of block effect representing differences between replications, $\mathbf{s}$ is a vector of random effect of genotype (with the assumption that was yielded from the Gaussian probability distribution $N(\mathbf{s} \mid \mathbf{0}, \sigma_u^2\mathbf{I})$, where σ_u^2 is a genetic variance and $\mathbf{I}$ is an identity matrix), $\mathbf{e}$ is a vector of residuals (with the assumption $N(\mathbf{e} \mid \mathbf{0}, \sigma_e^2\mathbf{I})$ where σ_e^2 is a residual variance), $\mathbf{L}$ and $\mathbf{Q}$ are design matrices, and $\mathbf{g}$ is a vector of the genotypic values. The R package lme4 (ver. 1.1-20) was used to solve for Eq. 4-4. For canopy height and canopy area, Eq. 4-4 was applied separately for each observation date.

4.2.6 Genomic relationship matrix and heritability

Whole-genome sequencing data of all 198 accessions were available (Kanegae et al., manuscript submitted for publication). This genotyping data identified of 4,776,813 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). A genomic relationship matrix $\mathbf{G}$ was estimated as $\mathbf{G} = \mathbf{X}\mathbf{X}^{\mathrm{T}} / c$, 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 & Jannink, 2012). Then, the genetic heritability was estimated for all traits with the genomic best linear unbiased prediction (G-BLUP) model,

$$\mathbf{g} = m\mathbf{1} + \mathbf{Zu} + \boldsymbol{\varepsilon} \quad (4\text{-}6),$$

where $\mathbf{g}$ is a vector of genotypic values estimated with Eq. 4-4 and 4-5, m is a mean value, $\mathbf{u}$ is a vector of random genetic effect which follows $N(\mathbf{u} \mid 0, \sigma_u^2\mathbf{G})$, $\boldsymbol{\varepsilon}$ is a vector of residuals which follows $N(\boldsymbol{\varepsilon} \mid \mathbf{0}, \sigma_\varepsilon^2\mathbf{I})$), and $\mathbf{Z}$ is a design matrix. The R package rrBLUP (version 4.6) (Endelman, 2011) was used to solve for Eq. 4-6. After solving this mixed model, genomic heritability was estimated as $h^2 = \sigma_g^2 / (\sigma_g^2 + \sigma_\varepsilon^2)$.

4.2.7 Genomic models predicting biomass

Three types of models were constructed to predict biomass (Fig. 4-4): GP, MGP, and TGP models. The GP model was the same as the model shown in Eq. 4-6, where only the data of biomass was used in the prediction.

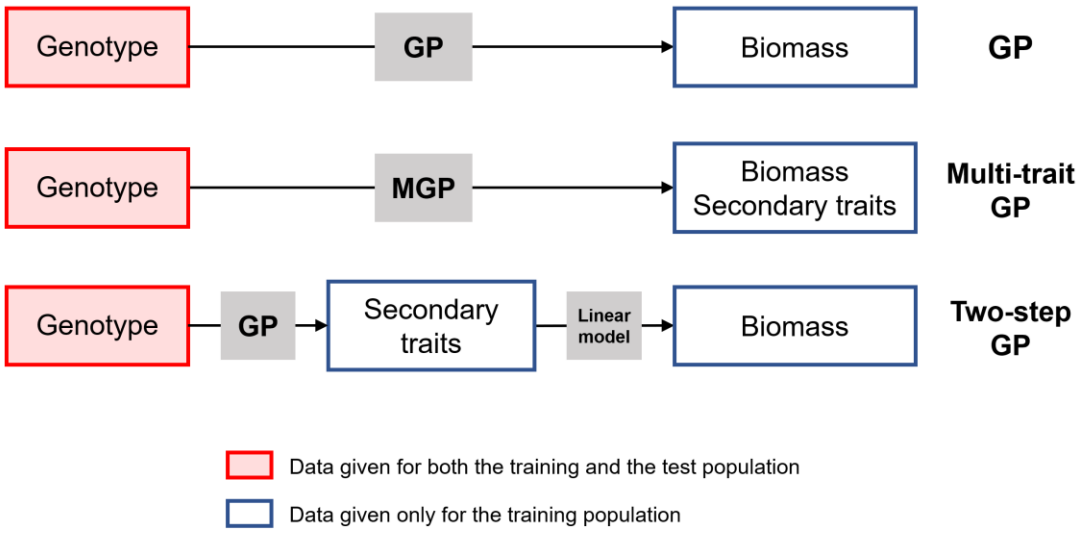

Figure 4-4. Comparison of the structures of three prediction models: genomic prediction (GP), multivariate GP (MGP), and two-step GP (TGP) models. The relationship between genotype data, secondary traits (the traits measured by UAV-RS or growth parameters), and biomass are shown.

The MGP model was an extension of the GP model, and attempted to explain the genetic values of several traits simultaneously (Calus & Veerkamp, 2011; Jia & Jannink, 2012). If secondary traits, which are involved in a model along with biomass, are both highly correlated with biomass and demonstrate high heritability, incorporation of these traits is expected to improve the prediction accuracy of biomass (Calus & Veerkamp, 2011). This model can be expressed as

$$\begin{pmatrix}\mathbf{y}_1\\ \vdots\\ \mathbf{y}_D\end{pmatrix}=\begin{pmatrix}\boldsymbol{\mu}_1\\ \vdots\\ \boldsymbol{\mu}_D\end{pmatrix}+\begin{pmatrix}\mathbf{Z}_1 & \cdots & \mathbf{0}\\ \vdots & \ddots & \vdots\\ \mathbf{0} & \cdots & \mathbf{Z}_D\end{pmatrix}\begin{pmatrix}\mathbf{g}_1\\ \vdots\\ \mathbf{g}_D\end{pmatrix}+\begin{pmatrix}\mathbf{e}_1\\ \vdots\\ \mathbf{e}_D\end{pmatrix} \quad (4\text{-}7),$$

where D is the number of traits in the model, and $\mathbf{y}_d$, $\boldsymbol{\mu}_d$, $\mathbf{g}_d$, and $\mathbf{e}_d$ are vectors of genotypic values, mean values, random genetic effects, and residuals of d^{th} trait, respectively. Assumptions for random effects were included in the model: $(\mathbf{g}_1{}^{\mathrm{T}}, \ldots, \mathbf{g}_D{}^{\mathrm{T}})^{\mathrm{T}} \sim \mathrm{N}(\mathbf{0}, \mathbf{A} \otimes \mathbf{H})$ and $(\mathbf{e}_1{}^{\mathrm{T}}, \ldots, \mathbf{e}_D{}^{\mathrm{T}})^{\mathrm{T}} \sim \mathrm{N}(\mathbf{0}, \mathbf{I} \otimes \mathbf{R})$, where $\mathbf{H}$ is a genomic variance-covariance matrix between traits, and $\mathbf{R}$ is a residual variance-covariance matrix between traits. The R package EMMREML was used to solve Eq. 4-7, in which the high-speed approximation algorithm EMMA (Kang et al., 2008) was used. Two models, one in which the measured values of canopy height and area were used as secondary traits (MGPuav), and the other in which the growth parameters estimated by model (2, 3) were used (MGPpar), were tested. The observed values of secondary traits in the training population were used to build a prediction model, while the observed values of secondary traits in the test population were not used in the prediction of biomass in the test population.

The TGP model used a two-step approach. In the first step, intermediate traits were predicted using GP. In the second step, biomass was calculated from predicted intermediate traits using multiple regression analysis. Genetic values of biomass and those of intermediate traits in the training population were used to estimate coefficients of multiple regression, whereas the values of intermediate traits in the test population predicted by GP were used to predict values of biomass in the test population. Note that only measured values of intermediate traits of the training sets were used; phenotypic values of intermediate traits were not used in the prediction for the test set by replacing them with genomic predicted values. To compare to the MGP model, two cases, one in which the observed values of canopy height and area were used as the intermediate traits (TGPuav), and one in which the growth parameters estimated by model (2, 3) were used (TGPpar), were tested.

Ten-fold cross-validation was repeated 10 times to validate the prediction accuracies of the prediction models. Prediction accuracies were evaluated using correlation coefficients of the genetic values from Eq. 4-4 and predicted values.

4.3 Results

4.3.1 Estimation accuracy of crop traits by UAV-RS

The estimation bias in canopy height measured by the image processing was so large that the longitudinal tendency of growth could not be tracked (Fig. 4-5a). However, the measured values of canopy height were correlated with the values of plant height measured manually on each observation date (Fig. 4-2). Upon correction of the canopy height using the manually-measured plant height and model (1), the estimation bias decreased, as shown by RMSE and RRMSE values (Table 4-1), and the growth processes were then able to be tracked (Fig. 4-5b), accepting only those results obtained on 10 Aug. 2016.

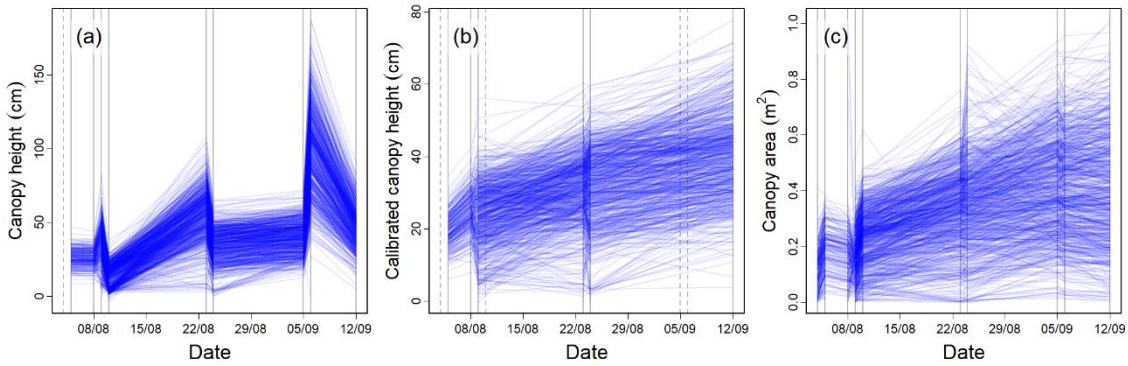

Figure 4-5. Longitudinal patterns of the canopy height and canopy area. (a) Canopy height obtained from the image processing, (b) canopy height corrected with plant height manually measured, (c) and canopy area. Vertical lines indicate the date of UAV-RS; solid lines indicate the dates when data were available, whereas dashed lines indicate the dates when data were unavailable due to failure in the construction of digital surface models (DSM) in Pix4Dmapper.

Compared with the height, the growth pattern in the canopy area was clear without correction (Fig. 4-5c). The canopy area as measured by UAV-RS at harvest and the leaf area as measured in the destructive sampling demonstrated a significant correlation (adjusted $R^2 = 0.540$) (Fig. 4-6).

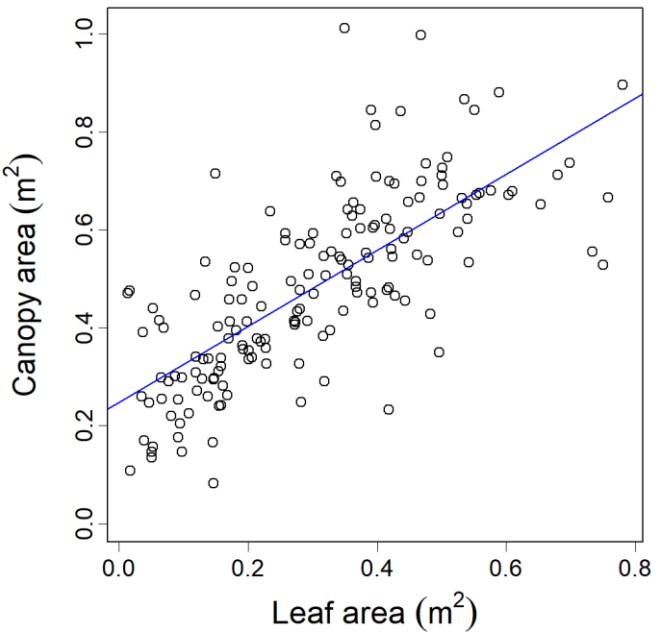

Figure 4-6. Comparison of canopy area and leaf area. The values of the leaf area of one plant are plotted on the horizontal axis, whereas the values of the canopy area of one plot are plotted on the vertical axis. The blue line represents a regression line of canopy area to leaf area.

Date of UAV-RS	Date of manual measurement	Number of samples	CH		Calibrated CH	
			RMSE	RRMSE	RMSE	RRMSE
5 Aug.	3 Aug.	87	7.17	0.432	2.85	0.172
8 Aug.	9 Aug.	180	3.72	0.157	3.75	0.159
9 Aug.	9 Aug.	321	13.43	0.569	6.58	0.279
10 Aug.	10 Aug.	357	12.58	0.539	12.05	0.516
23 Aug.	23 Aug.	177	26.95	0.914	7.45	0.253
24 Aug.	23 Aug.	177	7.09	0.240	6.85	0.232
12 Sept.	12 Sept.	671	10.65	0.276	8.53	0.221

Table 4-1. Estimation accuracy of canopy height. The values of canopy height obtained from image processing (CH) and canopy height calibrated with the plant height were used to evaluate estimation accuracy. Root mean squared errors (RMSE) and RMSE divided by the mean of CH (RRMSE) are shown. The nearest dates of manual measurement from the dates of UAV-RS were chosen for comparison.

4.3.2 Correlation with biomass and heritability of traits as measured by UAV-RS

Correlation coefficients between canopy height and biomass ranged from 0.314 to 0.697, and those between canopy area and biomass ranged from 0.401 to 0.864 (Table 4-2). Those values tended to increase closer to the date of destructive sampling, and to display higher values in canopy area than in canopy height.

Canopy height and area showed increasing heritability, ranging from 0.188 to 0.853 in canopy height and from 0.000 to 0.854 in the canopy area. The canopy area of the WW treatment group in the early growth stage showed notably low heritability due to the effect of weeds; compared with heritability in biomass (0.383 in the WW treatment group and 0.373 in the LW treatment group), heritability in canopy height and area in the late growth period was higher.

	Correlation with biomass				**Heritability**			
Trait	**Canopy height**		**Canopy area**		**Canopy height**		**Canopy area**	
Treatment	**WW**	**LW**	**WW**	**LW**	**WW**	**LW**	**WW**	**LW**
4 Aug.	-	-	0.444	0.600	-	-	0.000	0.275
5 Aug.	0.314	-	0.401	-	0.223	-	0.000	-
8 Aug.	0.343	-	-	-	0.188	-	-	-
9 Aug.	0.440	0.446	0.467	0.595	0.229	0.327	0.000	0.181
10 Aug.	-	-	0.498	0.606	-	-	0.000	0.156
23 Aug.	0.681	0.685	0.752	0.833	0.371	0.394	0.239	0.231
24 Aug.	0.681	0.691	0.800	0.742	0.355	0.47	0.262	0.178
5 Sept.	-	-	0.844	0.864	-	-	0.681	0.572
6 Sept.	-	-	0.833	0.855	-	-	0.797	0.596
12 Sept.	0.572	0.697	0.829	0.840	0.853	0.675	0.854	0.663
GSP	-	-	−0.237	−0.394	-	-	0.152	0.312
GTP	0.078	0.251	0.496	0.447	0.384	0.399	0.825	0.836
k	0.580	0.593	0.223	0.207	0.333	0.291	0.185	0.000
ymax	0.632	0.725	0.830	0.854	0.851	0.595	0.729	0.606

Table 4-2. Correlation coefficients between canopy height/area and biomass, and genomic heritability of canopy height/area. Observed values at each date and growth parameters as calculated using Eq. 4-2 and 4-3 were used. The values of well-watered and less-watered treatment groups (WW/LW) are shown separately.

4.3.3 Parameters of growth models

The correlation between the phenotypic values measured by UAV-RS and the parameters of growth models showed various patterns in growth parameters (Fig. 4-7). The correlation between y_{max} and the observed values of canopy height and area increased, as did GTP. The correlation between *k* and the observed values was stable over the observed time period.

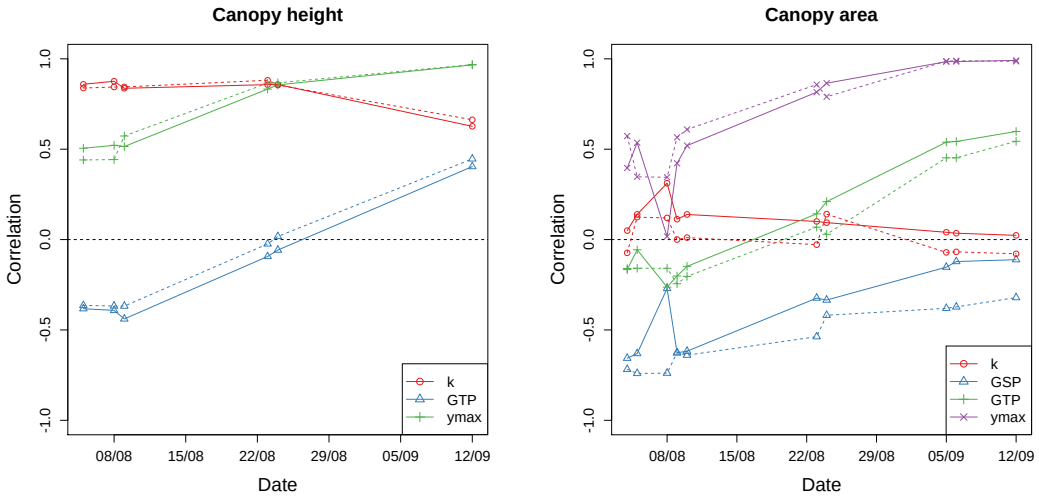

Figure 4-7. Longitudinal pattern of correlation coefficients between observed UAV-RS values and growth parameters. The results of canopy height (left) and canopy area (right) are shown. Solid and dashed lines indicate the results of WW and LW treatments, respectively. Vertical dashed lines indicate the date UAV-RS values were obtained.

The correlation coefficient between y_{max} and biomass was the highest in the parameters of canopy height and canopy area (Table 4-2). Heritability of GTP of canopy height and y_{max} of the canopy area was the highest among all parameters, while *k* of canopy height and GSP or *k* of the canopy area was the lowest.

By comparison with the flowering date, it was found that GTP of both canopy height and area tended to be longer than the number of days to flowering; in other words, termination of growth of canopy height and area were estimated to be later than their flowering (Fig. 4-8).

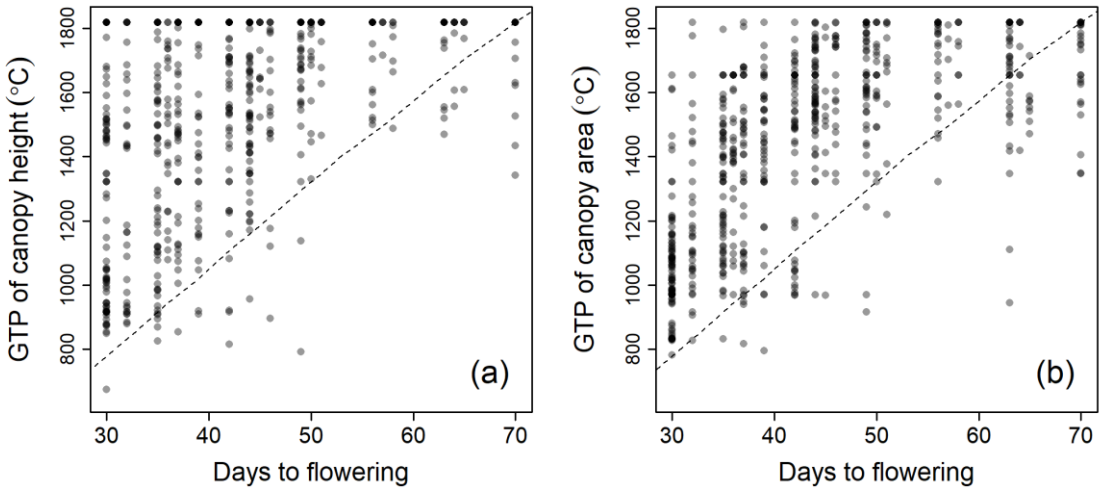

Figure 4-8. Comparison of GTP and days to flowering. Dotted lines indicate the correspondence of the date and heat unit.

4.3.4 Genomic prediction

In the WW treatment group, the prediction accuracy of the MGPuav model was the highest, at 21.2% higher than the GP model. On the other hand, in the LW treatment group, the prediction accuracy of the TGPuav model was the highest at 24.1% higher than the GP model (Fig. 4-9). Accuracies of the models using the growth parameters (MGPpar, TGPpar) were lower than the models with the values calculated using data observed by UAV (MGPuav, TGPuav), excepting the accuracy of the MGP model in the LW treatment group, for which the accuracy of the MGPuav model was notably low.

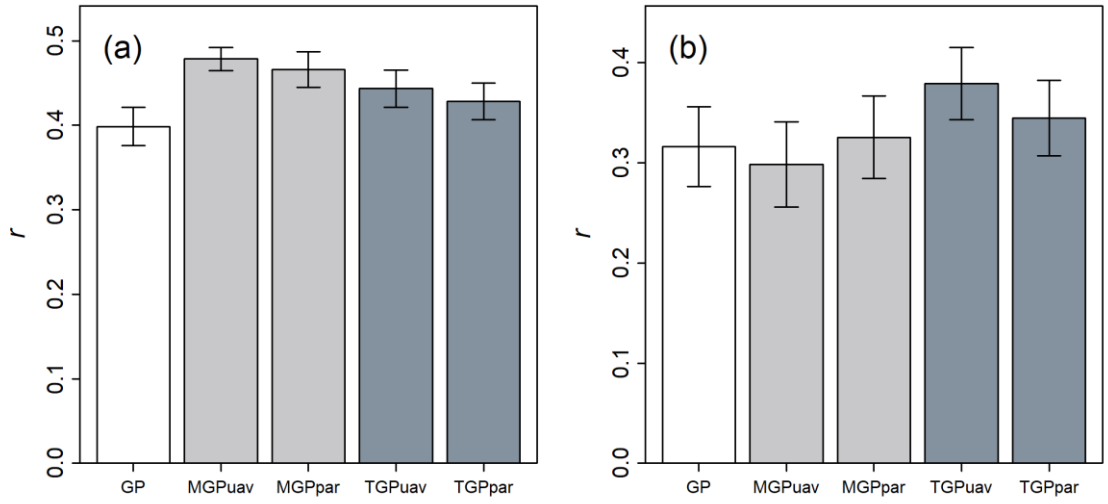

Figure 4-9. Prediction accuracies of biomass. Correlation coefficients of observed and predicted values were plotted. The accuracies of WW and LW treatment groups are shown on the left and right sides, respectively. Error bars indicate ± 1 s.d.

4.4 Discussion

4.4.1 Acquisition of longitudinal growth traits through UAV-RS

A large estimation bias was observed in canopy height due to the effects of wind or change in radiation levels. Such bias, which is inevitable in the field of RS, hinders and prevents the tracking of detailed plant growth process. However, as shown in (Hu et al., 2018) and by this research, calibration using manual measurements can greatly reduce this bias. As plant height is strongly related to biomass (Bendig et al., 2015), such a calibration method can play an essential role in breeding to observe the longitudinal growth of plants.

On the other hand, the canopy area showed lower bias than canopy height, and did not require any correction with manual measurements, suggesting that canopy area can be a robust measurement of temporal growth of soybean plants. Also, the correlation between the canopy area and biomass was high from an early stage of growth (Table 4-2). A similar trait, projected leaf area, was already shown to be highly correlated with biomass in a high-throughput phenotyping platform (Golzarian et al., 2011). Plot-level vegetation indices are widely used in field RS (Jessica Rutkoski et al., 2016; Tattaris et al., 2016; Duan et al., 2017), but plant-level traits such as canopy area are also a suitable trait for the observation of plant conditions, due to their robustness and correlation with biomass.

4.4.2 Growth model

Large differences were observed among the heritability of the growth parameters of the canopy area (Table 4-2). The parameters k and GSP demonstrated low heritability, possibly because these parameters reflected the influence of weeds during the early growth period. On the other hand, high heritability was observed in the GTP and y_{max} parameters, indicating that these parameters reflected the genetic characteristics of growth patterns of the canopy area. y_{max} was especially suitable as a secondary trait in GP, due to its high heritability and correlation with biomass. The growth parameter canopy height showed varied heritability, whereas y_{max} had high heritability and correlation with biomass, but no parameter showed higher heritability and correlation with biomass than those of GTP of canopy area.

It was additionally shown that the GTP of canopy height and area was later than flowering (Fig. 4-8). This result matches a characteristic of soybean, of which the stems and leaves keep growing after the beginning of flowering. The growth models of Eq. 4-2 and 4-3 were found to appropriately extract these growth characteristics from UAV-RS data.

4.4.3 Prediction accuracies of MGP and TGP models

Both MGP and TGP models could improve the prediction accuracy of GP (Fig. 4-9). This improvement was not observed in previous studies in cases where secondary traits were unavailable in the test population (Rutkoski et al., 2016; Sun et al., 2017). One reason that these models improved accuracy in this study is the high heritability of the secondary traits, especially

the canopy area. It was reported that the MGP model could improve the prediction accuracy if secondary traits were highly correlated with the target trait and also display high heritability (Calus & Veerkamp, 2011); canopy area was fould to be a good secondary trait in predicting biomass from this point of view (Table 4-2).

Comparing the results of the MGP and TGP models, the prediction accuracies were generally higher in the MGP model. This is because MGP estimates the random genetic effect of multiple traits simultaneously, whereas TGP does so for each trait separately. However, the accuracy of the MGPuav model in the LW treatment group was particularly low, and the accuracy of the TGPuav model exceeded it. This was due to the failure of estimation of covariance structures in the MGPuav model, which was suggested by the lower heritability of biomass in the MGPuav model than in the G-BLUP of Eq. 4-6 (Fig. 4-10). An increase in the number of secondary traits can also cause such estimation failures. On the other hand, the prediction accuracy of the TGP model was assumed to be stable, because it does not require complex computation. TGP is a robust approach to incorporate an increase in the number of secondary traits.

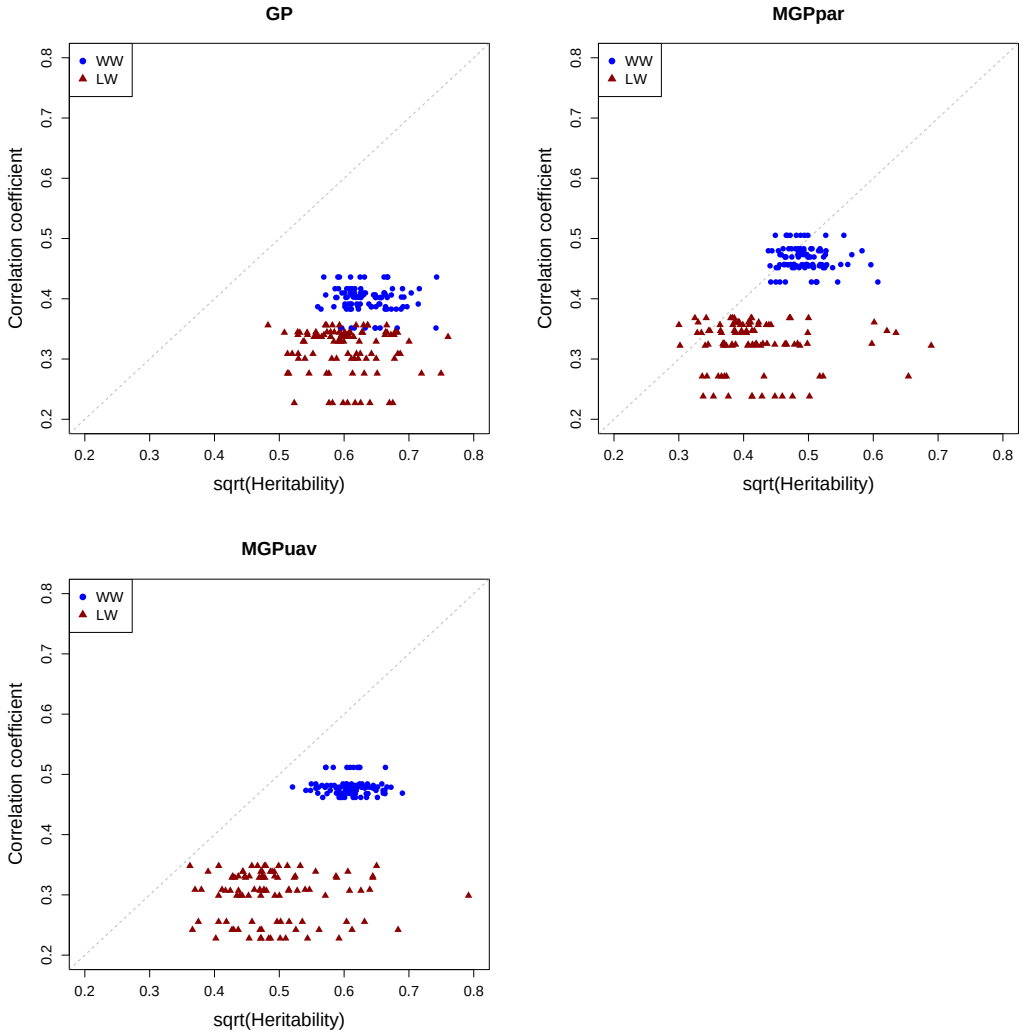

Figure 4-10. Comparison of the heritability (x-axis) and the prediction accuracy (y-axis) of biomass for each prediction model. Results of GP, MGPuav, MGPpar, TGPuav, and TGPpar are shown. Blue and red points indicate WW and LW treatments, respectively.

Linear regression was used in the second step of the TGP model in this study, but other statistical approaches can alternatively be used (Chapter 3). If the correlation among the intermediate traits is high, partial least squares regression or principal component regression are recommended for the second step of TGP to avoid a problem with multi-collinearity. Also, non-linear models, such as Random forest (Breiman, 2001), or support vector machines can be used to include a non-linear relationship between the intermediate traits and the target traits in the final model.

4.4.4 Use of growth parameters in the prediction model

The prediction accuracies of the MGPuav and TGPuav models were generally higher than those of the MGPpar and TGPpar models. The reduction in prediction accuracies in the MGPpar and TGPpar models likely occurred because the information in the original RS data was partially lost when growth models were fitted. However, the prediction accuracy of the MGPpar model in the LW treatment group was higher than that of the MGPuav model (Fig. 4-9), although the heritability of the biomass estimated with the MGPpar model was lower than that estimated with the MGPuav model (Fig. 4-10). The MGPuav and MGPpar models could not estimate the covariance matrix in this case, but the high heritability of the growth parameters (Table 4-2) increased the prediction accuracy in the MGPpar model. Although the prediction accuracy of the MGPpar model was not different from that of GP, it indicated that the incorporation of a growth model could effectively compress the longitudinal traits as growth parameters, and improve the predictive ability of the MGP model.

One alternative approach to incorporate longitudinal RS data is a random regression model, which is a statistical method to consider the longitudinal structure of traits in GP. The time-dependent structure of a covariance matrix is included in the model by explaining growth curves with base expansions using arbitrary functions (Kirkpatrick and Heckman, 1989). Several reports have shown that random regression was an effective method when applied to longitudinal RS data (Sun et al., 2017) or growth data measured by a high-throughput phenotyping platform (Campbell et al., 2018). Random regression can be understood as a dimension reduction of growth curves using functions such as Legendre polynomials or spline functions. Such functions have little restriction on their forms, and are applicable to a wide range of growth patterns. Still, even in exchange for flexibility, it is challenging to include prior information about the target growth pattern in random regression. One advantage of the proposed approach using growth models is that we could include prior knowledge about growth directly in the model, which led to the effective extraction of characteristics of the growth curves. Validation of the prediction ability of these methods using various datasets can help in choosing a more appropriate method when applied to real datasets.

Simple growth models were used to express the growth of the canopy height and area, but more complex models such as crop growth models can also be included in the framework. Crop growth models are simulation models where crop yield is expressed as an accumulation of carbon production during the cultivation period. Some reports have applied crop growth models in the framework of GP (Technow et al., 2015). Also, several methods have been developed to combine

crop growth models and longitudinal RS data via data assimilation (Jin et al., 2018; Kasampalis et al., 2018). It is expected that detailed growth characteristics can be estimated from longitudinal RS data using sophisticated models such as crop growth models.

5 Longitudinal growth analysis of soybean using UAV-based remote sensing and its application on genomic prediction

5.1 Introduction

Genetic mechanisms of growth processes have become a crucial topic in plant breeding. Genetic dissection of the formation process of target traits will provide a profound understanding of its mechanism, which will lead to the efficient genetic improvement of the target traits. 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). Most GP studies for crops have focused on traits at harvests, such as yield and quality. 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. Also, the growth prediction in the latter growth period using growth data of the early period will enable the selection in the early growth stages, which leads to the reduction of the cost for field trials.

However, cost- and labor-intensive phenotyping of measuring the longitudinal growth data of traits has been a major bottleneck for the genetic dissection of the formation process of target traits. In particular, it has been difficult to measure data for many genotypes grown in the field. Thanks to the rapid development of sensing technologies in recent years, high-throughput phenotyping has become available in plant breeding, and the measurement of longitudinal growth data is becoming more practical. Accurate and detailed acquisition of growth processes through high-throughput measurement is expected to lead to an improved genetic gain in plant breeding (Furbank & Tester, 2011; Cabrera-Bosquet et al., 2012; Araus & Cairns, 2014). For example, an automated phenotyping platform for measuring time-series three-dimensional plant growth in a greenhouse has enabled the genetic dissection of the growth processes with a longitudinal 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) has become available for measuring plant growth. Until recent years, however, studies on field remote sensing (RS) of plant growth have mainly focused on the application in field management. The application of the field RS to the genetic dissection of the plant growth process has still been limited (Blancon et al., 2019).

Several methods have been proposed to analyze longitudinal growth data. One common method is to fit a growth model function, such as Gompertz (Winsor, 1932) and logistic (Nelder, 1961) curves, to the data and use the parameters of the function for quantifying the pattern of growth. This method can be applied to various types of longitudinal growth data. Various methods of quantitative genetics, such as quantitative trait loci (QTL) analysis (Ma et al., 2002; Wu et al., 2002) and genome-wide association (GWAS) studies (Das, Li, Wang, et al., 2011; Crispim et al., 2015), has been applied to analyze genetic variations in the 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). Its application to GP, however, has not been discussed in previous studies.

In this study, a method integrating a growth model function and GP is proposed and applied to the soybean canopy area's longitudinal growth data. The growth patterns of soybean germplasm accessions were described with five parameters of the fitted growth model. Genetic variations in the growth pattern were quantified by decomposing the parameters into genetic and residual effects with mixed models. Finally, the integrated method of GP and the growth model were applied to the prediction of growth processes in several prediction schemes.

5.2 Material and methods

5.2.1 Field trials

Soybean accessions registered in the National Agriculture and Food Research Organization Genebank (https://www.gene.affrc.go.jp/index_en.php) were employed. From 2017 to 2019, the field trial was conducted in an experimental field with sandy soil at Arid Land Research Center, Tottori University (35°32' N lat, 134°12' E long, 14 m above sea level). One hundred eighty-six accessions were used in 2017, and 198 accessions were used in 2018–2019. Each plot consisted of four plants, and the distances between two rows, two plots, and two individuals were 50 cm, 80 cm, and 20 cm, respectively (Fig. 5-1d). Sowing was performed at the beginning of July, followed by thinning after two weeks (Fig. 5-2). Fertilizer (15, 6.0, 20, 11, 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 genetic variations in the responses to different environmental conditions. White mulching sheets (Tyvek, Dupond, US) were laid to prevent rainwater infiltration (Fig. 5-1b) to control soil conditions with artificial irrigation. Watering tubes were installed under the sheets to irrigate the field. Artificial irrigation was applied for five hours daily (7:00–9:00, 12:00–14:00, 16:00–17:00) starting the day after the thinning in treatment C, but no artificial irrigation was performed in treatment D. In the following text, an abbreviation for denoting a specific combination of the level of the treatment and the year of the experiment is used; e.g., treatment C in 2017 is abbreviated as "2017-C".

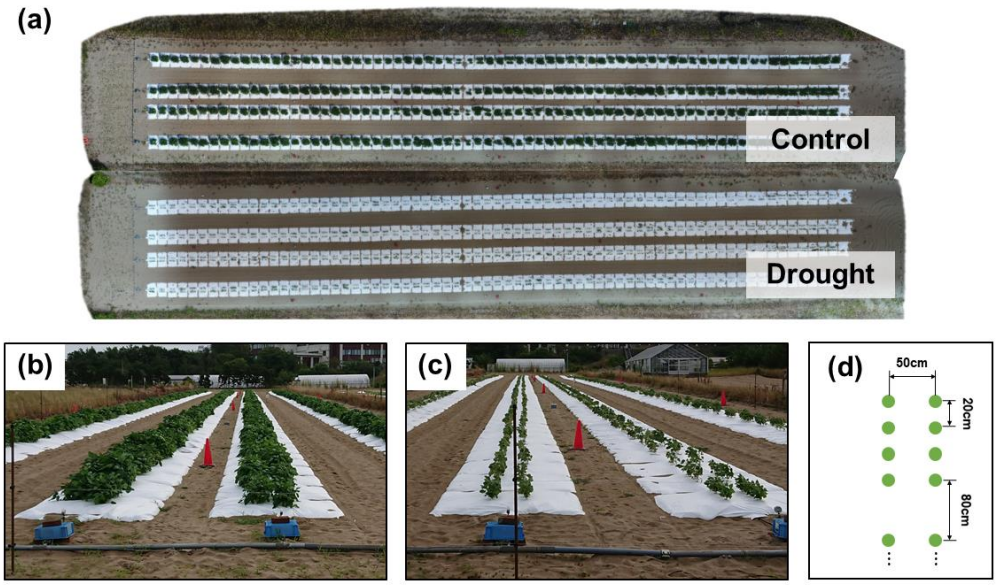

Figure 5-1. Explanation of the field design in the experiment. (a) An ortho-mosaic image (OMI) of the field, obtained on 25 Aug. 2018. The OMIs were created for each treatment and synthesized. (b, c) Photos of treatment C and D were taken from the ground. (d) The layout of plots and plants.

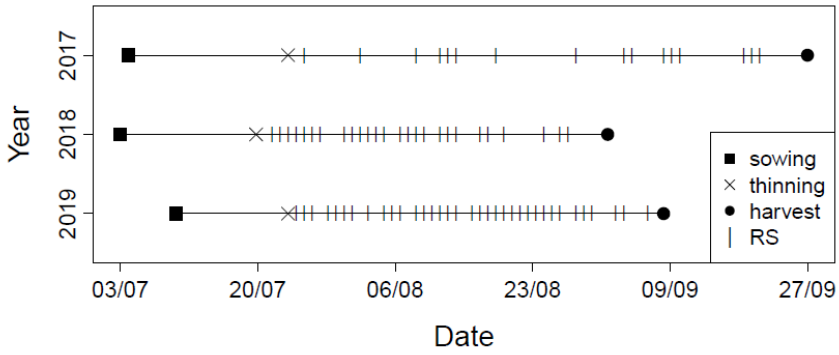

Figure 5-2. Calendar of field management and remote sensing in the field experiment. Dates of sowing, thinning, harvest, and remote sensing (RS) are shown.

5.2.2 Remote sensing and image analysis

UAV-RS started after the thinning and was performed 16–35 times in the cultivation period (Fig. 5-2). A consumer drone (DJI Phantom 4 Advanced, Shenzhen, China) was used for the RGB image collection. The UAV flew 12–14 m above the ground and took images with an interval of two seconds with autofocus function. A single UAV flight took about 15 minutes and collected 500 to 600 images for each treatment.

Digital surface models (DSM) and ortho-mosaic images (OMI) of the field of each UAV-RS were constructed using the images corrected in the UAV-RS and Pix4Dmapper (Pix4D, Switzerland). The geolocation information was used to segment individual plots from the DSM and OMI. Next, NDVI-based thresholds were used to segment the canopy and ground regions from the individual plots' OMI. Finally, the canopy area of each plot was estimated based on the DSM and OMI, respectively. The canopy area was estimated as the area of the canopy projected onto OMI. Python 2.7 (https://www.python.org) was used for the image analysis. For data in 2019, a similar procedure was used by Hiphen inc. (France).

5.2.3 Growth-process modeling

A growth model (Koetz et al., 2005) was applied to the canopy area's time-series measurements to quantify each plot's growth pattern with a small number of parameters. R (ver. 4.0.3) (R Core Team, 2020) was used for the analyses. By fitting the model to observed canopy area values, parameters were estimated for each plot. In this model, the leaf area index (LAI) on day d after sowing (LAI_d) is expressed as follows (Fig. 5-3):

$$\mathrm{LAI}_d = \mathrm{LAI}_{\mathrm{amp}} \left\{ \frac{1}{1 + \exp\left(-r_g\left(\mathrm{HU}_d - \mathrm{T}_g\right)\right)} - \exp\left(r_s\left(\mathrm{HU}_d - \mathrm{T}_s\right)\right) \right\} \qquad (5\text{-}1).$$

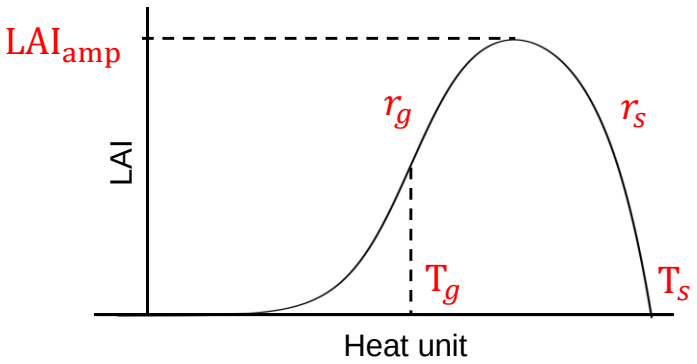

Figure 5-3. An explanation of the growth model Eq. 5-1.

The first and second terms in the parenthesis represent the logistic growth and exponential senescence, respectively. HU_d is a heat unit on day d. The heat unit is an indicator of the plant growth calculated as the cumulation of daily mean temperature adjusted by base temperature after the sowing date. The base temperature was set to 8°C in this study (Soltani & Sinclair, 2012). Thus,

$$\mathrm{HU}_d = \sum_{i=1}^{d} (\mathrm{Temp}_i - 8°\mathrm{C}) \quad (5\text{-}2),$$

where Temp_i is the daily mean temperature on the day i. LAI_{amp}, r_g, r_s, T_g, T_s in Eq. 5-1 are the parameters of this model, estimated for each plot. LAI_{amp} is the maximum value of the LAI in the growth curve; r_g is growth speed rate; r_s is senescence speed rate; T_g is heat unit when growth speed reaches the maximum; T_s is heat unit when LAI becomes zero, respectively. LAI values on day d can be converted to the canopy area (CA_d),

$$\mathrm{CA}_d = S\{1 - \exp(-k\ \mathrm{LAI}_d)\} \quad (5\text{-}3).$$

Eq. 5-3 is based on a famous equation of the relationship between LAI and the amount of light interception (Soltani & Sinclair, 2012). k is an extinction coefficient that is known to differ among plant species. In this model, the value of commonly used soybean, $k = 0.5$, was assigned (Soltani & Sinclair, 2012). S is the area for each plot ($S = 1.68$ m^2).

The five parameters, LAI_{amp}, r_g, r_s, T_g, and T_s, should be estimated for each plot. However, fitting the growth model to each plot's data was difficult because the canopy area in treatment D was so small that the noise disturbed the search for the best parameter sets. Therefore, the estimation was conducted along with the following steps.

1. **Estimation of genotype-specific values**. For each genotype (an accession in the evaluated collections), the same value of a parameter was applied to the time-series canopy area data of all plots, except LAI_{amp}, which is assumed to be different among plots. The optimal values of T_g and T_s were found with a grid search. T_g and T_s's search ranges were set to (300, 1200) and (1400, 3000), respectively, and seven points were chosen in equal intervals as grid points of each parameter. At the same time, the optimal values of the other parameters were estimated with the Nelder-Mead method.

2. **Estimation of plot-specific values**. The parameter estimation was conducted for each plot. A grid search was used to optimize the values of T_g and T_s. In the grid search, the T_g and T_s's optimized values obtained in Step 1 were used to determine the center of the searching spaces, and its range was narrower than Step 1 (estimated values ± 200 for T_g, estimated values ± 400 for T_s). The other parameters were estimated with the Nelder-Mead method, using estimated values in Step 1 as the initial values.

Usually, the minimization of the sum of squares is used to fit the growth model. However, in the canopy area, the variances of measurement noises showed heterogeneity and were

proportional to each day's mean values (Fig. 5-4). Therefore, in the estimation of parameters, the adjusted sum of squares was minimized:

$$\sum_d \sum_i \frac{\left(y_{i,d} - \hat{y}_{i,d}\right)^2}{\bar{y}_d} \quad (5\text{-}4),$$

where $y_{i,d}$ is the canopy area of plot i on day d, $\hat{y}_{i,d}$ is the fitted value of the canopy area with the growth model, and $\bar{y}_d$ is the mean value of the canopy area on day d. In Fig. 5-4, the variances of noise were estimated with the fitted values of the growth model: $\sum_i \left(y_{i,d} - \hat{y}_{i,d}\right)^2 / N$ where N is the number of plots.

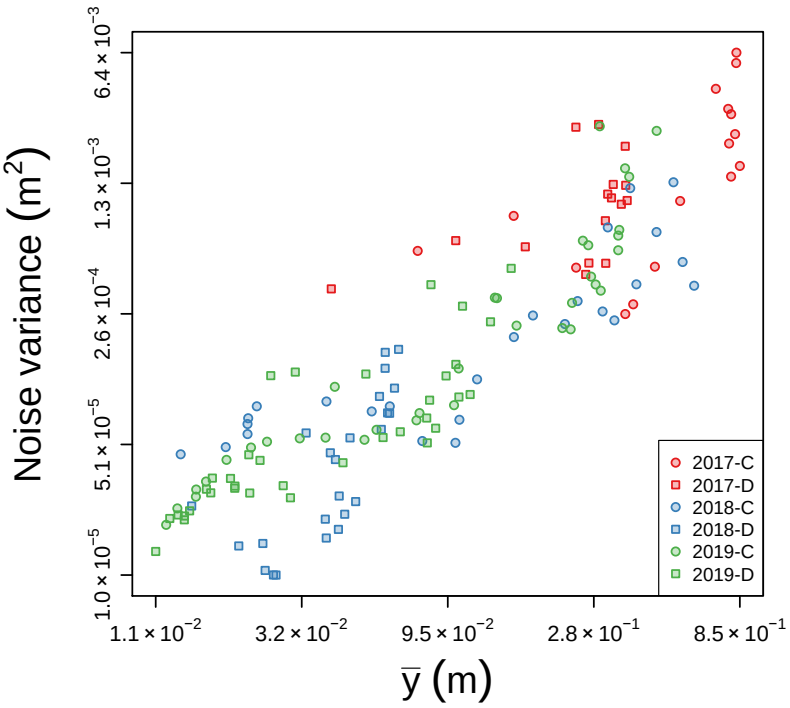

Figure 5-4. Scatterplot of the noise variances on the x-axis and mean values ($\bar{y}_d$) of the canopy area on the y-axis. Noise variances were estimated with the fitted values of the growth model: $\sum_i \left(y_{i,d} - \hat{y}_{i,d}\right)^2 / N$ where N is the number of plots. The logarithmic scale was used for both axes.

5.2.4 Estimation of genotypic values

Genotypic values of the canopy area and growth parameters were estimated for use in GP. For each combination of a trait (canopy area or a growth parameter) and a treatment (C or D), the following mixed model was fitted (cf. Eq. 2-2):

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{L\beta} + \mathbf{Qs} + \mathbf{e} \quad (5\text{-}5),$$

where $\mathbf{y}$ is a vector of the phenotypic values, μ is a mean, $\boldsymbol{\beta}$ is a vector of block effect representing differences between replications, $\mathbf{s}$ is a vector of genotypic values which follows $N(\mathbf{s} \mid \mathbf{0}, \sigma_s^2\mathbf{I})$, σ_s^2 is a genotypic variance, $\mathbf{e}$ is a vector of residuals which follows $N(\mathbf{e} \mid 0, \sigma_e^2\mathbf{I})$, σ_e^2 is a residual variance, $\mathbf{1}$ is a vector in which all the elements are one, $\mathbf{I}$ is an identity matrix, and $\mathbf{L}$ and $\mathbf{Q}$ are design matrices. The genotypic value ($\mathbf{g}$) was then calculated by (cf. Eq. 2-3)

$$\mathbf{g} = \mu\mathbf{1} + \mathbf{s} \quad (5\text{-}6).$$

The R package lme4 (ver. 1.1-20) was used to solve Eq. 5-5. For the canopy area, the genotypic value estimation was applied separately for each date of UAV-RS.

5.2.5 Genomic relationship matrix and genetic analysis

Whole-genome resequencing data of all 198 accessions were available (Kanegae et al., manuscript submitted for publication). A genomic relationship matrix $\mathbf{G}$ was estimated using the marker genotype data. The detailed information is the same as Section 5.2.6 (Endelman & Jannink, 2012). Then, the genetic heritability was estimated for all traits with the genomic best linear unbiased prediction (G-BLUP) model,

$$\mathbf{g} = m\mathbf{1} + \mathbf{Zu} + \boldsymbol{\varepsilon} \quad (5\text{-}7),$$

where $\mathbf{g}$ is a vector of genotypic values estimated with Eq. 5-5 and 5-6, m is a mean, $\mathbf{u}$ is a vector of random genetic effect which follows $N(\mathbf{u} \mid \mathbf{0}, \sigma_u^2\mathbf{G})$, $\mathbf{e}$ is a vector of residuals which follows $N(\mathbf{e} \mid \mathbf{0}, \sigma_\varepsilon^2\mathbf{I})$, σ_u^2 and σ_ε^2 are 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 Eq. 5-7. After solving the mixed model, genomic heritability was estimated as $h^2 = \sigma_u^2 / (\sigma_u^2 + \sigma_\varepsilon^2)$.

5.2.6 Prediction of the growth process

Three cross-validation schemes were conducted to validate if the estimated growth parameters can improve the growth process prediction (Fig. 5-5a). The first scheme is the cross-validation over genotypes (CV1). In this scheme, the data from a subset of genotypes in any treatments or years were left out of training data. A prediction model built with the training data was validated with the subset of left-out genotypes. The left-out genotypes were randomly selected. In this study, 10-fold cross-validation was employed.

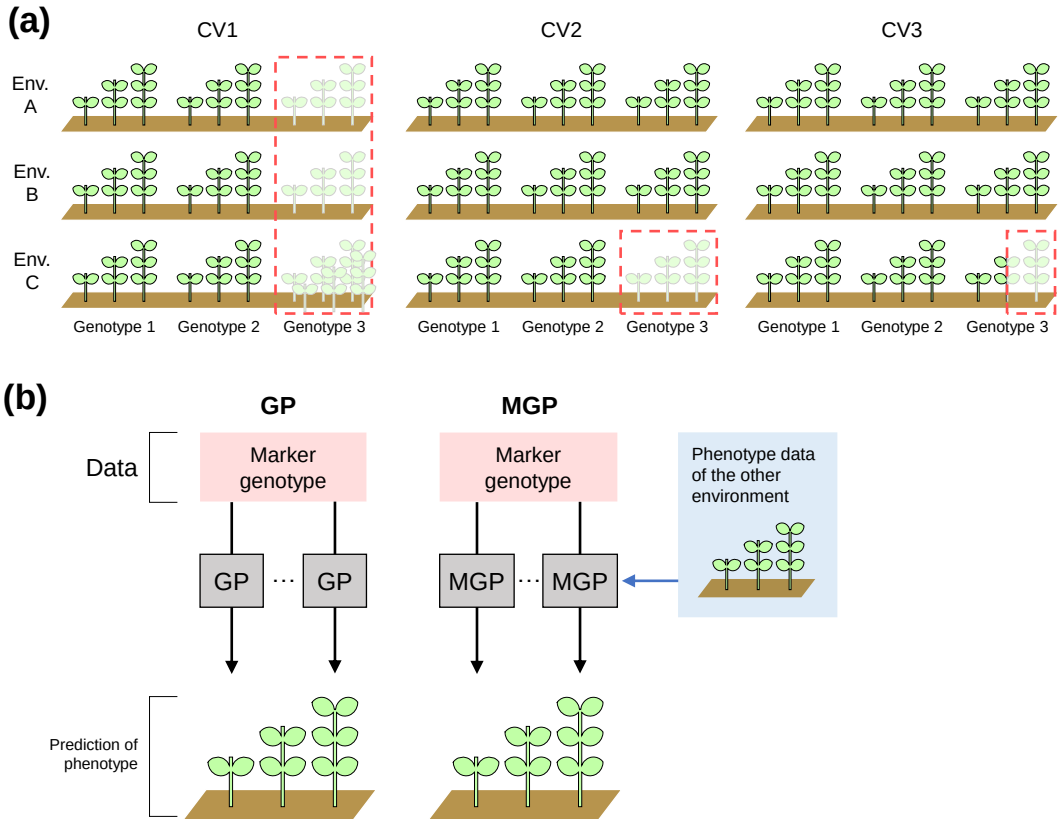

Figure 5-5-1. Visualization of prediction schemes and models. (a) Cross-validation schemes (CV1, CV2, and CV3). Left-out data in cross-validation is highlighted by a rectangle with dashed red lines. (b) Prediction models (GP and MGP).

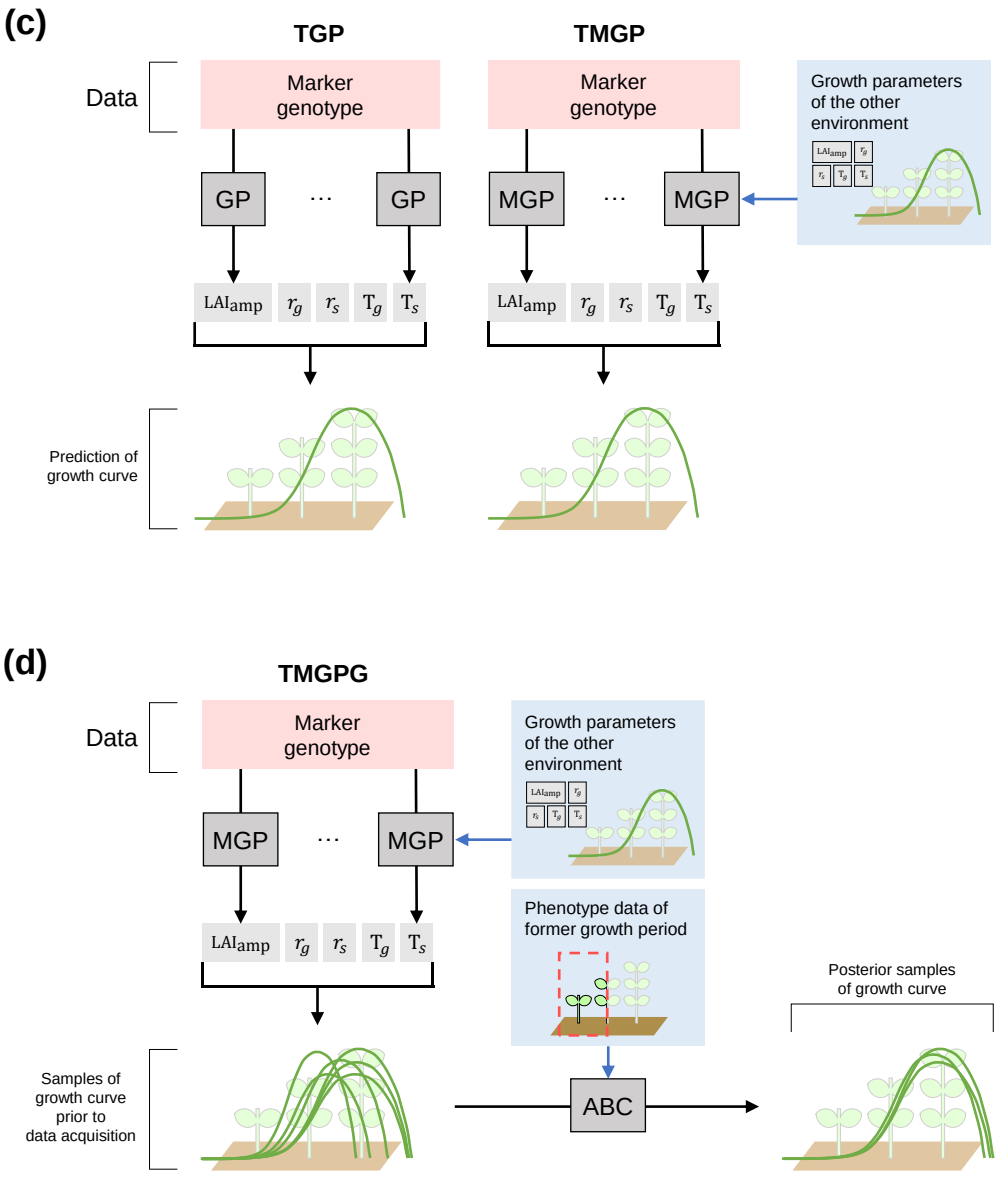

Figure 5-5-2. Visualization of prediction schemes and models. (c) Prediction models (TGP and TMGP). (d) Prediction models (TMGPG).

In CV1, the prediction accuracies of four models were compared (Fig. 5-5b and 5-5c); genomic prediction (GP), two-step GP (TGP), multivariate GP (MGP), and two-step multivariate GP (TMGP). For GP, a mixed model expressed as Eq. 5-7 was applied to the canopy area each day in the training data. Then, the random genetic values **g** of left-out genotypes were used as predicted values.

TGP consisted of two steps. The same GP model (Eq. 5-7) was applied to the growth parameters at first. Then, canopy areas of left-out genotypes on each day were calculated using the growth model and the predicted growth parameters (Eq.5-1 and 5-2).

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

$$\begin{pmatrix}\mathbf{g}_1\\ \vdots \\ \mathbf{g}_K\end{pmatrix} = \begin{pmatrix}m_1\mathbf{1}\\ \vdots \\ m_K\mathbf{1}\end{pmatrix} + \begin{pmatrix}\mathbf{Z}_1 & \cdots & \mathbf{0}\\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{Z}_K\end{pmatrix}\begin{pmatrix}\mathbf{u}_1\\ \vdots \\ \mathbf{u}_K\end{pmatrix} + \begin{pmatrix}\boldsymbol{\varepsilon}_1\\ \vdots \\ \boldsymbol{\varepsilon}_K\end{pmatrix} \quad (5\text{-}8),$$

where K is the number of variates in the model, $\mathbf{g}_k$, $\mathbf{u}_k$, and $\boldsymbol{\varepsilon}_k$ are vectors of genotypic values, random genetic effects, and residuals of variate k, respectively, and m_k is a mean of variate k. Assumptions for the random effects were included in which $\mathbf{g}_{all} = (\mathbf{g}_1{}^T, \ldots, \mathbf{g}_D{}^T)^T$ follows $N(\mathbf{g}_{all} \mid \mathbf{0}, \mathbf{K} \otimes \mathbf{G})$ and $\mathbf{e}_{all} = (\mathbf{e}_1{}^T, \ldots, \mathbf{e}_D{}^T)^T$ follows $N(\mathbf{e}_{all} \mid \mathbf{0}, \mathbf{R} \otimes \mathbf{I})$. **K** is a genomic variance-covariance matrix between variates, and **R** is a residual variance-covariance matrix between variates. The R package MTM (ver. 1.0.0) was used to solve Eq. 5-8 based on the Markov-chain Monte-Carlo (MCMC) method.

Since the canopy area was measured repeatedly in each environment, the number of observations was 152 in total. It was difficult to include such many phenotype data in MGP because the variance-covariance matrix would be too large to be appropriately estimated. Thus, a strategy was applied where MGP was repeated for each environment's observation date, and ten additional variates were selected from the whole data to support the prediction every time. In other words, eleven variates included in the MGP every time consisted of one target variate and ten supporting variates. The criterion to select supporting variates is based on the heritability and the correlation with the target variate. These two factors were essential to improve the prediction accuracy in MGP (Calus & Veerkamp, 2011). Top-10 observations of the following criterion were selected as supporting variates:

$$s(h^2) + s(|r|) \quad (5\text{-}9),$$

where s(.) 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.

As TGP, TMGP consisted of two steps, i.e., MGP of the growth parameters and the calculation of the canopy area using predicted growth parameters. The same criterion of the variate selection (Eq. 5-9) was applied to select ten supporting variates.

The second prediction scheme is the cross-validation for combining a genotype and an environment (CV2), where data of the target genotype in one environment were left out (Fig. 5-5a). As in CV1, four models' prediction accuracies (GP, TGP, MGP, and TMGP) were compared. In this scheme, the combination of a treatment and a year was treated as an environment; there were six environments (two treatments × three years). GP and TGP outputted the same predicted values as CV1 since they did not use data of non-target environments. On the other hand, MGP and TMGP could utilize data in non-target environments by selecting them as supporting traits, which might improve their prediction accuracy.

The third prediction scheme (CV3) is similar to CV2 but focused on the later growth period. As in CV2, data of the target genotypes in one treatment on one year were left out, but data in the former half of the growth period were given. In this scheme, the accuracy of GP, TGP, MGP, and TMGP for the prediction of future growth (TMGPG, Fig. 5-5d) was compared. The model for MGP in CV3 is the same as in CV2, but the selected supporting variates were different. Seven out of ten supporting variates were selected using the selection criterion (Eq. 5-9). Additionally, the canopy area of the latest three days in the former half of the target environment's growth period was chosen as the supporting variates. The canopy area in the former half of the growth period in the target environment was expected to improve the accuracy due to high correlations with the canopy area, which was the target of the prediction, even when they had low heritability.

TMGPG consisted of three steps. The first two steps are the same as TMGP. In these steps, the growth parameters were predicted without using the data from the former half of the growth period. One different point was that the MCMC samples of the growth parameters were saved, whereas the samples' average was used as the predicted values in TMGP. As a result, 60,000 samples of the predicted growth curves could be obtained for each genotype. Those samples can be understood as samples from a probability distribution before the growth data acquisition. Then, the former half of the growth period's data were reflected using the approximate Bayesian computation (ABC) method. Sixty samples of growth curves were selected for each genotype by evaluating each sample by the sum of squares of the differences between the predicted growth curves and the given growth data. Finally, the mean values of the 60 samples of growth curves were used as the predicted values.

Cross-validation was repeated three times for the combination of a cross-validation scheme and a model. The correlation coefficient between genotypic values ($\mathbf{g}$) and their predicted values ($\mathbf{u}$) of the canopy area was used to evaluate the prediction accuracy.

5.3 Results

5.3.1 Time-series canopy area data measured with UAV-RS

The canopy area's longitudinal growth processes could be obtained using UAV-RS (Fig. 5-6). Large variations in growth patterns of each plot were observed. Because of severe heat stress in the experimental field of 2019, the canopy area of 2019-D was smaller than the others.

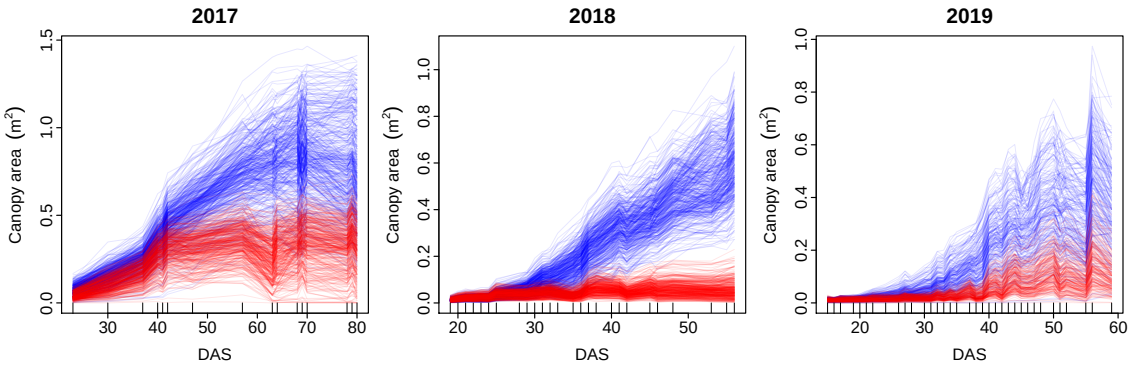

Figure 5-6. Longitudinal process of the canopy area measured with UAV-RS. The canopy area in treatment C and D are shown in blue and red lines, respectively. The number of days after sowing (DAS) was used as the x-axis.

The time-series heritability of the canopy area showed a U-shape pattern in all the environments (Fig. 5-7) where the value decreased until 40–50 days after sowing and increased from 50 days after sowing.

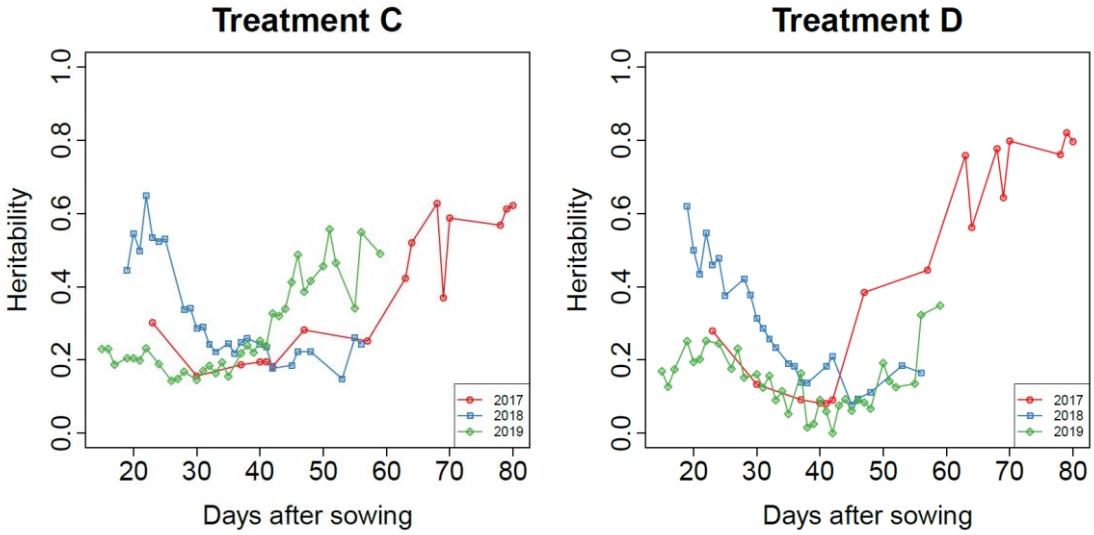

Figure 5-7. Heritability of the time-series canopy area of treatment C and D. Red, blue, and green lines indicate the values in 2017, 2018, and 2019, respectively.

5.3.2 Growth parameter estimation

The growth model fitted a wide range of canopy area growth patterns (Fig. 5-8). However, the goodness of fit was not sufficient in a few cases.

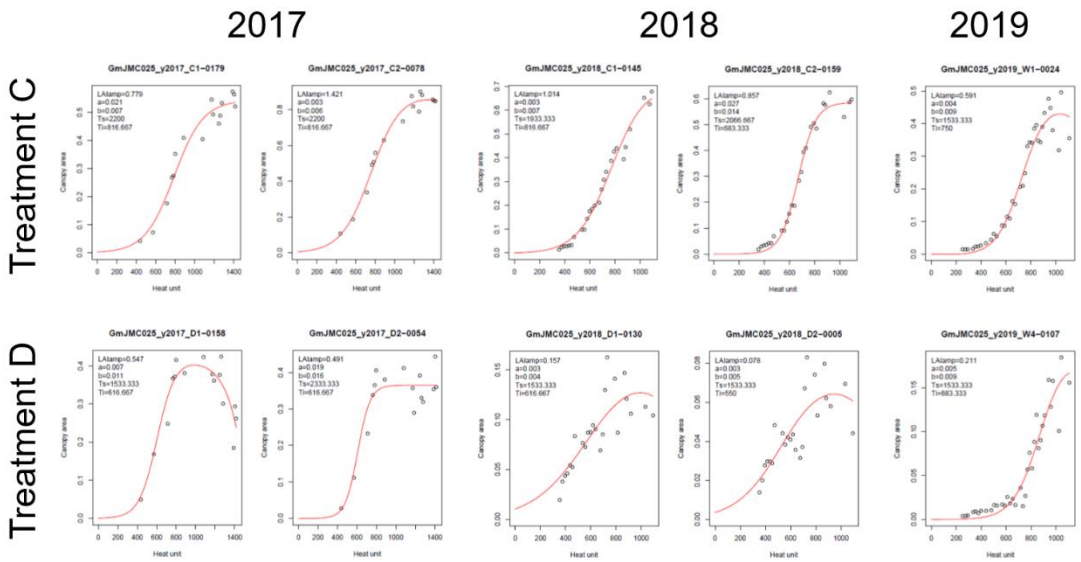

Figure 5-8. Longitudinal growth data of the canopy area of the genotype "Enrei," a famous Japanese cultivar, and growth curves fitted to the data. The results of each plot are displayed separately.

The heritability of the growth parameters varied among the treatments and the years (Fig. 5-9). The heritabilities of LAI_{amp} and T_g were relatively higher than the others. The heritability was the highest in 2017 when that of LAI_{amp} was around 0.6 and the lowest in 2018 when that of LAI_{amp} was around 0.2. The heritabilities of r_g and r_s were low, except for those of 2017-D.

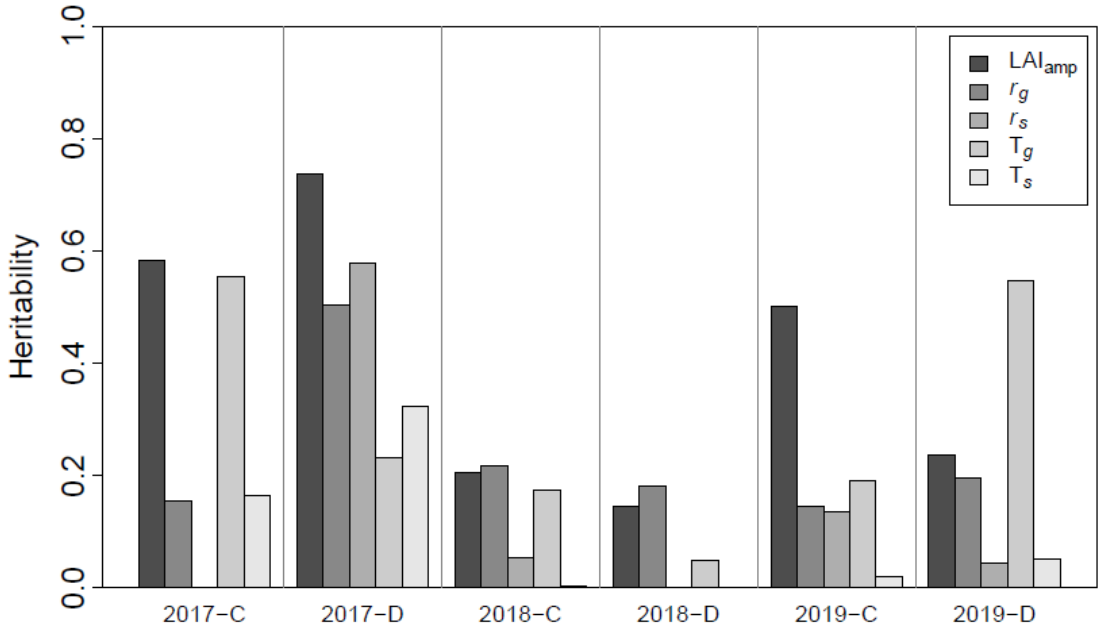

Figure 5-9. Heritability of the growth parameters in each environment (a combination of treatment and year).

5.3.3 Prediction of growth patterns

In CV1, the prediction accuracy of TGP and MGP was higher than GP (Fig. 5-10 and 5-11). Significant improvement of prediction accuracy was observed in 2019-D, where the accuracy of GP was close to zero in the latter half of the growth period. Comparing MGP and TGP, the accuracy of MGP tended to be slightly higher than that of TGP. The accuracy of TMGP differed among environments; it was lower than the accuracy of GP when predicting the canopy area in 2018, while it was higher than the accuracy of MGP when predicting the canopy area in the latter half of the growth period in 2019.

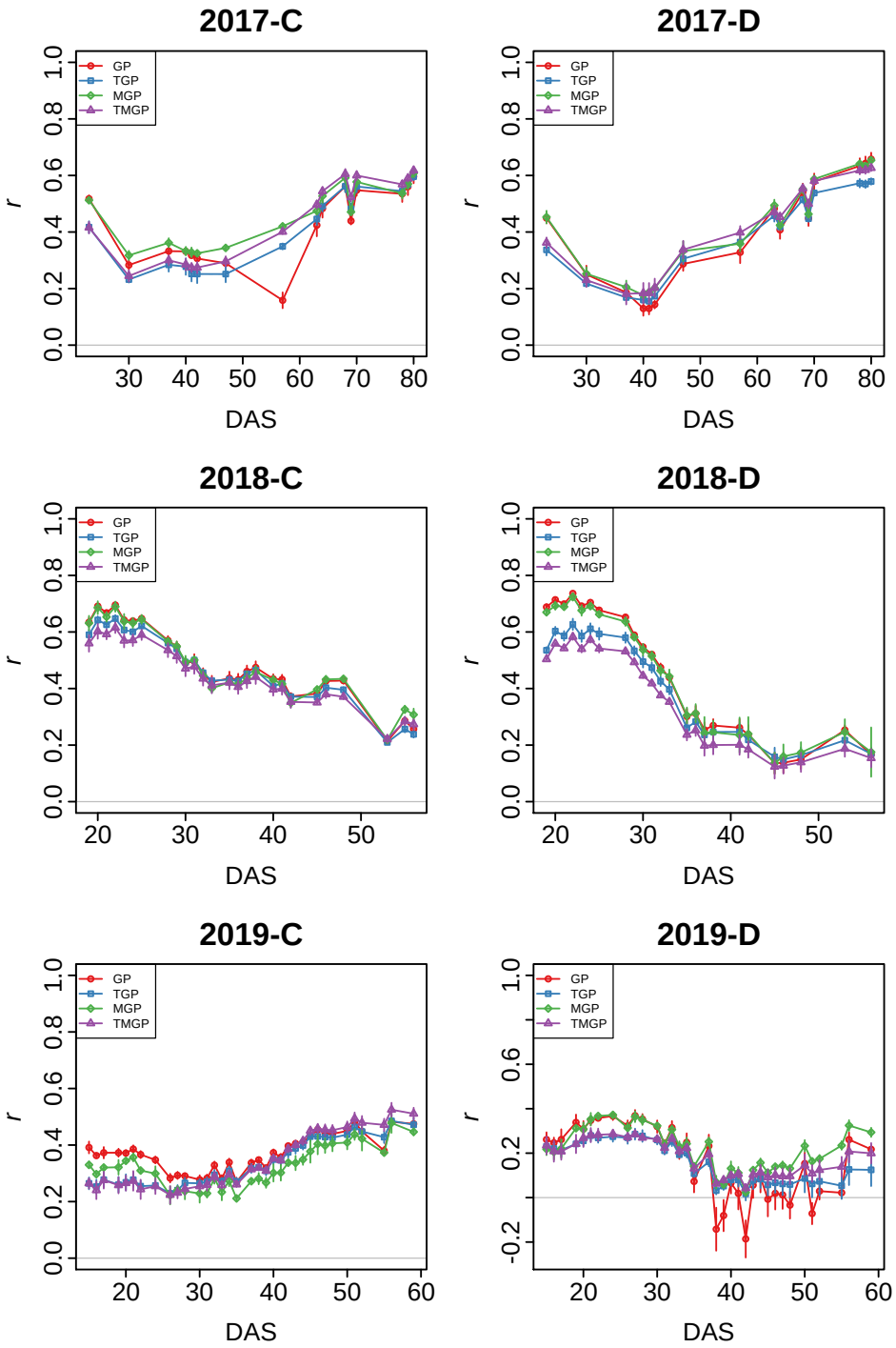

Figure 5-10. The prediction accuracy of the canopy area using four models in CV1. The number of days after sowing (DAS) was used as the x-axis, whereas the correlation coefficients (*r*) between observed and predicted values yielded were plotted as the y-axis. The mean values of correlations yielded from repeated cross-validation were plotted as dots, and the range of ±1 standard deviations was expressed as vertical bars added to each dot. Red, blue, green, and purple lines correspond to the accuracy of GP, MGP, TGP, and TMGP, respectively.

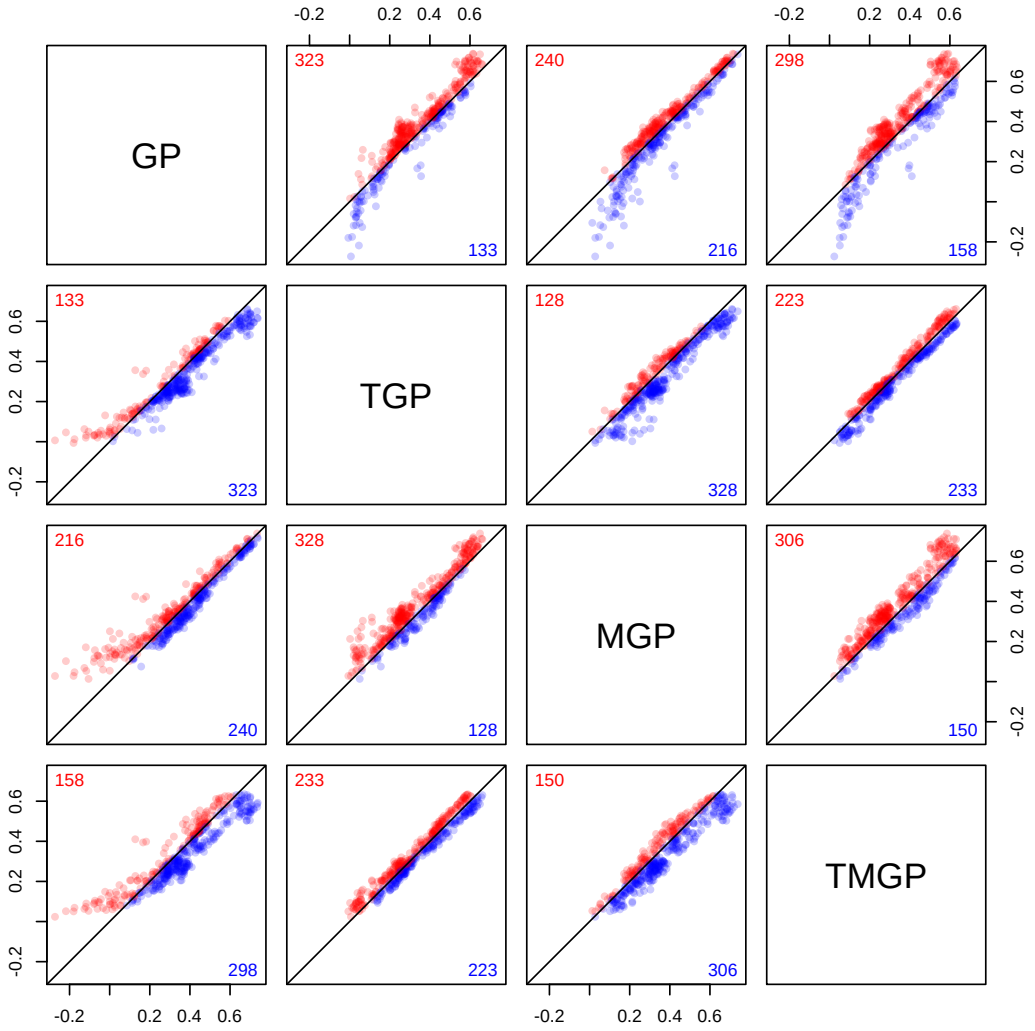

Figure 5-11. Comparison of the prediction accuracy in CV1. Two out of four models (GP, MGP, TGP, and TMGP) were selected in each panel. The correlation coefficients between genotypic values and their predicted values were compared between the two models. Solid lines indicate where the accuracy was equal between the two models. The colors of the points (red/blue) indicate whether each point is upper or lower than the solid lines. The red and blue numbers are the number of red and blue points, respectively.

Predicted values with GP and TGP in CV2 are equal to those in CV1 because they did not utilize data in environments other than their targets. Thus, for CV2, focus will be on the accuracy of MGP and TMGP. The accuracy of MGP was higher in CV2 than CV1 and was significantly higher than that of TGP (Fig. 5-12 and 5-13). The accuracy of TMGP was lower in CV2 than CV1 in 2018, while it was higher in CV2 than 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.

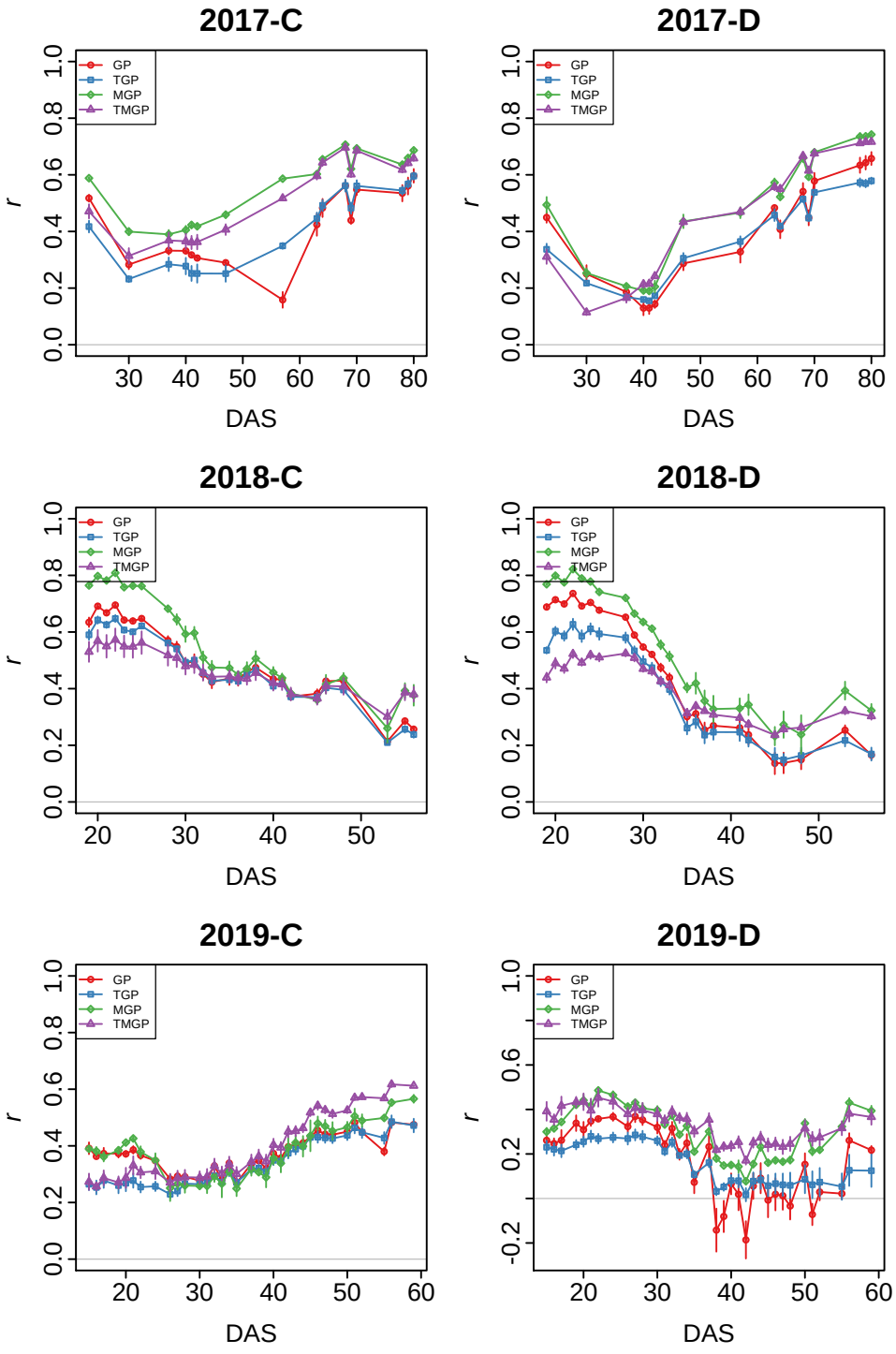

Figure 5-12. The prediction accuracy of the canopy area using four models in CV2. The number of days after sowing (DAS) was used as the x-axis, whereas the correlation coefficients (*r*) between observed and predicted values yielded were plotted as the y-axis. The mean values of correlations yielded from repeated cross-validation were plotted as dots, and the range of ±1 standard deviations was expressed as vertical bars added to each dot. Red, blue, green, and purple lines correspond to the accuracy of GP, MGP, TGP, and TMGP, respectively.

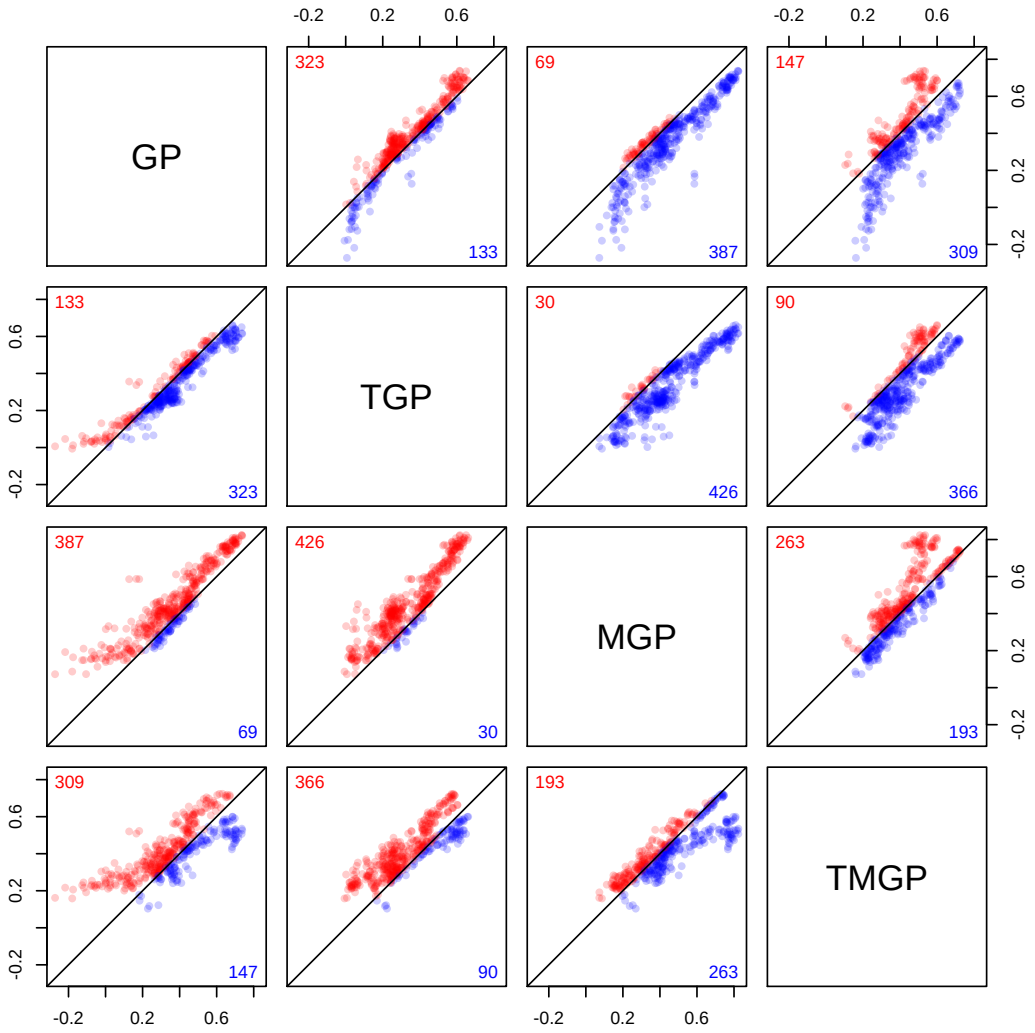

Figure 5-13. Comparison of the prediction accuracy in CV2. Two out of four models (GP, MGP, TGP, and TMGP) were selected in each panel. The correlation coefficients between genotypic values and their predicted values were compared between the two models. Solid lines indicate where the accuracy was equal between the two models. The colors of the points (red/blue) indicate whether each point is upper or lower than the solid lines. The red and blue numbers are the number of red and blue points, respectively.

In CV3, the prediction accuracy of TMGPG was the best in all environments and the whole growth period (Fig. 5-14 and 5-15). 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 and TGP but lower than that of TMGPG. As in CV2, predicted values with GP and TGP in CV3 are the same as CV1.

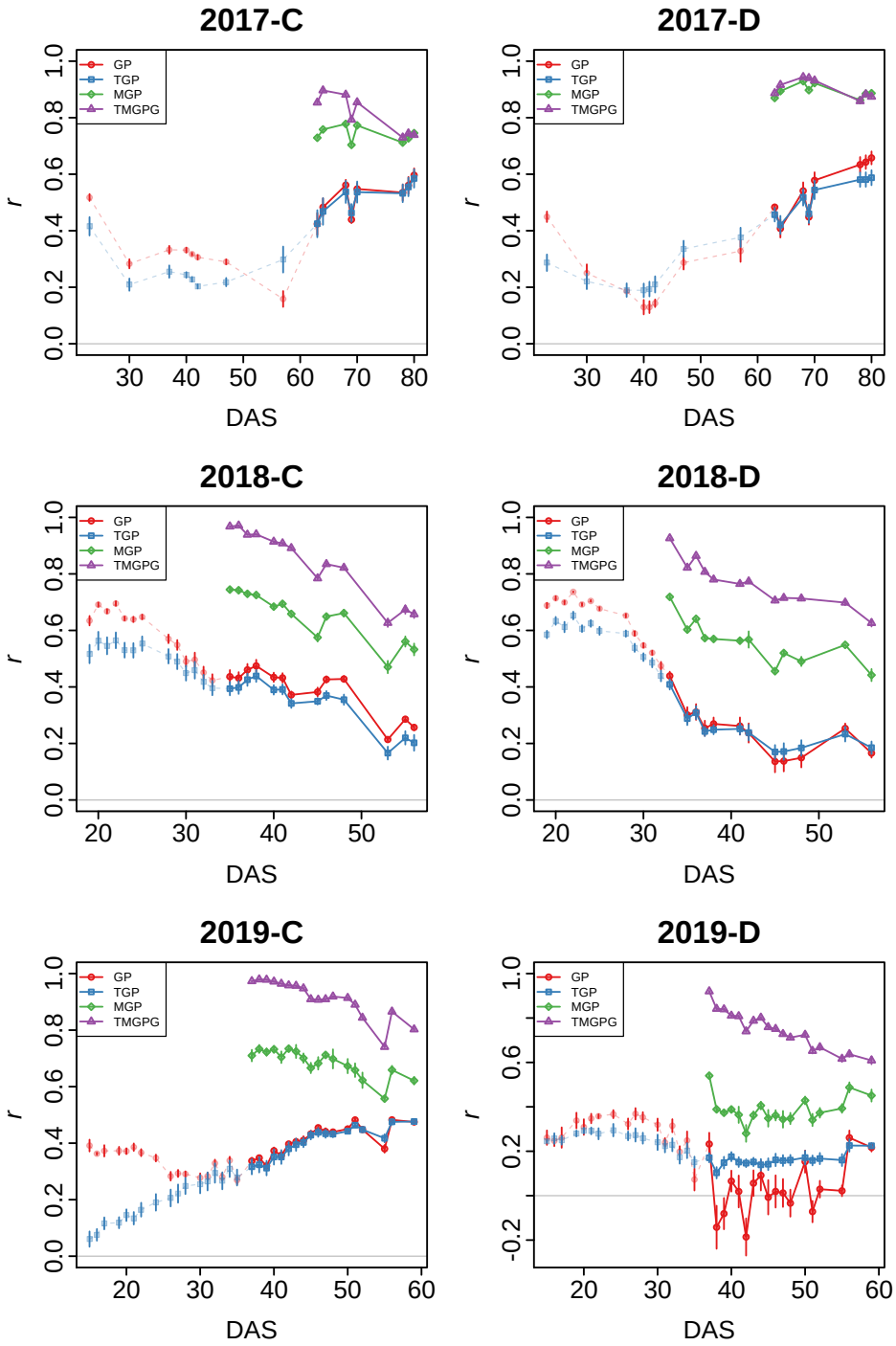

Figure 5-14. The prediction accuracy of the canopy area using four models in CV3. The number of days after sowing (DAS) was used as the x-axis, whereas the correlation coefficients (*r*) between observed and predicted values yielded were plotted as the y-axis. The mean values of correlations yielded from repeated cross-validation were plotted as dots, and the range of ±1 standard deviations was expressed as vertical bars added to each dot. Red, blue, green, and purple lines correspond to the accuracy of GP, MGP, TGP, and TMGPG, respectively. The accuracy in the former half of the growth period of GP and MGP is shown in fine points.

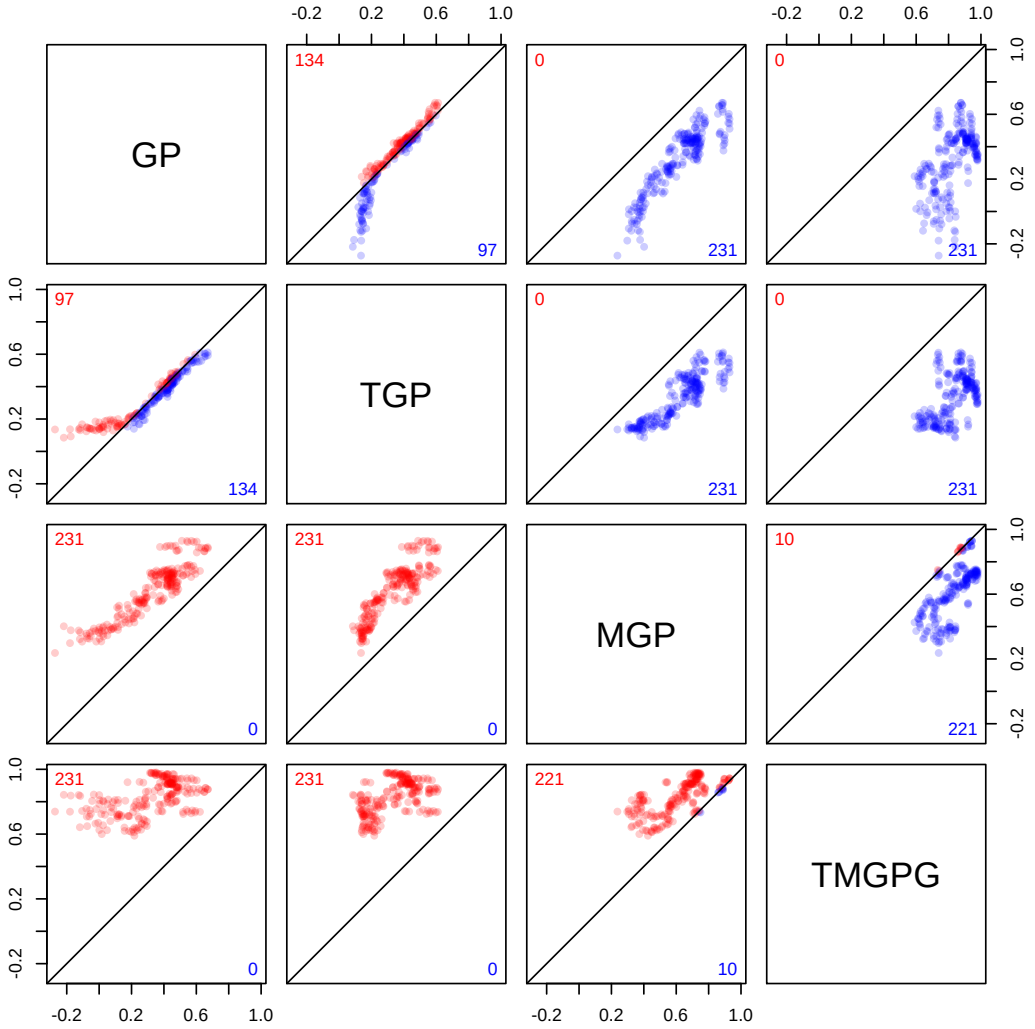

Figure 5-15. Comparison of the prediction accuracy in CV3. Two out of four models (GP, MGP, TGP, and TMGPG) were selected in each panel. The correlation coefficients between genotypic values and their predicted values were compared between the two models. Solid lines indicate where the accuracy was equal between the two models. The colors of the points (red/blue) indicate whether each point is upper or lower than the solid lines. The red and blue numbers are the number of red and blue points, respectively.

5.4 Discussion

5.4.1 UAV-RS as a tool to evaluate growth patterns

This study showed that the canopy area's UAV-RS measurement could be used to assess genetic diversity in soybean growth patterns. A U-shape longitudinal pattern was observed in the heritability of the canopy area in all the environments (Fig. 5-7). The reason for the U-shape heritability patterns can be explained in three steps. In the early stage of growth, the canopy area seemed to be determined by a few factors regarding initial growth speed, such as radiation use efficiency, which results in high heritability of the canopy area. At about 25 days after sowing, several factors such as growth phenology and plant structure related to the determination of growth phenology started to affect the canopy area, which decreased the heritability. Then, a saturation of the canopy area occurred around 45–60 days after sowing. In this period, the confounding factors related to growth phenology became weak, leading to increased heritability.

5.4.2 Fitting of growth models

The growth model was flexible enough to represent various growth patterns in different environments (Fig. 5-8). By fitting the model, the time-series canopy area was decomposed into five parameters. The genetic variation in the growth process was examined in detail, focusing on each parameter.

Genetic analysis of the parameters showed that the heritability of LAI_{amp} was high (Fig. 5-9). This result explains why the heritability of the canopy area increased in the later stages of growth. T_g, which describes the stage-shift timing of the canopy area, also showed high heritability in 2017-C and 2019-D.

For the senescence stage, the heritability was low in 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 succeeded in that environment. Heritability of r_s was higher than T_s, which means that the change of canopy area in the senescence stage was mainly determined by its speed, r_s, rather than the timing of senescence, T_s. It is expected that the senescence pattern will be evaluated more precisely in 2017 and 2018 by extending the cultivation periods to observe the whole process of senescence.

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

Another possible improvement of the model is the inclusion of other environmental effects, such as soil moisture and drought stress. In the growth model, the effect of temperature on growth stages was considered. However, the inclusion of other factors may allow for improved fitting and the simultaneous parameter estimation of multiple environments. For example, the low heritability in 2018 of the growth parameters was because severe heat stress in the summer of 2018 made the growth slower than usual years, and thus the sigmoid pattern in growth was truncated at the end of the cultivation (Fig. 5-6). Considering other environmental factors will allow simultaneous parameter estimation with other environments, leading to stability in the estimated parameters.

5.4.3 Prediction of growth curves

In CV1, the accuracy was close between GP and TGP (Fig. 5-11). This result suggests that the growth model used in TGP could extract sufficient genetic variations from phenotypic variations in the longitudinal growth pattern to achieve the same predictive accuracy as the GP.

Models with multivariate GP yielded better accuracy than those with univariate GP; the accuracy of MGP and TMGP were better than those of GP and TGP, respectively (Fig. 5-11). High correlations among variates, a typical property of longitudinal data, suggest that multivariate GP improves the prediction accuracy because MTG and TMGP can leverage the among-characteristics correlation. In the following paragraphs, the focus is on the comparison of MGP and TMGP.

In 2018, the accuracy of TMGP was lower than that of MGP in CV1 and CV2 (Fig. 5-10 and 5-12) due to the low heritability of the growth parameters, as described in Section 5.4.2. On the other hand, the accuracy of TMGP was higher than that of MGP in 2019 in CV2. TMGP was better than MGP due to the higher heritability of growth parameters than the canopy area in 2018. Extraction of genetic variance in growth patterns in 2019 was successful as $\mathrm{LAI}_{\mathrm{amp}}$ in 2019-C and T_g in 2019-D, leading to improved prediction accuracy.

In CV3, the prediction accuracy of TMGPG outperformed the other models (Fig. 5-15). The higher prediction accuracy than 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 growth model. Correlation coefficients between the predicted values of TMGPG and the genotypic values exceeded 0.6 in most cases in all environments, indicating that TMGPG is robust to changes in the environment. Similar prediction accuracy of MGP and TMGPG in 2017-D may be due to the lack of change in the canopy area in the second half of this environment's growth period. This approach to future prediction through growth models has potential application for selection in early growth stages in crop breeding.

In this study, the growth model and GP/MGP were used separately in TGP/TMGP, but they can be integrated as one hierarchical model. Several reports have shown the effectiveness of hierarchical models in the analysis of longitudinal traits (Onogi et al., 2019), QTL analysis (Ma et al., 2002), and GWAS (Das, Li, Wang, et al., 2011; Crispim et al., 2015). It is expected that the

joint analysis will make the parameter estimation robust. In this study, two steps are required to estimate the growth parameters, but joint estimation may simplify the estimation process.

A random regression model is also known as a regression method of longitudinal data with a mixed model structure, and it was used in GP of longitudinal growth data (Sun et al., 2017; Campbell et al., 2018). Although random regression cannot incorporate the growth curve structure like the growth model in this study, its formulation is simple (see Section 2.2.3). Random regression is assumed to perform better than MGP in the prediction scheme of CV3, but it will not be as accurate as TMGPG.

5.4.4 Growth analysis on remote sensing data for plant breeding

Applying the growth model to longitudinal growth data measured using UAV-RS data allows us to capture genetic variation in growth patterns. The integration of growth model functions with genetic analysis was shown to be a practical approach for analyzing field experiments growth processes. The integrated model of GP and the growth model was able to predict the future growth curve in the early growth stage. The future growth prediction can be used for selection in the early growth stages, which will reduce the cost for the field trials. This study suggested that data collection using UAV-RS and its analysis using growth models and mixed models will benefit crop breeding.

In the near future, it is expected that UAV-RS will play an active role in the plant breeding field 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 growth 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.

6 Prediction of soybean growth curves by modeling genetic and environmental effects on daily growth

6.1 Introduction

It is well known that many plant traits are affected by both genetic and environmental factors. Genotypic, environmental, and genotype-by-environment interaction (G × E) effects have been considered and evaluated simultaneously in plant breeding and genetics. For example, several studies have proposed models that include these three effects in genomic prediction (GP) (Schulz-Streeck et al., 2013; Technow et al., 2015; Onogi, Watanabe, et al., 2016; Jarquin et al., 2018). Previous studies on G × E analysis have mainly focused on traits at harvest because they directly determine the agricultural value of crops and genotypes (Kang, 2001). However, considering the actual mechanism, the G × E of traits at harvest is determined by the accumulation of G × E during the growth process. If we can evaluate G × E during the growth process, we can gain useful knowledge in plant breeding, such as environmental factors affecting a trait at any particular time during the growth process.

However, In crop breeding, it has been difficult to observe the growth process since many genotypes are usually tested simultaneously in a field trial. Thanks to the rapid development of sensing technologies in recent years, high-throughput phenotyping is becoming available in plant genetics and breeding, and the measurement of longitudinal growth data is becoming more practical. Accurate and detailed acquisition of the growth process by high-throughput measurements is expected to improve genetic gain in plant breeding (Araus & Cairns, 2014; Cabrera-Bosquet et al., 2012; Furbank & Tester, 2011). For field phenotyping, the application of remote sensing with unmanned aerial vehicles (UAV-RS) is expected because of its low cost of implementation and management. One of the problems is that the phenotypic data obtained from UAV-RS is noisy because UAVs usually obtain plant images from a distance of 10 to 100 meters. To properly trace the plant growth process, it is necessary to develop a noise filtering method suitable for UAV-RS data.

In this section, two topics were considered to develop a G × E analysis method of the growth process. The first is the development of a model to estimate the growth curves using UAV-RS data accurately. The proposed model assumed that the UAV-RS data is affected by two factors, noise and bias, and these were included in a hierarchical Bayesian model. The model was applied to the soybean canopy area and height, and the results were validated with ground-truth data of the canopy height.

The second topic is the modeling of the daily response of growth traits to environmental factors. Two approaches were proposed: an additive spline (AS) model and a machine learning (ML) model. In the AS model, the daily response curve to environmental effects was represented by additive splines whose coefficients were determined by considering the genotypic relationship matrix. On the other hand, in the ML model, environmental factors and marker genotypic data

were used equally as dependent variables in the Random forest. Although the ML model cannot explicitly model the plant traits' environmental response, it can incorporate complex interaction among environmental factors and may have higher prediction accuracy than the AS model.

6.2 Material and methods

6.2.1 Data acquisition

The data was obtained from the same field experiments as Chapter 4 (Fig. 6-1). In 2017 and 2018, no-watered (W0) and well-watered (WW) treatments were set to evaluate the influence of drought and control conditions on phenotypic variations, as explained in Chapter 4. In 2019, four watering levels were investigated. Two of them were equivalent to those in 2017 and 2018, i.e., W0 and WW, but different watering patterns were set for the rest two; five days watering followed by five days no-watering (W5), and ten days watering, and ten days no-watering (W10). Watering treatment started after thinning every year (Fig. 5-2). I use an abbreviation for the combination of the treatment and year; treatment WW in 2017 is called “2017-WW”.

Plant height of a subset of plots was measured manually as the ground truth data and was used to correct the UAV-RS-based canopy height measurement bias. Plant height was defined as the height of the top of a plant from the ground. In 2017, all plots were separated into three groups. The plant height of each group was measured in turn almost once a week. In 2018 and 2019, nine and eight plots were chosen from each block, respectively. The plant height of the selected plots was measured every day.

Minimum, mean, and maximum temperature, solar radiation, and transpiration were measured as environmental factors (Fig. 6-2). Also, soil moisture was measured using (TDR-341F, Fujiwara Seisakusho, Japan) every day at several points in the field.

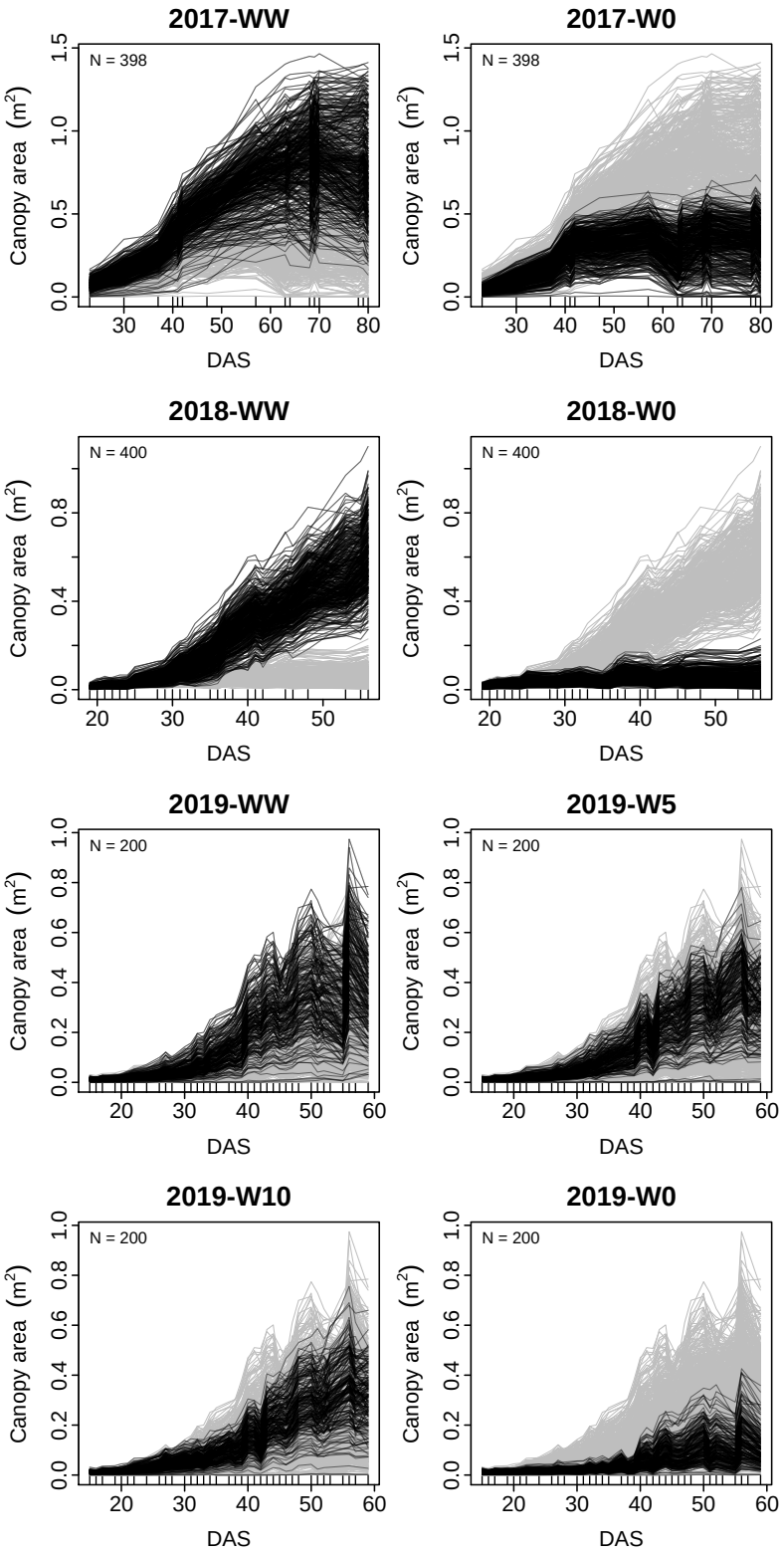

Figure 6-1-1. The canopy area measured by UAV-RS. Observed values of each treatment were plotted. Phenotypic data of the same year was plotted with gray lines.

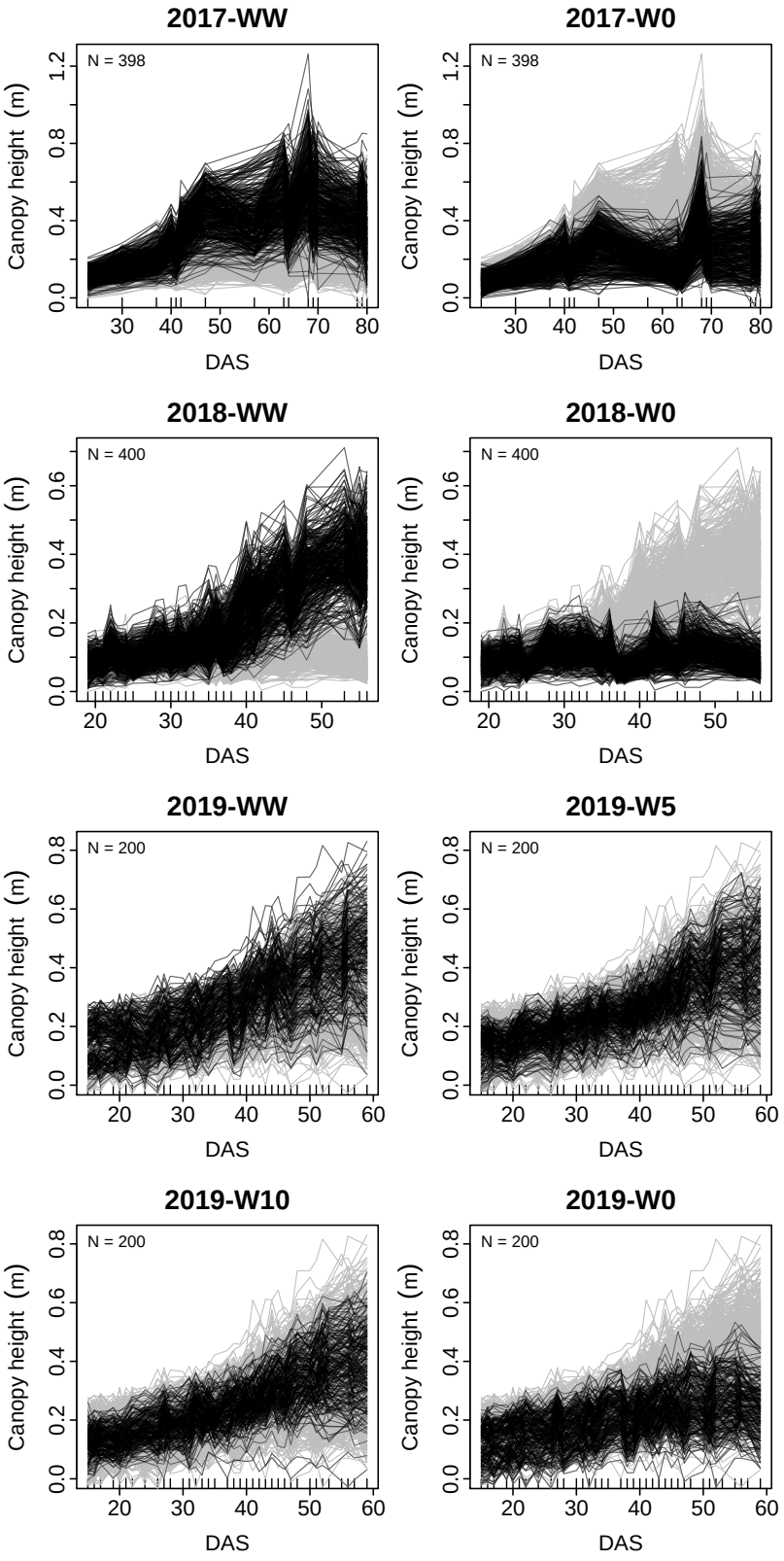

Figure 6-1-2. The canopy height measured by UAV-RS. Observed values of each treatment were plotted. Phenotypic data of the same year was plotted with gray lines.

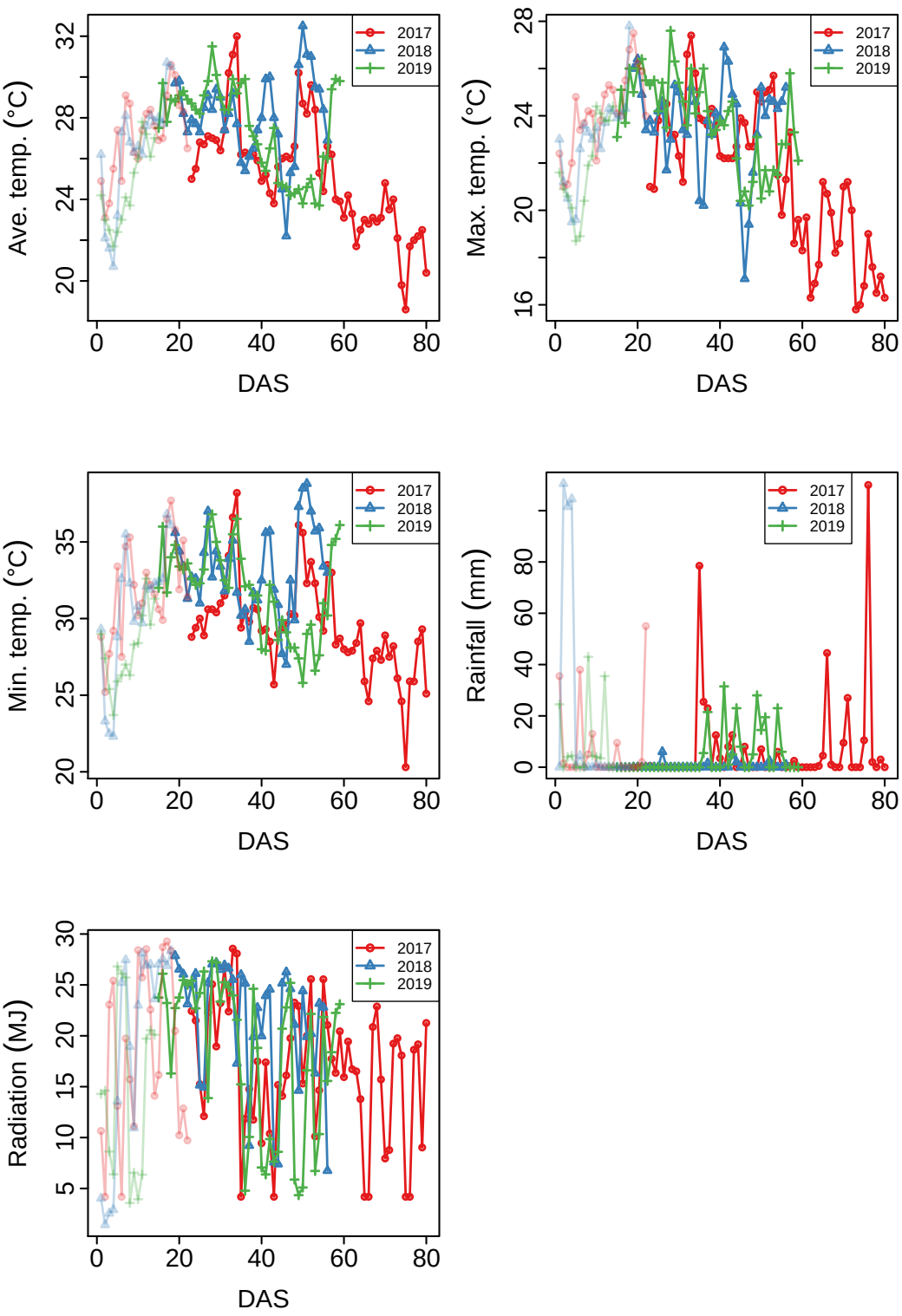

Figure 6-2. The longitudinal curves of the environmental factors. Average, maximum, and minimum daily temperature, rainfall, and radiation are shown. Data, before the observation started, are plotted in fine lines and points.

6.2.1 Filtering of noise and bias

The canopy area and height data were found to have large measurement noise and bias. However, standard noise filtering methods such as a spline fitting were not appropriate because the canopy area and height sometimes showed a sudden increase and decrease (Fig. 6-1). Therefore, a statistical method to remove the noise and bias was developed (Fig. 6-3). R (http://www.r-project.org) was used for the following analyses.

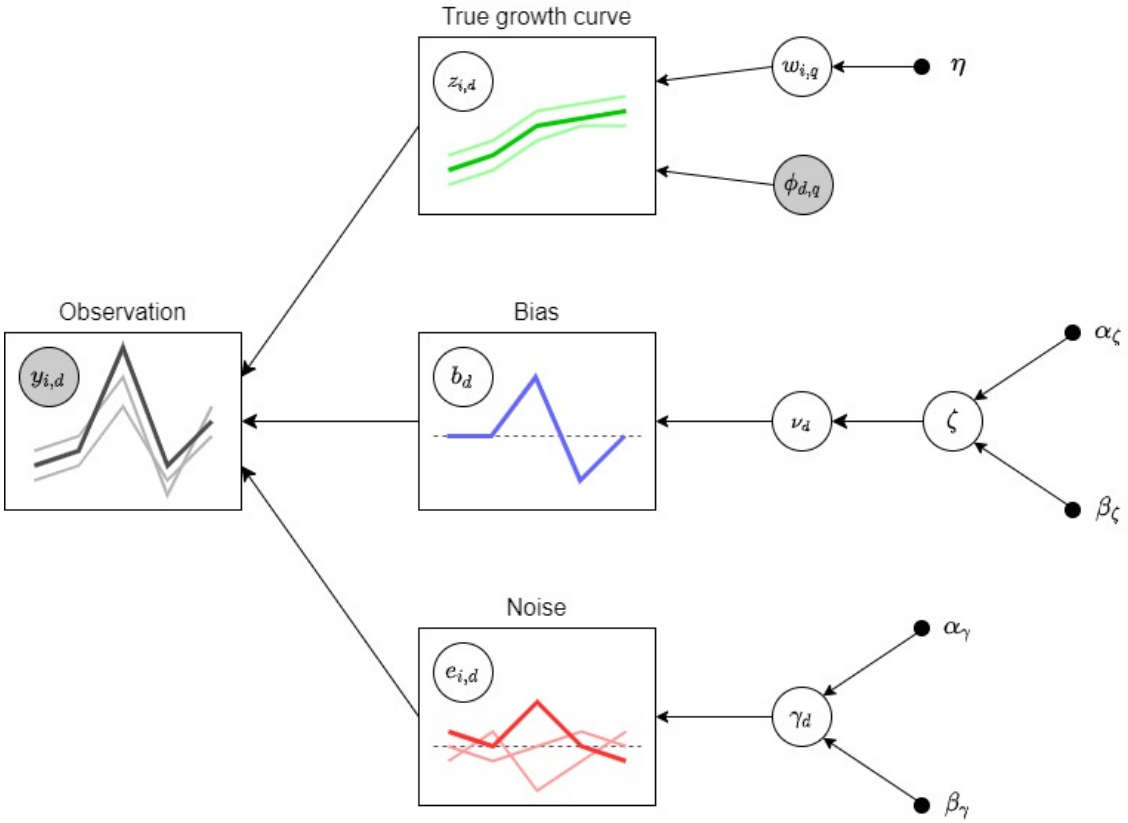

Figure 6-3. The model structure assumed to filter the measurement bias and noise. White, gray, and black nodes indicate latent variables, given variables, and hyperparameters, respectively.

First, the measured values were decomposed into several factors as

$$y_{i,d} = z_{i,d} b_d + e_{i,d} \tag{6-1},$$

where $y_{i,d}$ is the measured values of the canopy area or height of plot i on day d, $z_{i,d}$ is the true value of the canopy area or height, b_d is the measurement bias, and $e_{i,d}$ is the measurement noise. The measurement bias was assumed to be shared by all plots on each day. The measurement noise $e_{i,d}$ was assumed to be independent:

$$p(e_{i,d}) = N(e_{i,d}|0, \gamma_d^{-1}) \tag{6-2},$$

$$p(\gamma_d) = \mathrm{Gamma}(\gamma_d|\alpha_\gamma, \beta_\gamma) \tag{6-3},$$

where $N(. \mid 0, \gamma_d^{-1})$ is the Gaussian distribution with mean 0 and variance γ_d^{-1}, $\mathrm{Gamma}(. \mid \alpha, \beta)$ is the Gamma distribution with parameter α and β,

$$\mathrm{Gamma}(x|\alpha, \beta) = \frac{1}{\Gamma(\alpha)} \beta^\alpha x^{\alpha-1} \exp(-\beta x) \tag{6-4}.$$

The hyperparameters were set as $\alpha_\gamma = 0.1$ and $\beta_\gamma = 0.1$ so that the prior would be close to noninformative distribution. The bias, b_d, seems to be one on most days (when plant height was estimated without bias) but may not be one on the others (when plant high was estimated with bias). Such characteristics can be expressed with a hierarchical model as

$$p(b_d|\nu_d) = \begin{cases} \delta(b_d - 1) & \text{if } \nu_d = 0 \\ \mathrm{Unif}(b_d|0, 2) & \text{if } \nu_d = 1 \end{cases} \tag{6-5},$$

$$p(\nu_d|\zeta) = \mathrm{Bern}(\nu_d|\zeta) \tag{6-6},$$

$$p(\zeta) = \mathrm{Beta}(\zeta|\alpha_\zeta, \beta_\zeta) \tag{6-7},$$

where $\delta(.)$ is the Dirac's delta function value of which is zero when the input is not zero, $\mathrm{Unif}(. \mid \alpha, \beta)$ is the uniform distribution value of which is constant when the input is larger than α and smaller than β, $\mathrm{Bern}(. \mid \zeta)$ is the Bernoulli distribution with success rate ζ, and $\mathrm{Beta}(. \mid \alpha, \beta)$ is the Beta distribution with parameter α and β,

$$\mathrm{Beta}(x|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1 - x)^{\beta-1} \tag{6-8}.$$

Thus, ν_d is a latent variable determining whether there was a bias on day d, and ζ indicates the probability of the occurrence of the bias. The hyperparameters were set as $\alpha_\zeta = 0.01$ and $\beta_\zeta = 0.01$ so that the prior would be close to noninformative distribution.

The true values of the canopy area and height, $z_{i,d}$, of a plot were assumed to be on a spline curve. Since the spline function can be described with a linear combination of basis functions, its structure can be written as

$$z_{i,d} = \sum_{q=1}^{Q} w_{i,q}\phi_{d,q} \quad (6\text{-}9),$$

$$p(w_{i,q}) = N(w_{i,q}|0, \eta^{-1}) \quad (6\text{-}10),$$

where $\phi_{d,q}$ is the q^{th} basis of the B-spline on day d, Q is the number of bases (Q = 6), $w_{i,q}$ is the linear combination coefficient, and η is a hyperparameter determining the variance of $w_{i,q}$. The hyperparameter was set as $\eta = 10^{-6}$ so that the prior would be flat.

The variational Bayes method was used to estimate the parameters. The joint distribution of all the parameters is

$$\begin{aligned} &p(\mathbf{Y}, \mathbf{Y}_0, \mathbf{Z}, \mathbf{b}, \boldsymbol{\gamma}, \mathbf{w}, \boldsymbol{\Phi}, \boldsymbol{\nu}, \zeta) \\ &= p(\mathbf{Y}|\mathbf{Z}, \mathbf{b}, \boldsymbol{\gamma})p(\mathbf{Y}_0|\mathbf{Z}, \mathbf{b}, \boldsymbol{\gamma})p(\boldsymbol{\gamma})p(\mathbf{Z}|\mathbf{w}, \boldsymbol{\Phi})p(\mathbf{w})p(\mathbf{b}|\boldsymbol{\nu})p(\boldsymbol{\nu}|\zeta)p(\zeta) \end{aligned} \quad (6\text{-}11),$$

where $\mathbf{Y}$, $\mathbf{Z}$, $\mathbf{b}$, $\boldsymbol{\gamma}$, $\mathbf{w}$, $\boldsymbol{\Phi}$, and $\boldsymbol{\nu}$ are matrices or vectors $y_{i,d}$, $z_{i,d}$, b_d, γ_d, $w_{i,q}$, $\phi_{d,q}$, and v_d, respectively. In those variables, $\mathbf{Y}$ and $\boldsymbol{\Phi}$ are the given values as data, and the rest are the parameters to be estimated. Here, the growth process data was separated in $\mathbf{Y}$ and $\mathbf{Y}_0$, the observed and missing values, to distinguish given values and parameters to be estimated. The missing observation data appeared when the image processing of UAV-RS data with the software Pix4Dmapper failed. It was assumed that the joint distribution Eq. 6-11 could be approximated with a probability density

$$q(\mathbf{Y}_0, \mathbf{Z}, \mathbf{b}, \boldsymbol{\gamma}, \mathbf{w}, \boldsymbol{\nu}, \zeta) = q_1(\mathbf{Z}, \mathbf{w})q_2(\mathbf{b}, \boldsymbol{\nu})q_3(\mathbf{Y}_0, \boldsymbol{\gamma}, \zeta) \quad (6\text{-}12),$$

under the assumption that the joint distribution can be decomposed into three distributions. Then, it is known that the $q_k^\star$ (k = 1, 2, and 3), which is the nearest approximation of joint distribution (Eq. 6-12), can be estimated by (Bishop, 2006)

$$\ln q_k^\star(\boldsymbol{\theta}_k) = \mathbb{E}_{k' \neq k}[\ln p(\mathbf{X}, \boldsymbol{\theta})] + \text{const.} \quad (6\text{-}13),$$

where $\boldsymbol{\theta}_k$ is the parameter vector of q_k, $\boldsymbol{\theta}$ is the parameter vector of q, $\mathbb{E}_{k' \neq k}[.]$ is the expectation with the parameters of $q_{k'}$ where $k' \neq k$, and $\mathbf{X}$ is the given data. In this study, $q_1^\star$ and $q_2^\star$ are

$$\begin{aligned} \ln q_1^\star(\mathbf{Z}, \mathbf{w}) &= \mathbb{E}_{\mathbf{Y}_0, \mathbf{b}, \boldsymbol{\gamma}, \boldsymbol{\nu}, \zeta}[\ln p(\mathbf{Y}, \mathbf{Y}_0, \mathbf{Z}, \mathbf{b}, \boldsymbol{\gamma}, \mathbf{w}, \boldsymbol{\Phi}, \boldsymbol{\nu}, \zeta)] + \text{const.} \\ &= \mathbb{E}_{\mathbf{Y}_0, \mathbf{b}, \boldsymbol{\gamma}}[\ln p(\mathbf{Y}, \mathbf{Y}_0|\mathbf{Z}, \mathbf{w}, \mathbf{b}, \boldsymbol{\gamma})] + \mathbb{E}[\ln p(\mathbf{Z}, \mathbf{w})] + \text{const.} \end{aligned} \quad (6\text{-}14),$$

$$\begin{aligned} \ln q_2^\star(\mathbf{b}, \boldsymbol{\nu}) &= \mathbb{E}_{\mathbf{Y}_0, \mathbf{Z}, \boldsymbol{\gamma}, \mathbf{w}, \zeta}[\ln p(\mathbf{Y}, \mathbf{Y}_0, \mathbf{Z}, \mathbf{b}, \boldsymbol{\gamma}, \mathbf{w}, \boldsymbol{\Phi}, \boldsymbol{\nu}, \zeta)] + \text{const.} \\ &= \mathbb{E}_{\mathbf{Z}, \boldsymbol{\gamma}}[\ln p(\mathbf{Y}|\mathbf{Z}, \mathbf{b}, \boldsymbol{\gamma})] + \ln p(\mathbf{b}|\boldsymbol{\nu}) + \mathbb{E}_\zeta[\ln p(\boldsymbol{\nu}|\zeta)] + \text{const.} \end{aligned} \quad (6\text{-}15).$$

Since $\mathbf{Y}_0$, $\boldsymbol{\gamma}$, and ζ are independent when other parameters are given, $q_3^\star$ can be decomposed into three terms:

$$\ln q_3^\star(\mathbf{Y}_0, \boldsymbol{\gamma}, \zeta) = \ln q_{3Y_0}^\star(\mathbf{Y}_0) + \ln q_{3\gamma}^\star(\boldsymbol{\gamma}) + \ln q_{3\zeta}^\star(\zeta) \quad (6\text{-}16),$$

$$\begin{aligned}\ln q^{\star}_{3Y_0}(\mathbf{Y}_0) &= \mathbb{E}_{\mathbf{Z},\mathbf{b},\boldsymbol{\gamma},\mathbf{w},\boldsymbol{\nu},\zeta}[\ln p(\mathbf{Y},\mathbf{Y}_0,\mathbf{Z},\mathbf{b},\boldsymbol{\gamma},\mathbf{w},\boldsymbol{\Phi},\boldsymbol{\nu},\zeta)] + \text{const.} \\ &= \mathbb{E}_{\mathbf{Z},\mathbf{b},\boldsymbol{\gamma}}[\ln p(\mathbf{Y}_0|\mathbf{Z},\mathbf{b},\boldsymbol{\gamma})] + \text{const.}\end{aligned} \quad (6\text{-}17),$$

$$\begin{aligned}\ln q^{\star}_{3\gamma}(\boldsymbol{\gamma}) &= \mathbb{E}_{\mathbf{Y}_0,\mathbf{Z},\mathbf{b},\mathbf{w},\boldsymbol{\nu},\zeta}[\ln p(\mathbf{Y},\mathbf{Y}_0,\mathbf{Z},\mathbf{b},\boldsymbol{\gamma},\mathbf{w},\boldsymbol{\Phi},\boldsymbol{\nu},\zeta)] + \text{const.} \\ &= \mathbb{E}_{\mathbf{Y}_0,\mathbf{Z},\mathbf{b}}[\ln p(\mathbf{Y},\mathbf{Y}_0|\mathbf{Z},\mathbf{b},\boldsymbol{\gamma})] + \ln p(\boldsymbol{\gamma}) + \text{const.}\end{aligned} \quad (6\text{-}18),$$

$$\begin{aligned}\ln q^{\star}_{3\zeta}(\zeta) &= \mathbb{E}_{\mathbf{Y}_0,\mathbf{Z},\mathbf{b},\boldsymbol{\gamma},\mathbf{w},\boldsymbol{\nu}}[\ln p(\mathbf{Y},\mathbf{Y}_0,\mathbf{Z},\mathbf{b},\boldsymbol{\gamma},\mathbf{w},\boldsymbol{\Phi},\boldsymbol{\nu},\zeta)] + \text{const.} \\ &= \mathbb{E}_{\boldsymbol{\nu}}[\ln p(\boldsymbol{\nu}|\zeta)] + \ln p(\zeta) + \text{const.}\end{aligned} \quad (6\text{-}19).$$

Then, the parameters were estimated with the variational Bayes method. In the method, the expectations of parameters of q_k were calculated as

$$\mathbb{E}\left[\boldsymbol{\theta}_k|\widehat{\boldsymbol{\theta}}_{-k}\right] = \int \boldsymbol{\theta}_k q\left(\boldsymbol{\theta}_k|\widehat{\boldsymbol{\theta}}_{-k}\right) d\boldsymbol{\theta}_k \quad (6\text{-}20),$$

where $\widehat{\boldsymbol{\theta}}_{-k}$ is a vector of the expectation of the parameters of the distribution other than q_k. By repeating the estimation of the expected values of the parameters (Eq. 6-20) of q_k for each k = 1, 2, and 3, the parameter values will converge to the joint distribution (Eq. 6-11). The estimation was applied to each block's data in the actual execution, and the estimation procedure was repeated ten times.

6.2.2 Interpolation of soil moisture

Since the soil moisture was measured in limited locations and dates, the soil moisture in all the plots and dates was interpolated. Spatial and temporal similarities of the soil moisture were modeled with the kernel method:

$$y_{i,j,d} = \sum_{j'}\sum_{d'} k_s\left(l_j - l_{j'}\right) k_t(d-d') I(i=i')\, y_{i',j',d'} \quad (6\text{-}21),$$

where y is the soil moisture data of row i, plot j on day t, $k_s(.)$ is the kernel function for spatial effect, $k_t(.)$ is the kernel function for temporal effect, l_i is the location (integer) of plot i, and $I(.)$ is a function to select the same row where $I(0) = 1$ and otherwise 0. It means that the soil moisture data measured only in the same row was used for the estimation because the soil moisture condition was assumed to depend on each watering tube (i.e., to be independent among different watering tubes). The Gaussian kernel was used for $k_s(.)$ and $k_t(.)$, thus

$$k_s(x) = \exp(-x^2/\lambda_s) \quad (6\text{-}22),$$

$$k_t(x) = \exp(-x^2/\lambda_t) \quad (6\text{-}23).$$

The bandwidths of the kernels, λ_s and λ_t, were selected from the candidate values (17 values from 0.1 to 1000) each year using leave-one-out cross-validation. Root mean squared errors (RMSE) were used to evaluate the accuracy of the estimation in cross-validation.

6.2.3 Daily growth model

Two models were used to explain the growth curve of the canopy area and height. One model is the Random forest (RF) model (Breiman, 2001), where daily growth was modeled with machine learning using genotypic and environmental factors as predictors,

$$y_{i,d} = \mathrm{RF}\left(\hat{y}_{i,d-1}, \mathbf{a}_i, \mathbf{w}_{d-1}, s_{i,d-1}, d\right) + e_{i,d} \quad (6\text{-}24),$$

where RF(.) is a function of RF, $\hat{y}_{i,d-1}$ is a predicted value of the canopy area or height on day $d-1$, $\mathbf{a}_i$ is a column vector of genotypic relationship matrix **A** corresponding to the genotype in plot i, $\mathbf{w}_{d-1}$ is a vector of the weather variables on day $d-1$, $s_{i,\,d-1}$ is a measure of the soil moisture. Thus, this model includes the effect of genotype ($\mathbf{a}_i$), environment ($\mathbf{w}_{d-1}$, $s_{i,d-1}$), plant size one day before ($\hat{y}_{i,d-1}$), and growth stage (d). The column vector of **A** was used to represent the genotypic effect since the original whole-marker genotype data is too large to evaluate. The weather variables, $\mathbf{w}_{d-1}$, consisted of the minimum, mean, and maximum temperature, solar radiation, and transpiration. When training the model, the measured value of the canopy area or height on day $d-1$, $y_{i,d-1}$, was used instead of $\hat{y}_{i,d-1}$ as an input.

Another model is the additive spline (AS) model, where daily growth was expressed with a statistical model,

$$\Delta y_{i,d} = y_{i,d} - \hat{y}_{i,d-1} = \sum_{k=1}^{5} \mathrm{SP}_{k,i,d}\left(w_{k,d-1}\right) + \mathrm{SP}_{6,i,d}\left(s_{i,d-1}\right) + e_{i,d} \quad (6\text{-}25),$$

where $\Delta y_{i,d}$, $\mathrm{SP}_{k,i,d}(.)$ is the spline function for variable k and plot i on day d, $\mathrm{w}_{k,d-1}$ is a k^{th} weather variable on day $d-1$. Since the spline function can be described with a linear connection of the basis functions,

$$\mathrm{SP}_{k,i,d}\left(w_{k,d-1}\right) = \sum_{q=1}^{Q} c_{q,k,i,d}\,\phi_q\left(w_{k,d-1}\right) \quad (6\text{-}26),$$

where Q is the number of basis functions of the spline, ϕ_q is q^{th} basis function, and $c_{q,k,i,d}$ is its coefficient. Here, the B-spline basis function was used as ϕ_q. Three values (3, 4, 5) were assigned to the number of the basis functions, Q.

The varying coefficient model (Hastie et al., 2009) was used to estimate coefficients $c_{q,k,i,d}$ with consideration of genotypic relationship. It was assumed that the coefficients $c_{q,k,i,d}$ of genetically close genotypes were similar, and those on close observation dates were similar. In other words, the coefficients are assumed to vary among genotypes and dates smoothly. These assumptions can be reflected in the estimation criterion of the coefficients. Here, the minimization of the weighted least squares was used for the estimation,

$$\min_{\substack{c_{q,k,i,d} \\ k=1,\ldots,6 \\ q=1,\ldots,Q}} \sum_{i'=1}^{N} \sum_{d'=1}^{D} a_{i,i'} k_d (d-d') \left(y_{i',d'} - \hat{y}_{i,d-1} - \Delta\hat{y}_{i,d}\right)^2 \quad (6\text{-}27),$$

$$k_d(x) = \exp(-x^2/\lambda_d) \quad (6\text{-}28),$$

where $a_{i,i'}$ is an element of matrix **A** corresponding to the genotypes of plot i and i', $k_d(.)$ is the Gaussian kernel of observation date, $y_{i',d'}$ is an observed value of plot i' on day d', $\Delta\hat{y}_{i,d}$ is a fitted value of Eq. 6-25 of plot i on day d, and λ_d is the kernel width. Therefore, data of variety i' or day d' were weighted based on the genotypic similarity $a_{i,i'}$ and closeness of date $k_d(d - d')$ when estimating the daily growth of variety i or day d. Three values (10, 30, 60) were assigned to the kernel bandwidth, λ_d. For ease of the calculation, zero was assigned to $a_{i,i'}$ if it was smaller than zero.

The model performances were assessed using the prediction accuracy with leave-one-environment-out cross-validation. One combination of the treatment and year (e.g., 2017-WW) was eliminated from the dataset, and the rest were used for model training. Then, the eliminated data was used to validate the prediction accuracy of the trained model. The model accuracy was evaluated with the correlation coefficients between predicted and observed values.

6.3 Results

6.3.1 Filtering of noise and bias

The canopy area and height's estimated true growth processes were smoother than the observed values (Fig. 6-4, 6-5). The measurement bias (b_d) and the probability of bias occurrence (v_d) were estimated for each block, which seemed appropriate for most cases. The correlation coefficients between the estimated bias and the true bias calculated from the canopy height measured by hand were more than 0.5 in 9 cases out of 12 (Fig. 6-6). The correlation was low in the estimation of the 2019 data. The estimated bias was larger than the ground-truth values in 2017 and 2018, whereas they were smaller in 2019. On the other hand, the growth curves estimated by the model (Fig. 6-4c, 6-5c) were not much different compared with those estimated by the simple fitting of B-splines (Fig. 6-4g, 6-5g).

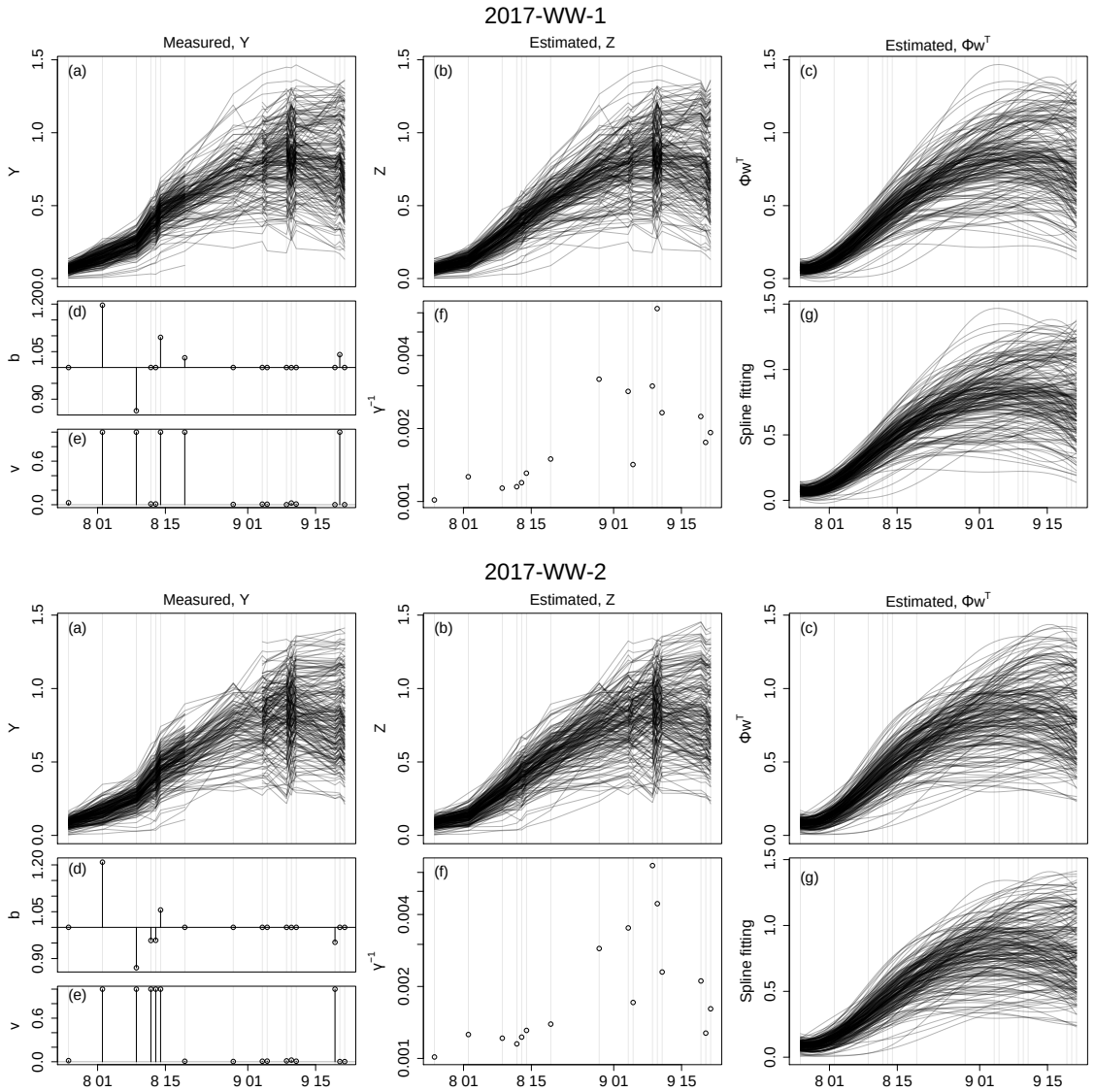

Figure 6-A4-1. The result of the noise and bias filtering of the canopy area in 2017-WW. (a–c) Observed canopy area (**Y**), estimated canopy area (**Z**), and estimated growth curve ($\mathbf{\Phi w}^T$). (d–f) The estimated values of the bias (**b**), the probability of bias occurrence (**v**), and the variance of noise ($\boldsymbol{\gamma}^{-1}$). (g) The estimated growth curve by fitting a simple spline. The number of degrees was the same as used in the estimation model.

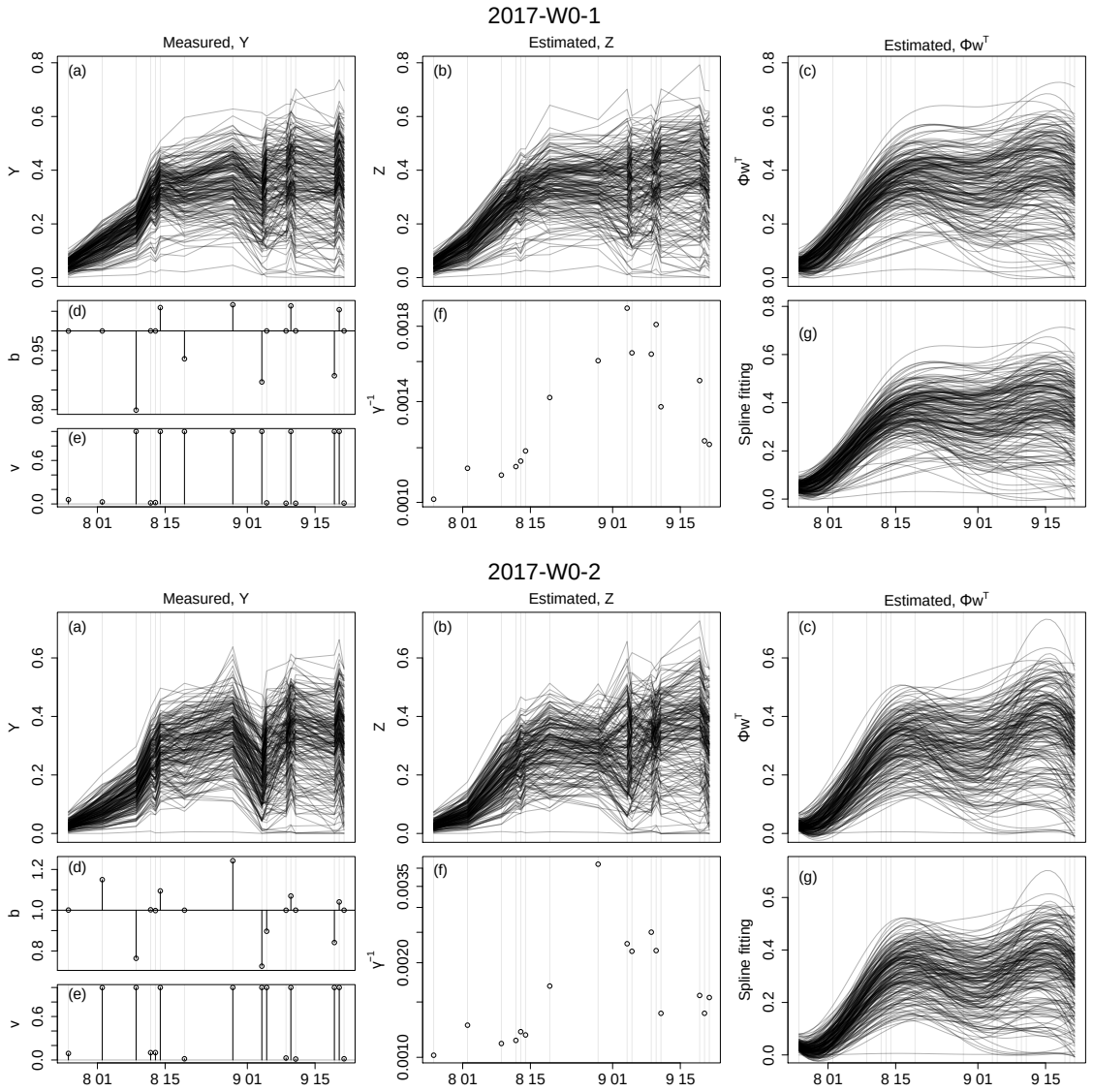

Figure 6-A4-2. The result of the noise and bias filtering of the canopy area in 2017-W0. (a–c) Observed canopy area (**Y**), estimated canopy area (**Z**), and estimated growth curve ($\mathbf{\Phi w}^{T}$). (d–f) The estimated values of the bias (**b**), the probability of bias occurrence (**v**), and the variance of noise ($\boldsymbol{\gamma}^{-1}$). (g) The estimated growth curve by fitting a simple spline. The number of degrees was the same as used in the estimation model.

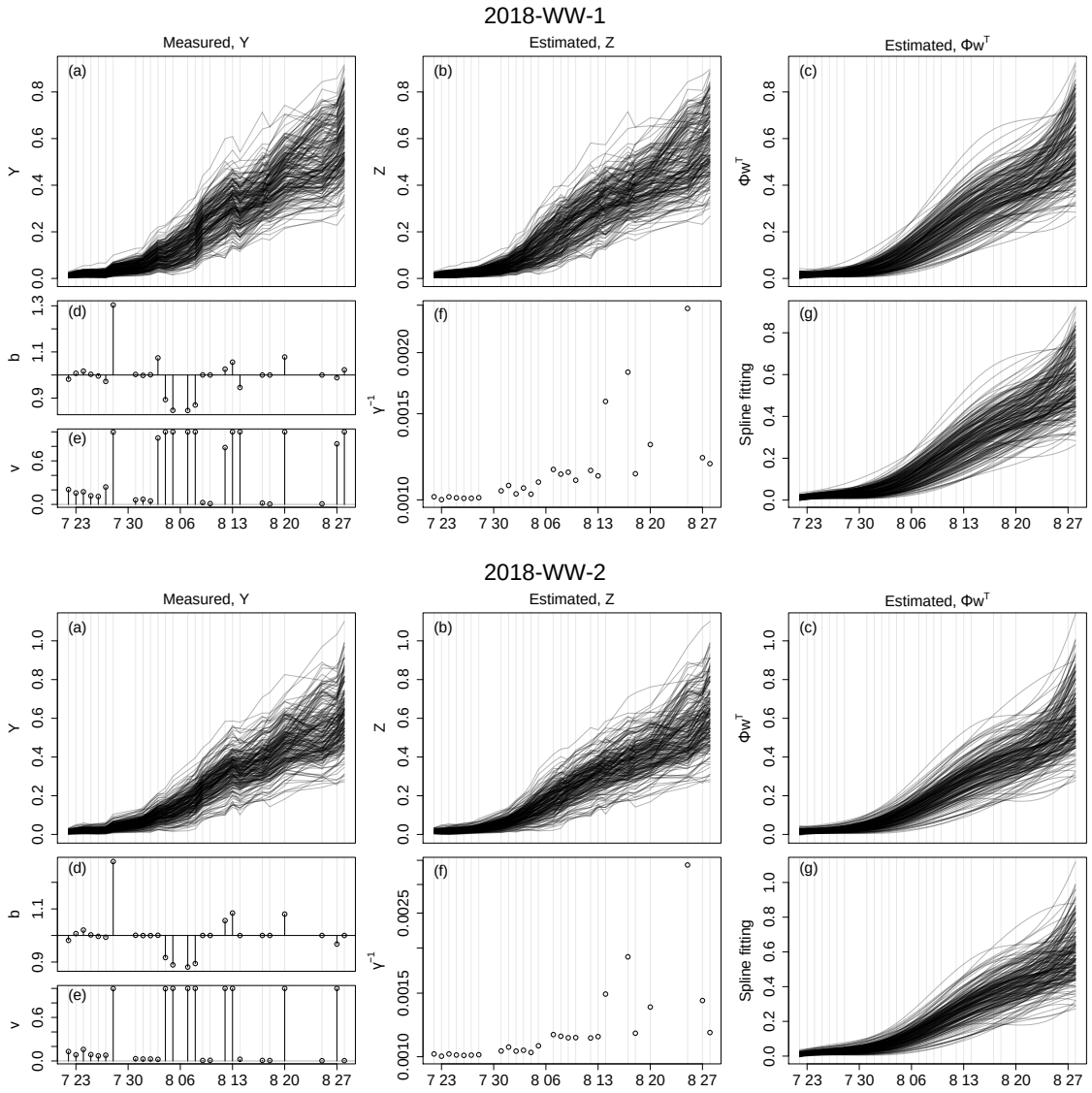

Figure 6-A4-3. The result of the noise and bias filtering of the canopy area in 2018-WW. (a–c) Observed canopy area (**Y**), estimated canopy area (**Z**), and estimated growth curve ($\mathbf{\Phi w}^T$). (d–f) The estimated values of the bias (**b**), the probability of bias occurrence (**v**), and the variance of noise ($\boldsymbol{\gamma}^{-1}$). (g) The estimated growth curve by fitting a simple spline. The number of degrees was the same as used in the estimation model.

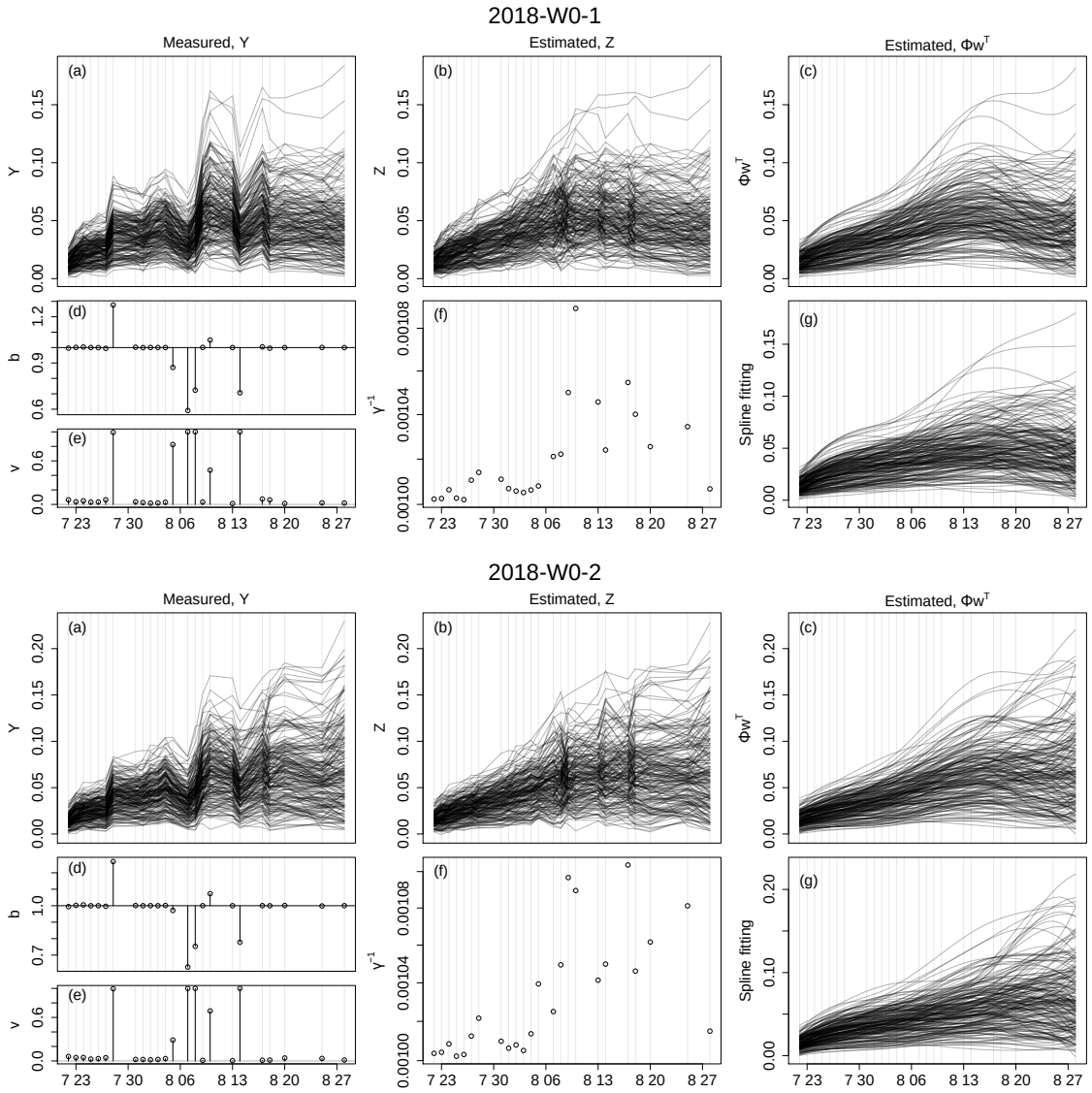

Figure 6-A4-4. The result of the noise and bias filtering of the canopy area in 2018-W0. (a–c) Observed canopy area (**Y**), estimated canopy area (**Z**), and estimated growth curve ($\mathbf{\Phi w}^T$). (d–f) The estimated values of the bias (**b**), the probability of bias occurrence (**v**), and the variance of noise ($\boldsymbol{\gamma}^{-1}$). (g) The estimated growth curve by fitting a simple spline. The number of degrees was the same as used in the estimation model.

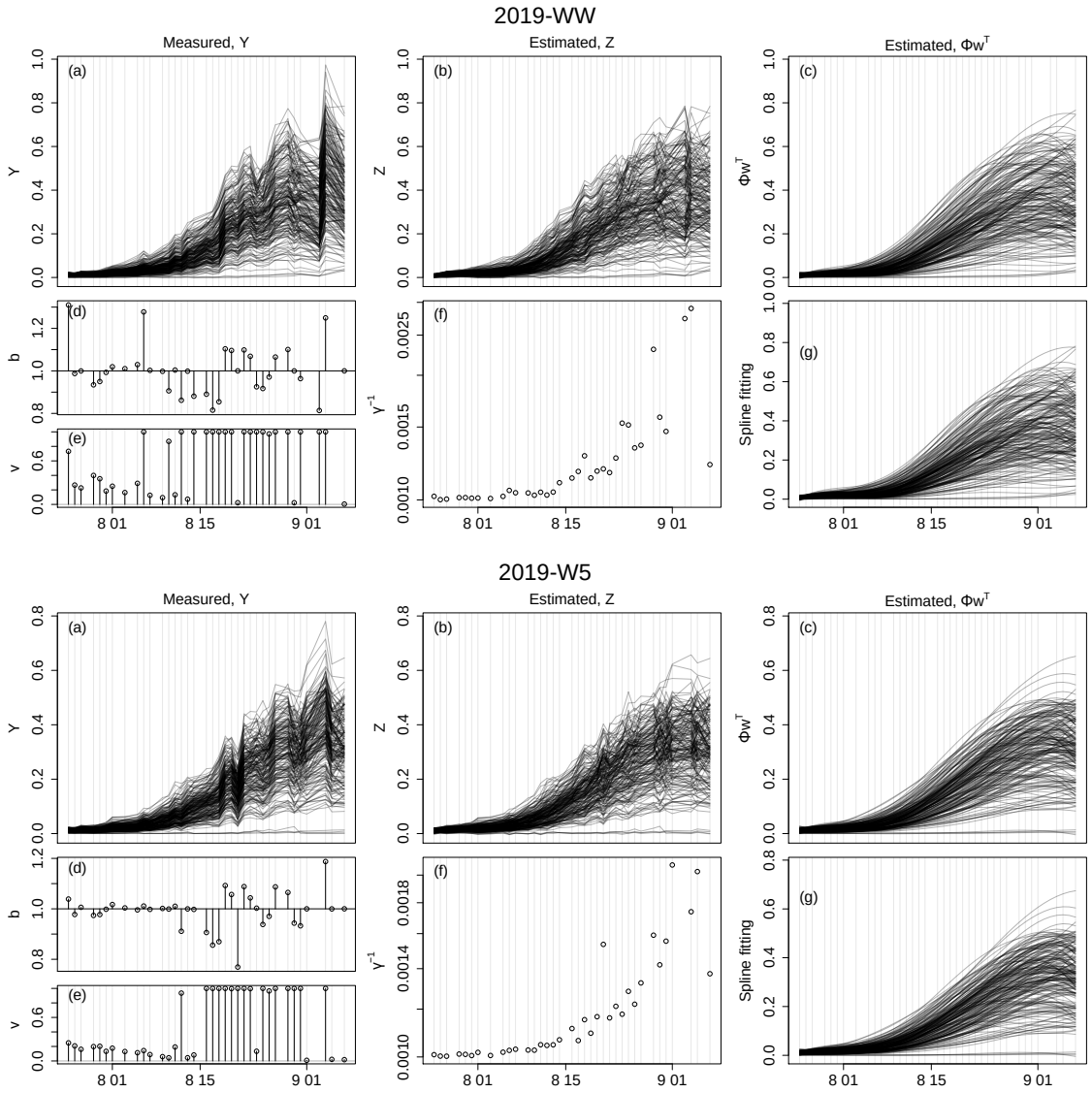

Figure 6-A4-5. The result of the noise and bias filtering of the canopy area in 2019-WW and 2019-W5. (a–c) Observed canopy area (**Y**), estimated canopy area (**Z**), and estimated growth curve ($\mathbf{\Phi w}^T$). (d–f) The estimated values of the bias (**b**), the probability of bias occurrence (**v**), and the variance of noise ($\boldsymbol{\gamma}^{-1}$). (g) The estimated growth curve by fitting a simple spline. The number of degrees was the same as used in the estimation model.

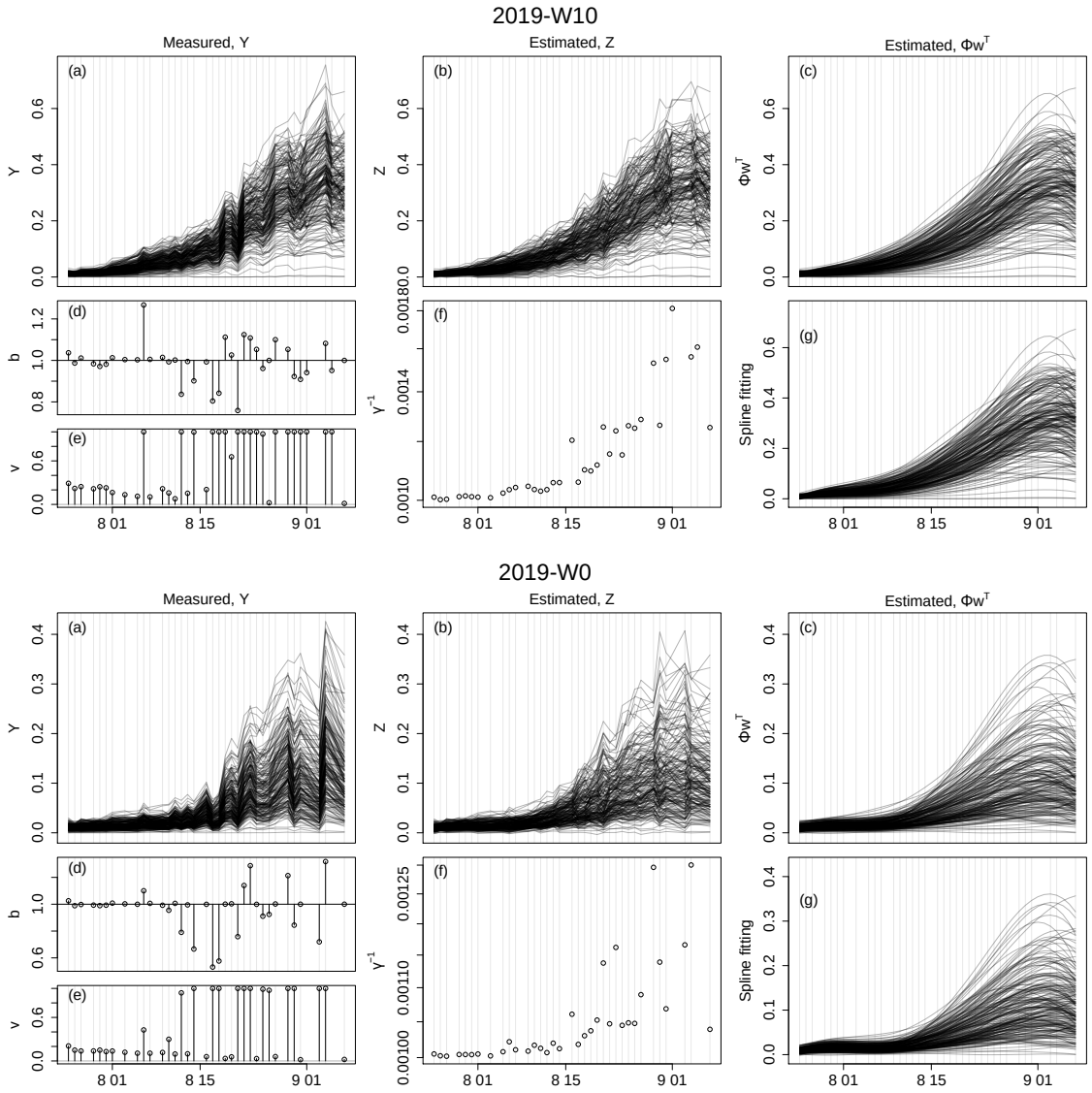

Figure 6-A4-6. The result of the noise and bias filtering of the canopy area in 2019-W10 and 2019-W0. (a–c) Observed canopy area (**Y**), estimated canopy area (**Z**), and estimated growth curve ($\mathbf{\Phi w}^T$). (d–f) The estimated values of the bias (**b**), the probability of bias occurrence (**v**), and the variance of noise ($\boldsymbol{\gamma}^{-1}$). (g) The estimated growth curve by fitting a simple spline. The number of degrees was the same as used in the estimation model.

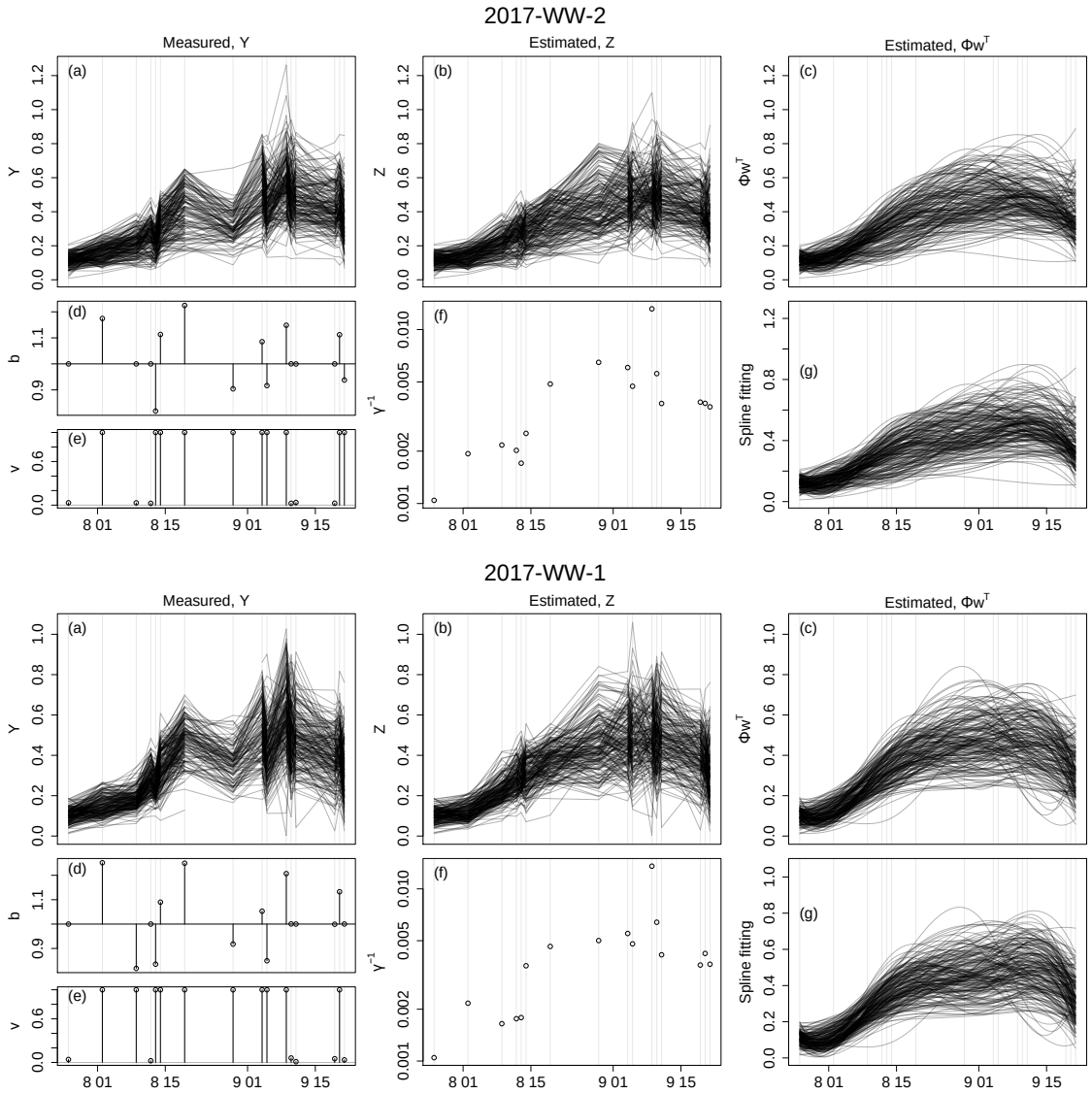

Figure 6-A5-1. The result of the noise and bias filtering of the canopy height in 2017-WW. (a–c) the measured canopy height (**Y**), the estimated values (**Z)**, and the estimated growth curve ($\mathbf{\Phi w}^T$). (d–f) the estimated values of the bias (**b**), the probability of bias occurrence (**v**), and the variance of noise ($\boldsymbol{\gamma}^{-1}$). (g) The estimated growth curve by fitting a simple spline. The number of degrees was the same as used in the estimation model.

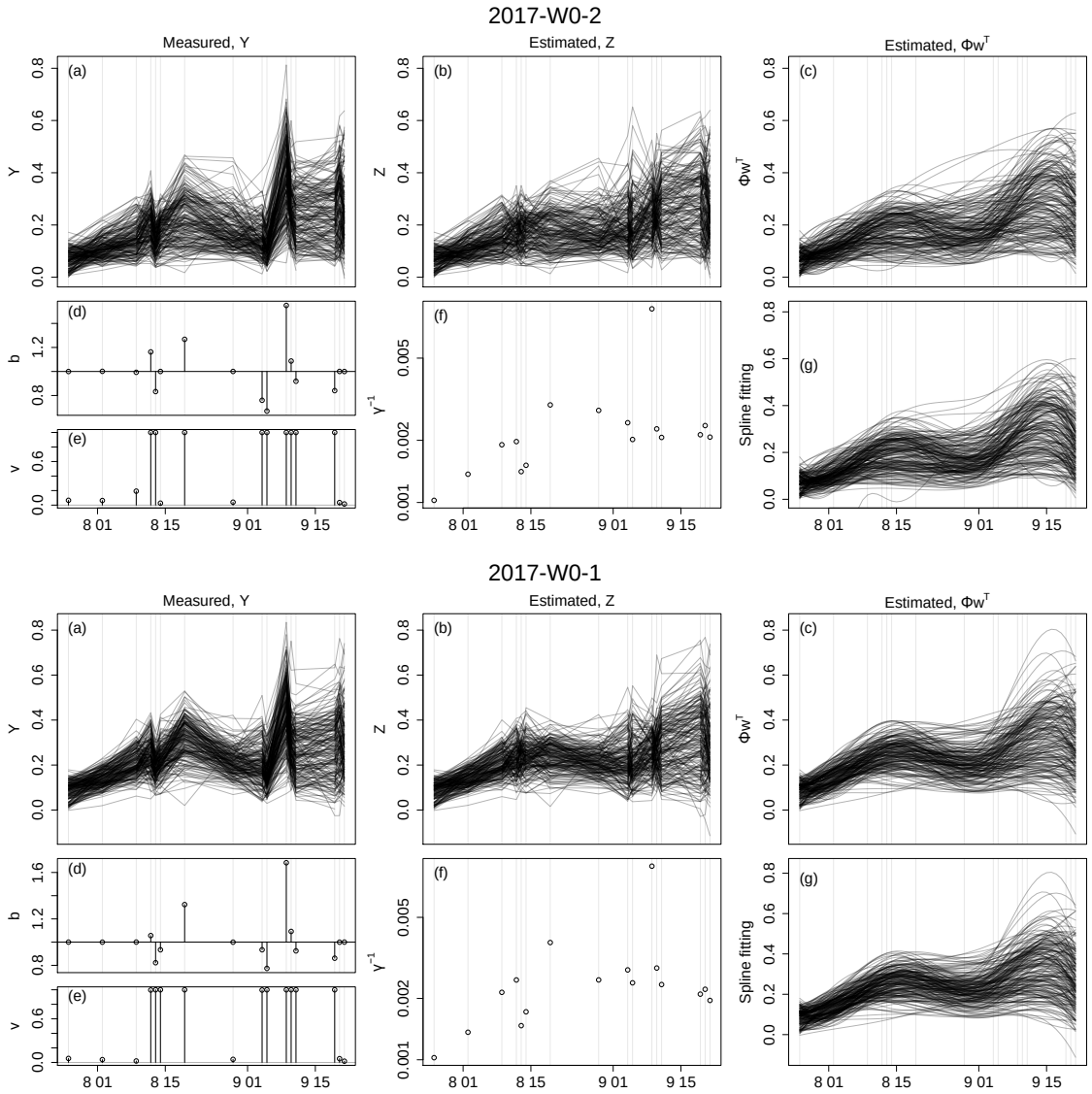

Figure 6-A5-2. The result of the noise and bias filtering of the canopy height in 2017-W0. (a–c) the measured canopy height (**Y**), the estimated values (**Z)**, and the estimated growth curve (**Φw**^{T}). (d–f) the estimated values of the bias (**b**), the probability of bias occurrence (**v**), and the variance of noise ($\boldsymbol{\gamma}^{-1}$). (g) The estimated growth curve by fitting a simple spline. The number of degrees was the same as used in the estimation model.

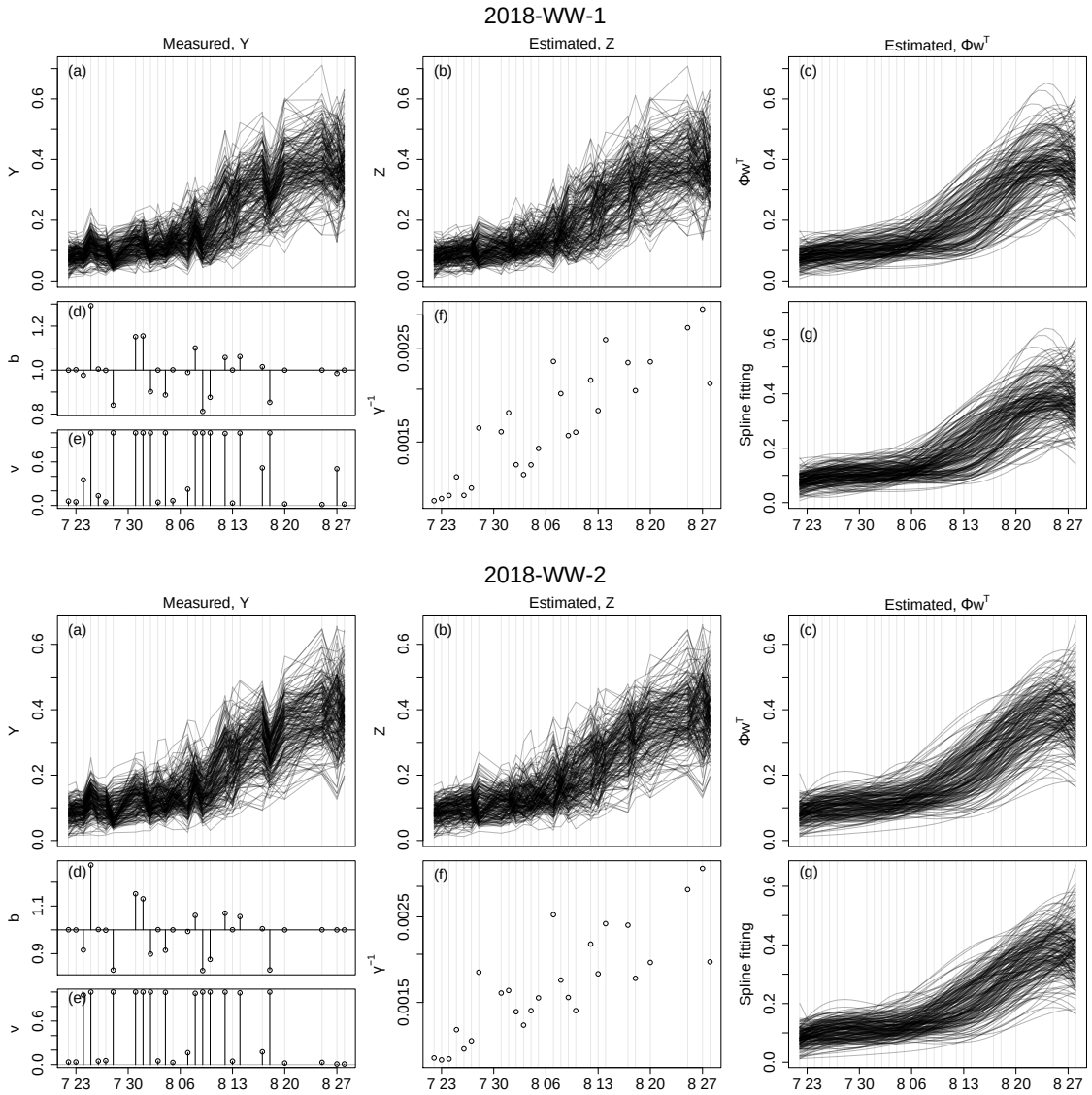

Figure 6-A5-3. The result of the noise and bias filtering of the canopy height in 2018-WW. (a–c) the measured canopy height (**Y**), the estimated values (**Z**), and the estimated growth curve ($\mathbf{\Phi w}^T$). (d–f) the estimated values of the bias (**b**), the probability of bias occurrence (**v**), and the variance of noise ($\boldsymbol{\gamma}^{-1}$). (g) The estimated growth curve by fitting a simple spline. The number of degrees was the same as used in the estimation model.

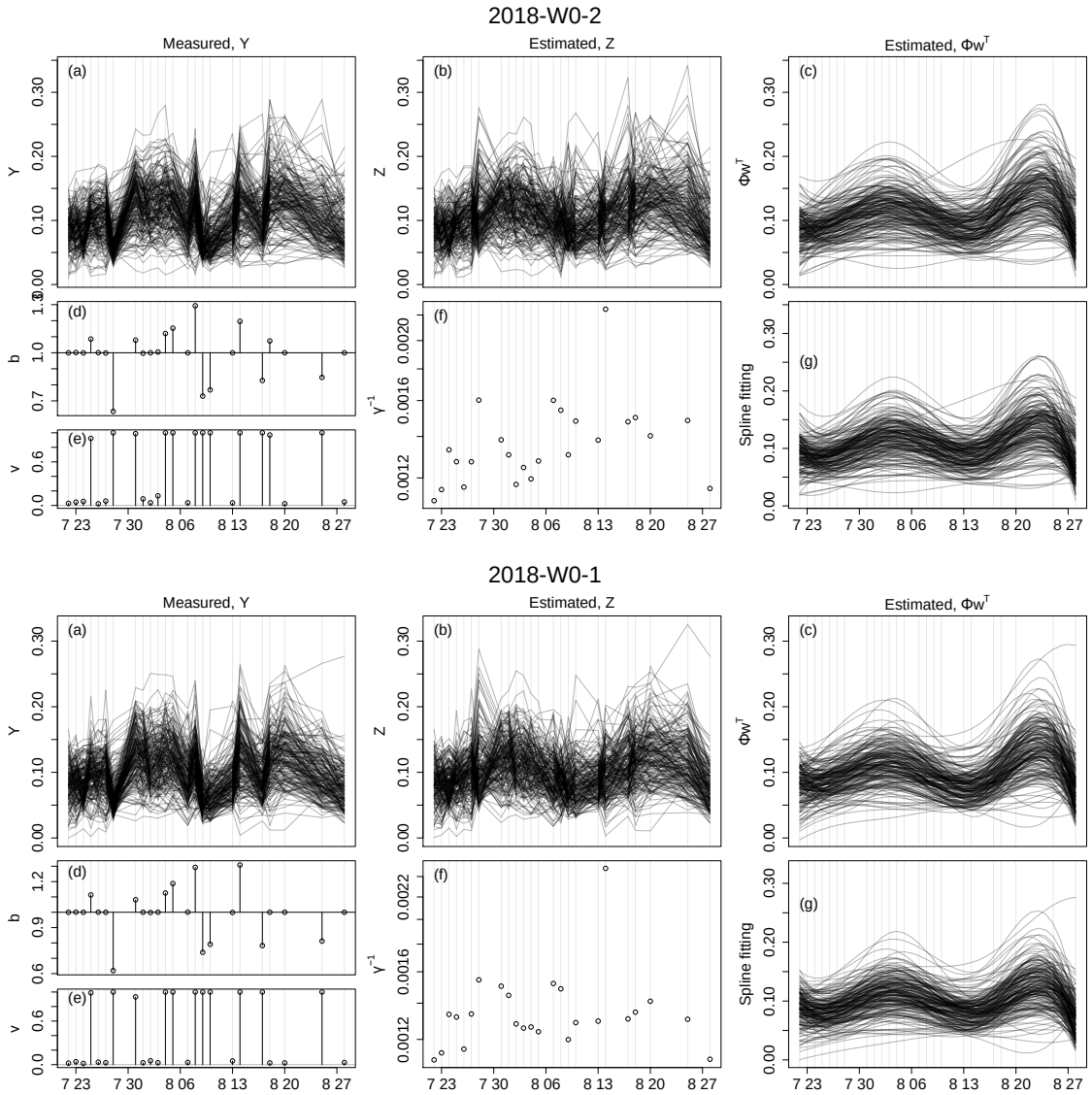

Figure 6-A5-4. The result of the noise and bias filtering of the canopy height in 2018-W0. (a–c) the measured canopy height (**Y**), the estimated values (**Z**), and the estimated growth curve ($\mathbf{\Phi w}^T$). (d–f) the estimated values of the bias (**b**), the probability of bias occurrence (**v**), and the variance of noise ($\boldsymbol{\gamma}^{-1}$). (g) The estimated growth curve by fitting a simple spline. The number of degrees was the same as used in the estimation model.

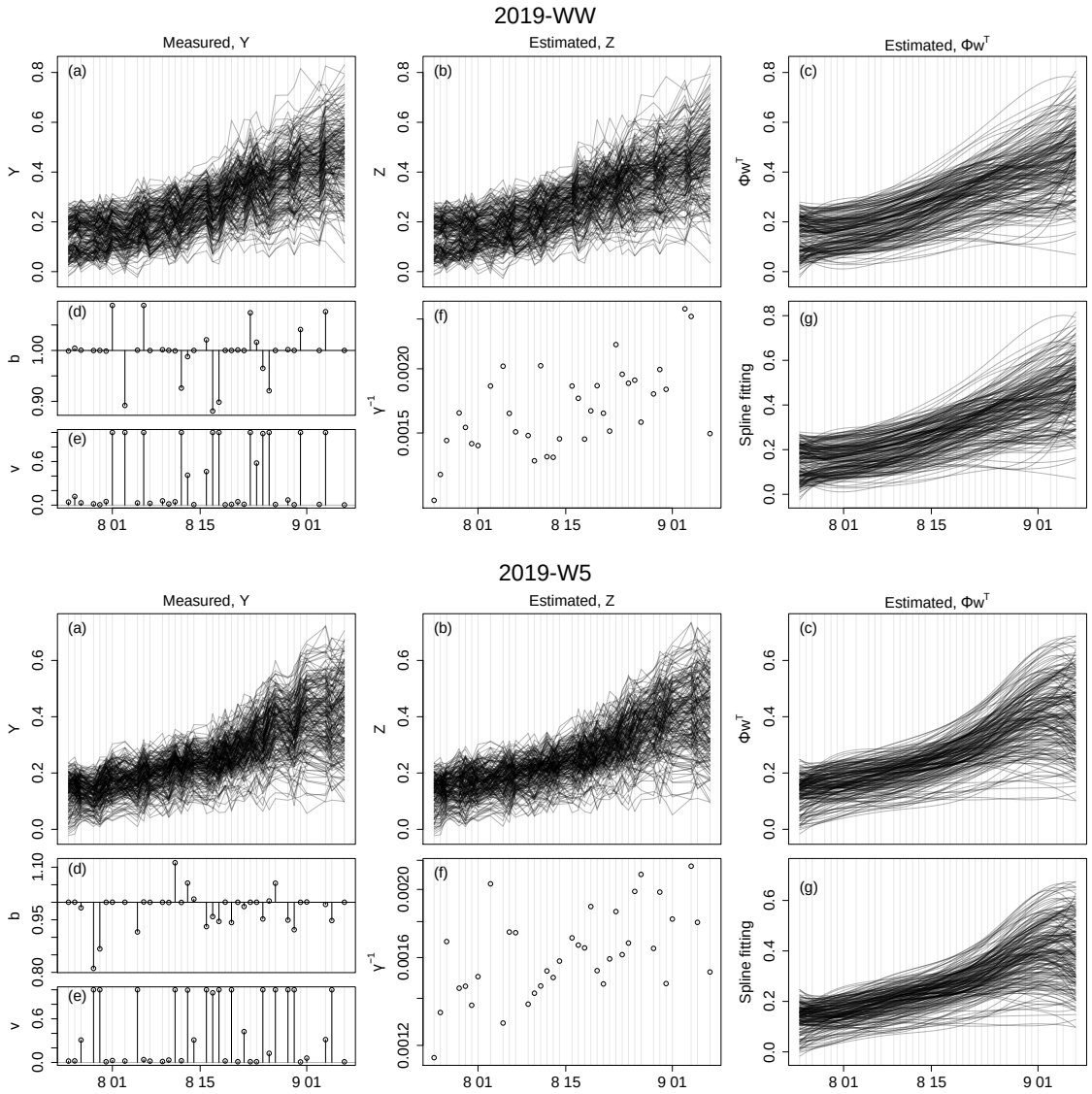

Figure 6-A5-5. The result of the noise and bias filtering of the canopy height in 2019-WW and 2019-W5. (a–c) the measured canopy height (**Y**), the estimated values (**Z**), and the estimated growth curve ($\mathbf{\Phi w}^T$). (d–f) the estimated values of the bias (**b**), the probability of bias occurrence (**v**), and the variance of noise ($\boldsymbol{\gamma}^{-1}$). (g) The estimated growth curve by fitting a simple spline. The number of degrees was the same as used in the estimation model.

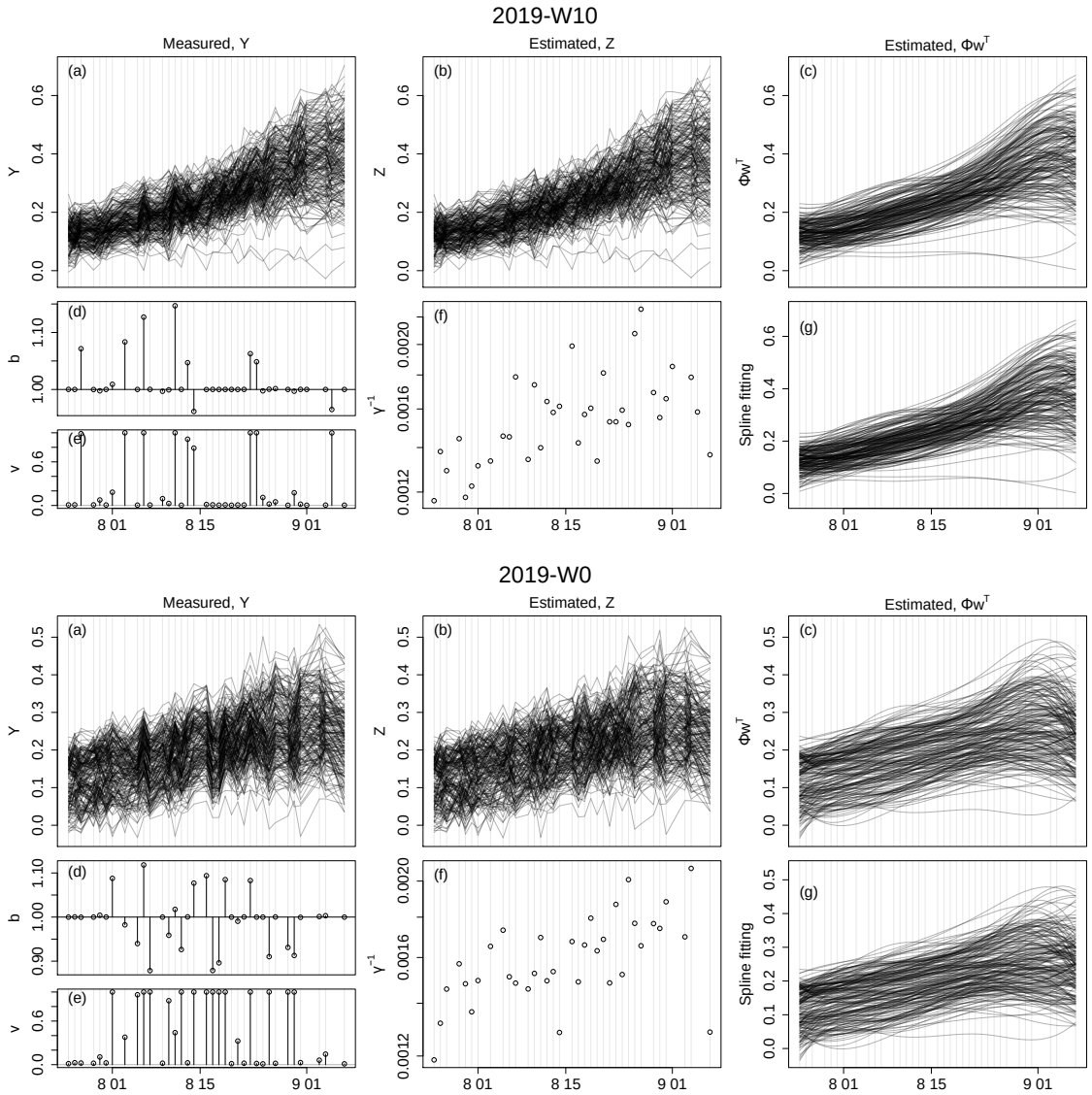

Figure 6-A5-6. The result of the noise and bias filtering of the canopy height in 2019-W10 and 2019-W0. (a–c) the measured canopy height (**Y**), the estimated values (**Z**), and the estimated growth curve ($\mathbf{\Phi w}^T$). (d–f) the estimated values of the bias (**b**), the probability of bias occurrence (**v**), and the variance of noise ($\boldsymbol{\gamma}^{-1}$). (g) The estimated growth curve by fitting a simple spline. The number of degrees was the same as used in the estimation model.

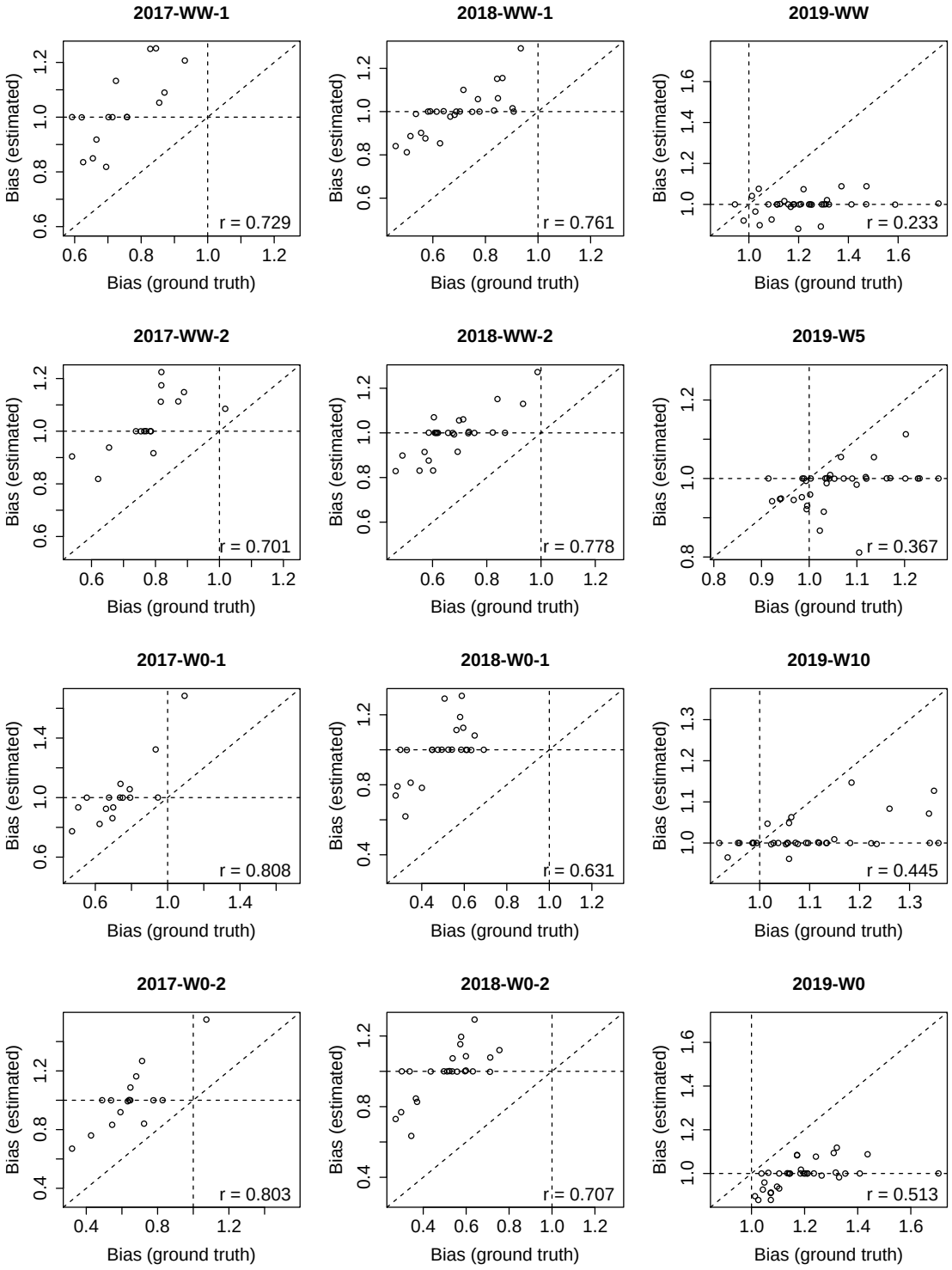

Figure 6-6. Comparison between the bias (**b**) estimated by the model and obtained from manual measurement in each block. Their correlation coefficients (*r*) are shown in the bottom right of each plot.

6.3.2 Interpolation of soil moisture

The selected bandwidths of the temporal effect, λ_t, were stable among years, whereas the selected bandwidths of the spatial effect, λ_s, were smaller in 2017 than those in the other years (Fig. 6-7). The differences in the optimal λ_s among years were caused mainly by the heterogeneity of soil moisture in 2017 in a specific row (Fig. 6-8-1, the lower row of WW-1). The estimated parameters, $\lambda_{t,}$ and $\lambda_{s,}$ were used to estimate the soil moisture in all the plots on all the dates (Fig 6-9). In 2017, the estimated soil moisture converged to the nearest locations' observed values because the kernel function $k_s(.)$ behaved as a nearest-neighborhood estimation with small λ_s.

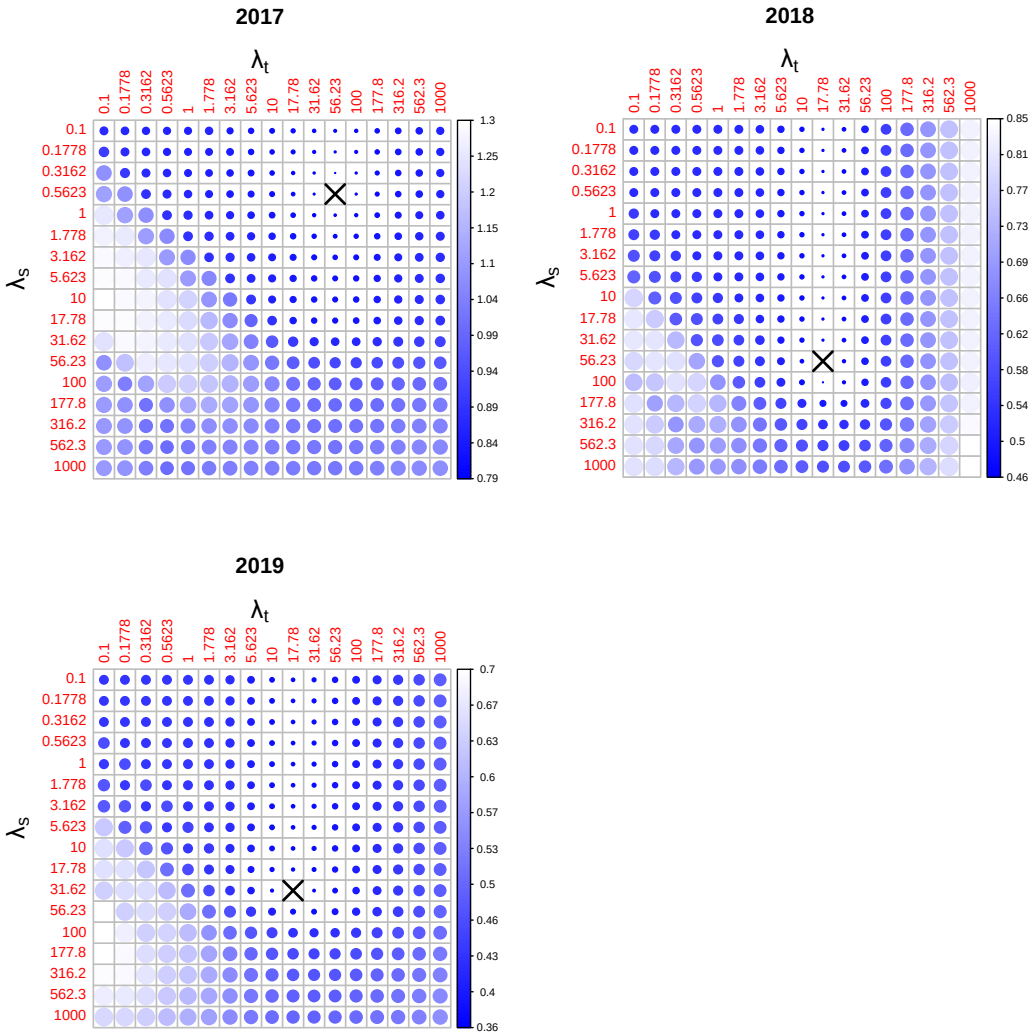

Figure 6-7. Heatmaps of RMSE for variant bandwidths of the kernels, $\lambda_{s,}$ and λ_t. The sets of bandwidths with the lowest RMSE in each year were marked with crosses.

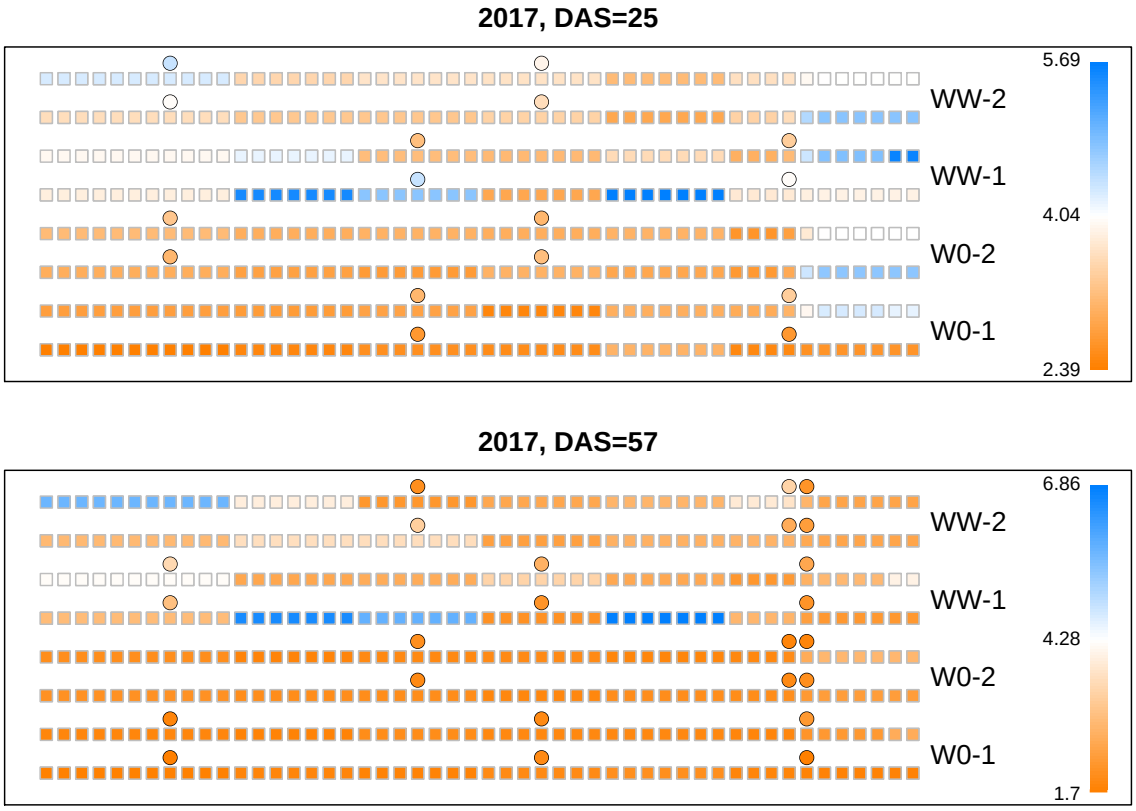

Figure 6-8-1. Heatmap of the interpolated soil moisture in 2017. The values of 25 and 57 days after sowing were chosen. The measured values were plotted as circles.

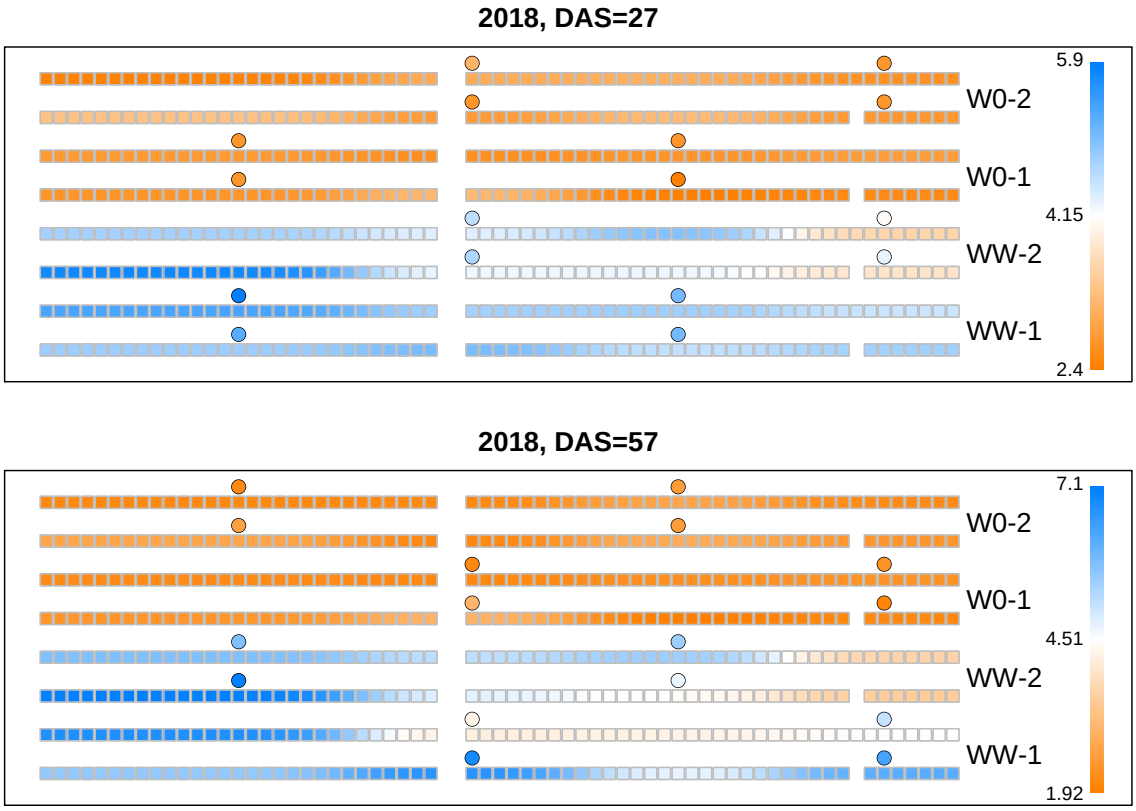

Figure 6-8-2. Heatmap of the interpolated soil moisture in 2018. The values of 27 and 57 days after sowing were chosen. The measured values were plotted as circles.

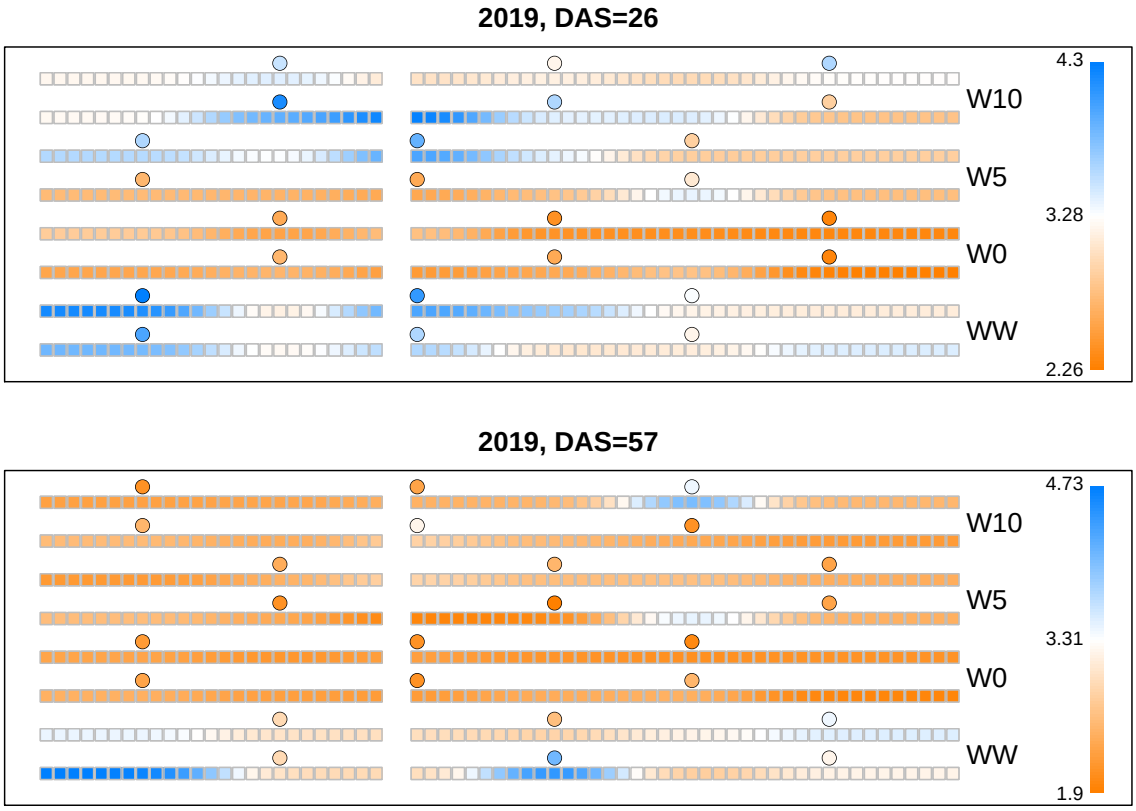

Figure 6-8-3. Heatmap of the interpolated soil moisture in 2019. The values of 26 and 57 days after sowing were chosen. The measured values were plotted as circles.

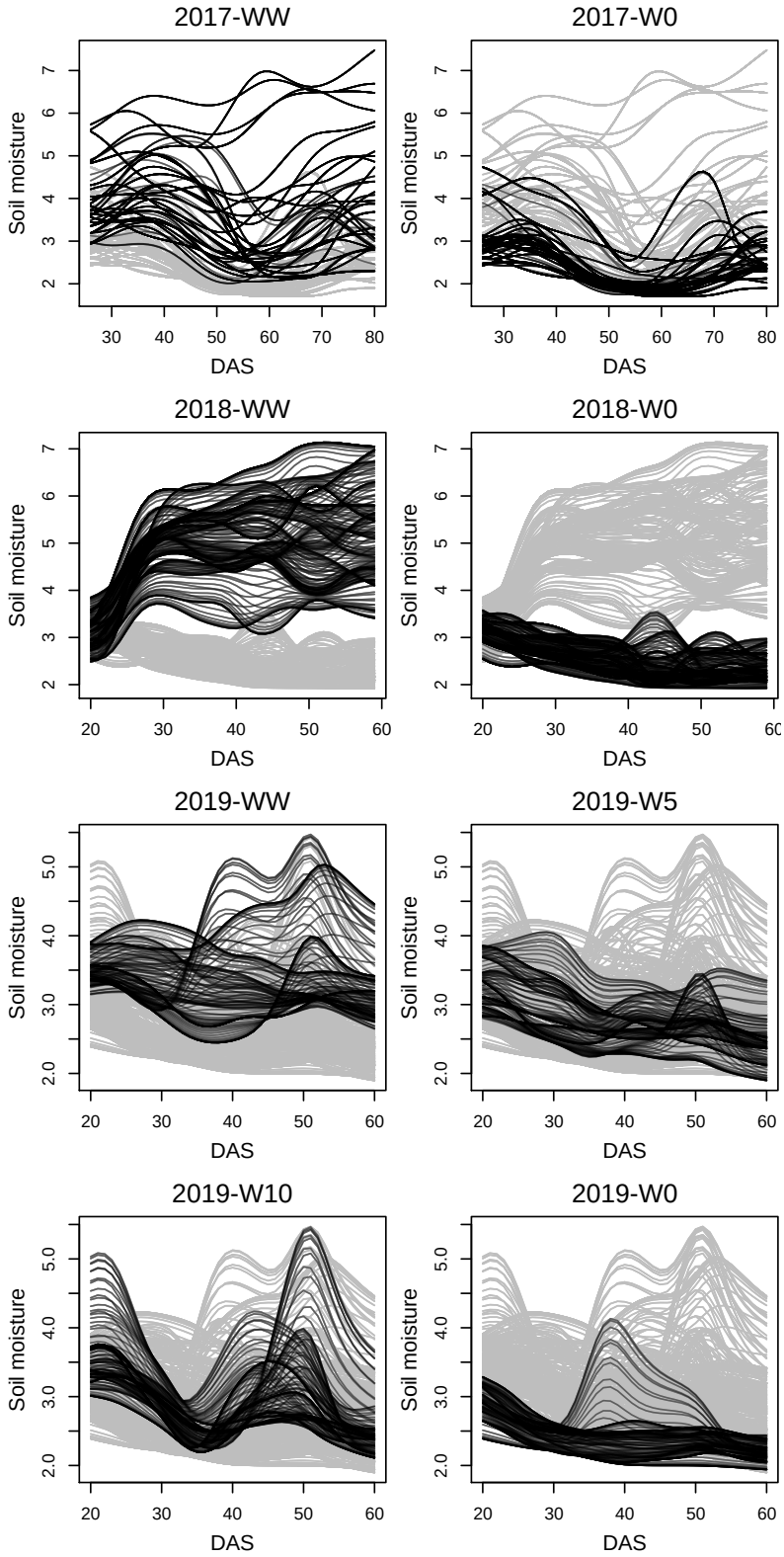

Figure 6-9. Time-series soil moisture interpolated for all the plots. Line colors identified different treatments. Observed values of each treatment were plotted. Phenotypic data of the same year was plotted with gray lines.

6.3.1 Daily growth model

First, nine sets of hyperparamterses (Q, λ_d) of the AS modelwere compared (Fig. 6-10). The best hyperparameters were different for each observation or environment, but (Q, λ_d) = (4, 10) and (5, 10) were the best hyperparameters in most cases for the prediction of canopy area and height, respectively. In the following paragraphs, these hyperparameters were used as a representative of the AS model.

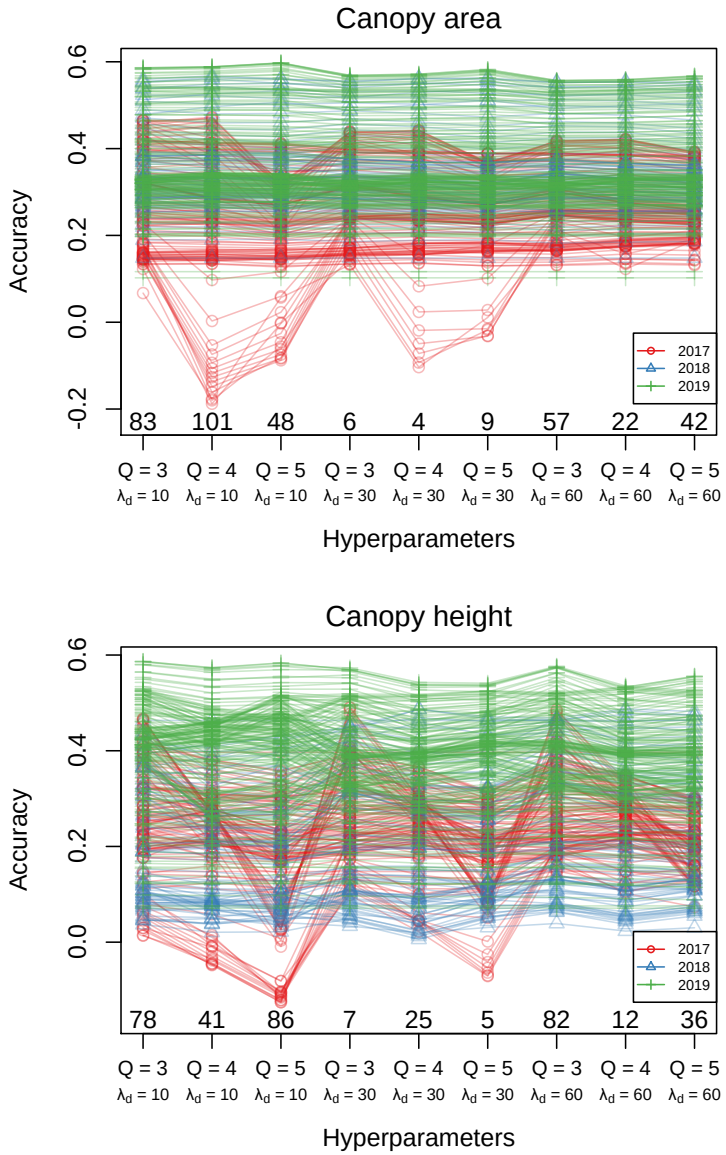

Figure 6-10. Comparison of the AS model's prediction accuracy with various sets of hyperparameters (Q and λ_d). The correlation coefficients of the predicted and observed values are used as the y-axis. The numbers at the bottom of each plot are the number of points, the accuracy of which was the best with each set of hyperparameters.

The RF model exceeded the AS model in predicting accuracy in almost all environments (Fig. 6-11). The accuracy of the RF model was higher than that of the AS model. The AS model's accuracy was especially low in the later growth period in 2017, where the data was not supplied from the other environments.

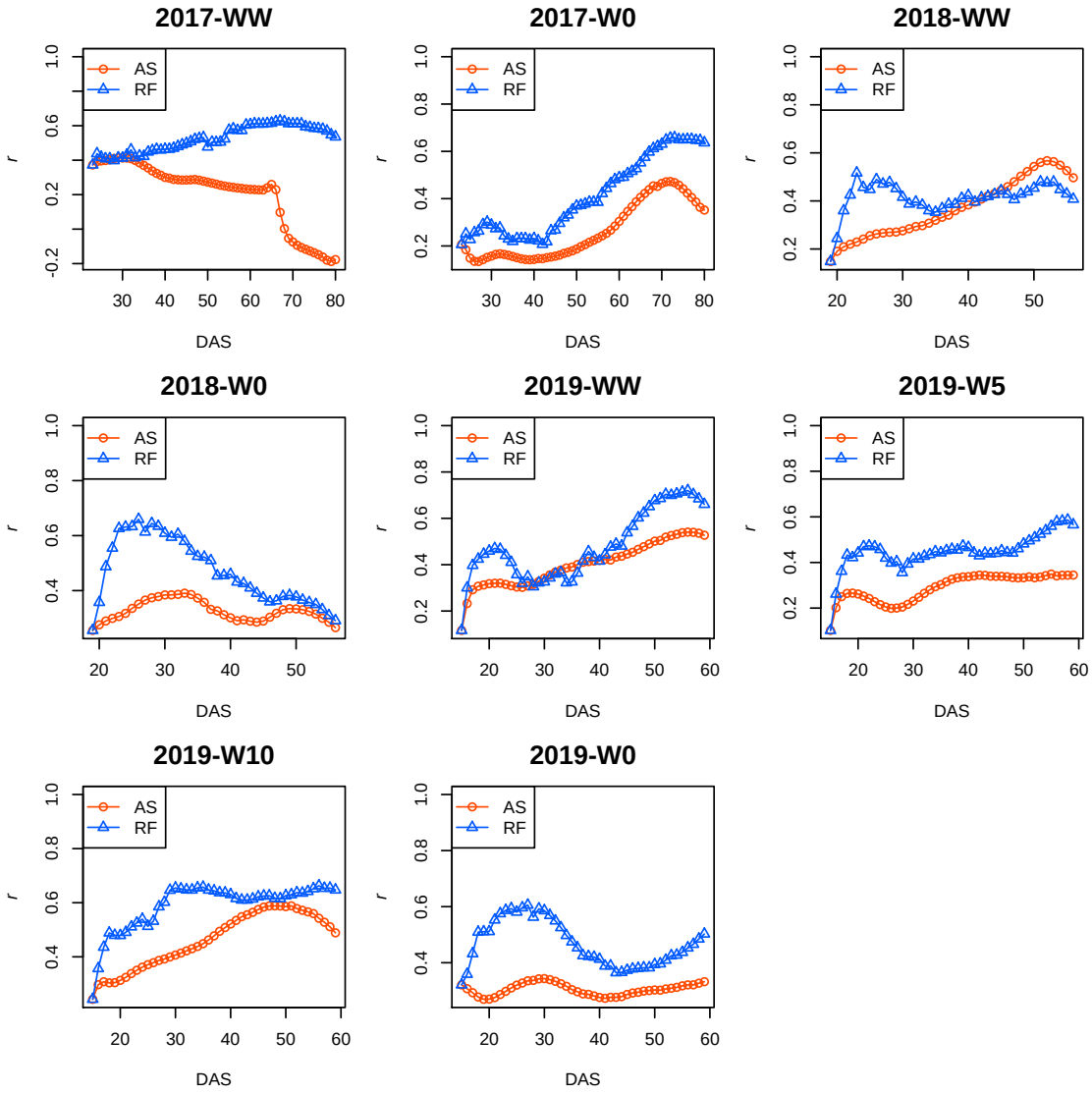

Figure 6-11. Prediction accuracy of the daily growth of the canopy area using the RF and AS models. The accuracy was measured as the correlation coefficient between observed and predicted values and plotted for each environment. For the AS model, the hyperparameters $Q = 4$ and $\lambda_d = 10$ were used. Red and blue lines correspond to the accuracy of the RF and the AS models, respectively.

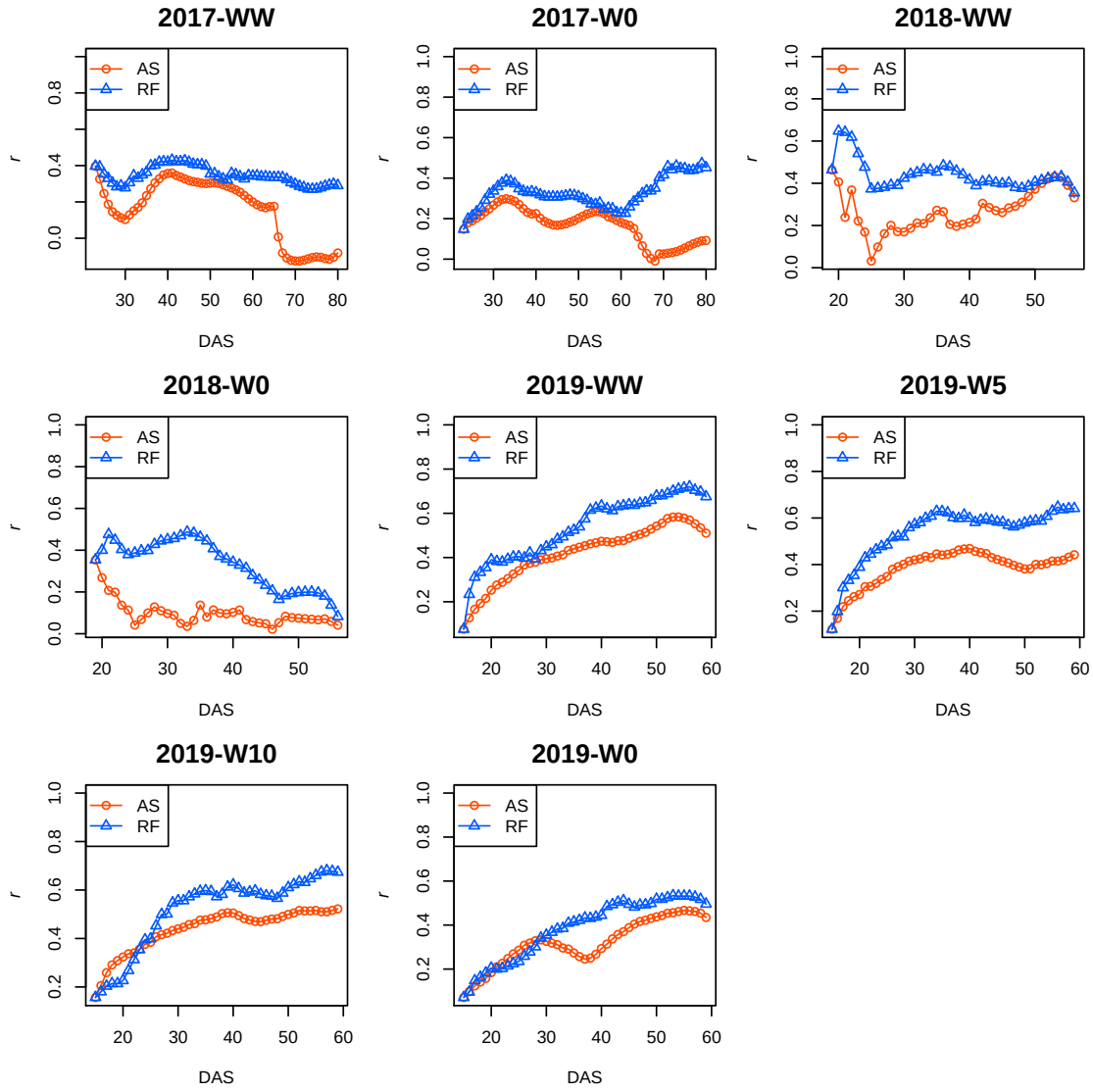

Figure 6-12. Prediction accuracy of the daily growth of the canopy height using RF and AS models. The accuracy was measured as the correlation coefficient between observed and predicted values and plotted for each environment. For the AS model, the hyperparameters $Q = 5$ and $\lambda_d = 10$ were used. Red and blue lines correspond to the accuracy of the RF and the AS models, respectively.

The estimated environmental response in each plot was drawn as a curve using the estimated coefficients. The curves of all plots were overlayed in Fig. 6-13, 6-14. There were tendencies that daily growth increased with soil moisture in the canopy area and height, whereas common tendencies could not be found in the other factors. The estimated curves were not stable where the data were not obtained around the day.

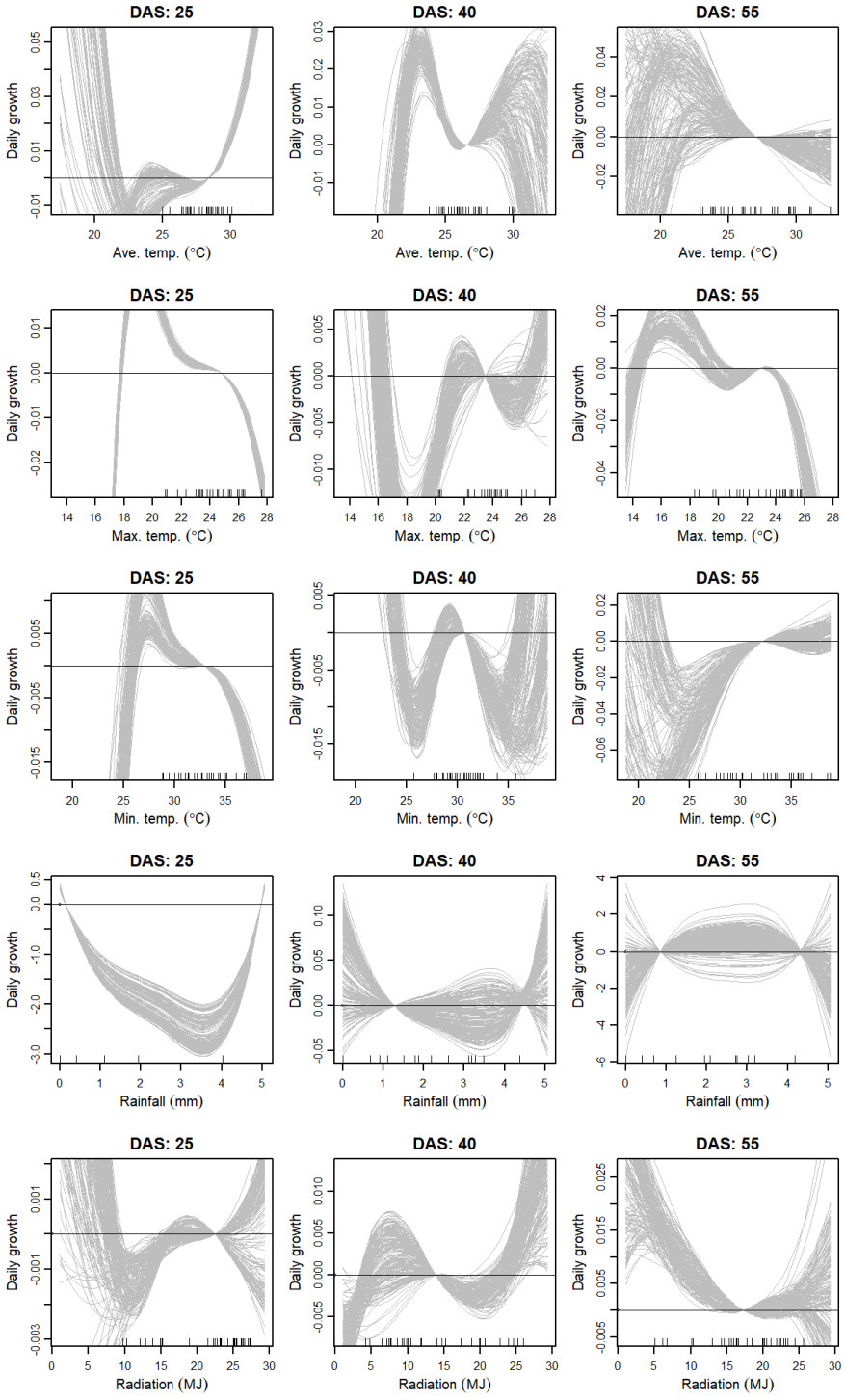

Figure 6-13-1. The estimated relationship between environmental factors (except for the soil moisture) and the canopy area's daily growth. The results at the hyperparameter values of $Q = 4$ and $\lambda_d = 10$ were selected. The drawn curves were scaled so that the y-value becomes zero when the predictor is equal to its mean values. Tick marks were put on the data points observed five days before and after. Six dates were selected for the visualization based on days after sowing (DAS).

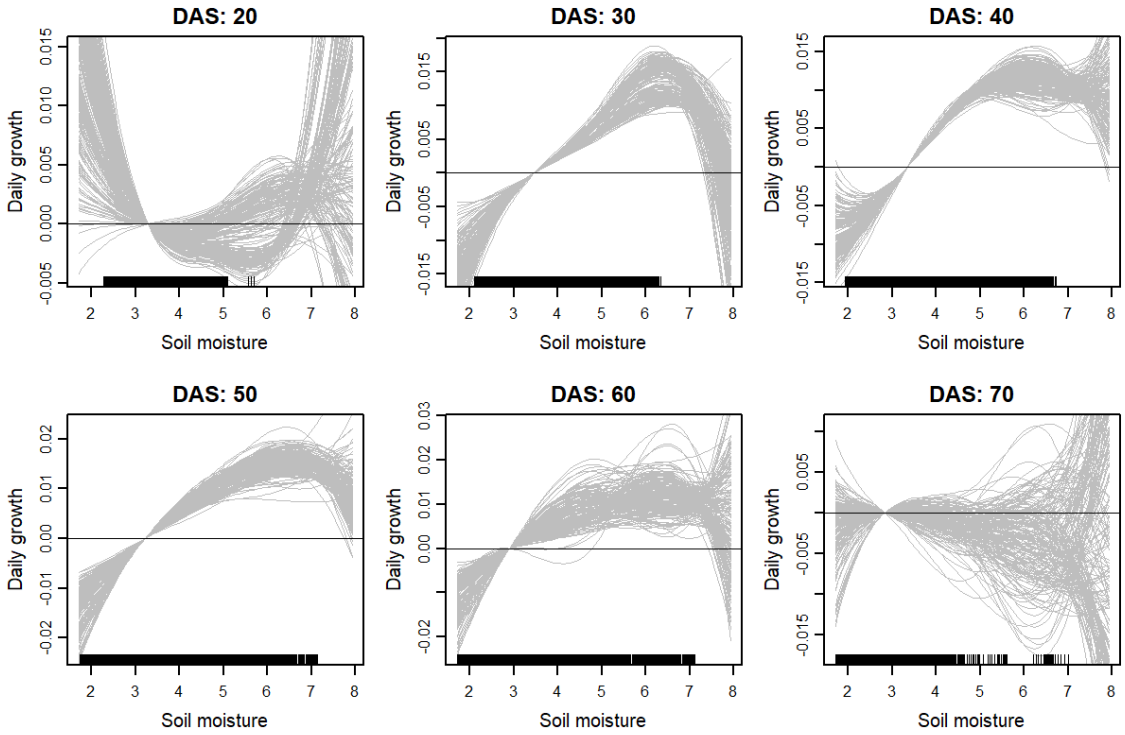

Figure 6-13-2. The estimated relationship between the soil moisture and the daily growth of the canopy area. The results at the hyperparameter values of $Q = 4$ and $\lambda_d = 10$ were selected. The drawn curves were scaled so that the y-value becomes zero when the predictor is equal to its mean values. Tick marks were put on the data points observed five days before and after. Six dates were selected for the visualization based on days after sowing (DAS). Data in 2017 and 2018 were only available in 20 DAS, and data in 2019 was only available in 60 and 70 DAS.

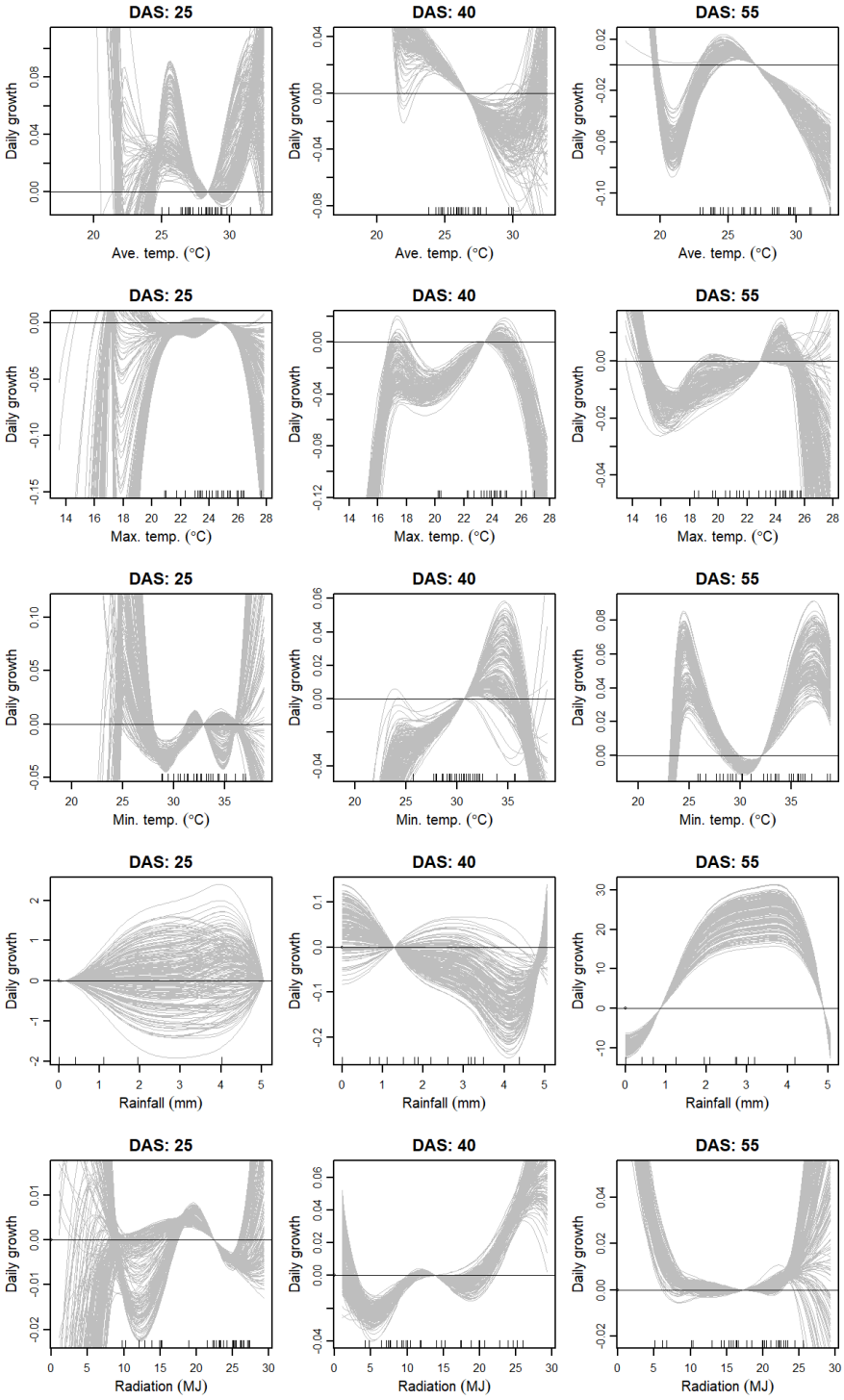

Figure 6-14-1. The estimated relationship between environmental factors (except for the soil moisture) and the canopy height's daily growth. The results at the hyperparameter values of Q = 5 and λ_d = 10 were selected. The drawn curves were scaled so that the y-value becomes zero when the predictor is equal to its mean values. Tick marks were put on the data points observed five days before and after. Six dates were selected for the visualization based on days after sowing (DAS).

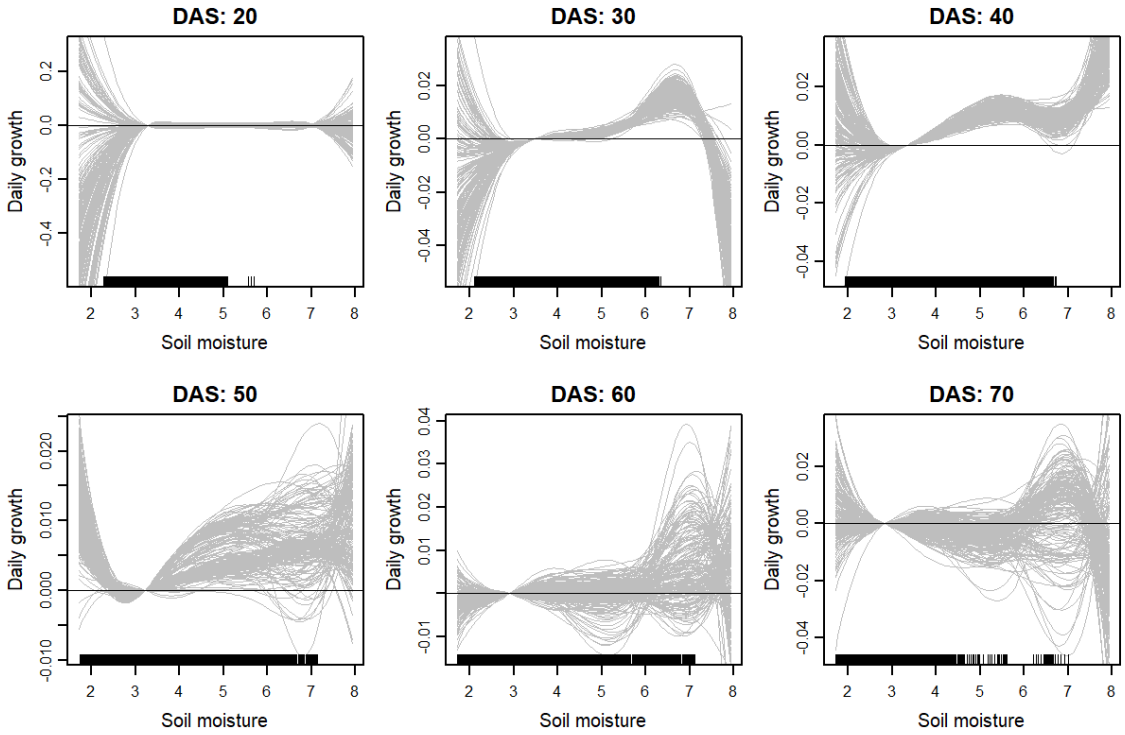

Figure 6-14-2. The estimated relationship between the soil moisture and the daily growth of the canopy height. The results at the hyperparameter values of $Q = 5$ and $\lambda_d = 10$ were selected. The drawn curves were scaled so that the y-value becomes zero when the predictor is equal to its mean values. Tick marks were put on the data points observed five days before and after. Six dates were selected for the visualization based on days after sowing (DAS). Data in 2017 and 2018 were only available in 20 DAS, and data in 2019 was only available in 60 and 70 DAS.

6.4 Discussion

6.4.1 Filtering of noise and bias

Correlation coefficients between the estimated and the ground-truth bias were high ($r > 0.5$ in 2017 and 2018), indicating that the model could yield appropriate estimates about the measurement bias (Fig. 6-6). The estimated bias term (b_d) was higher than the estimated value with ground-truthing. This is because the bias was assumed not to exist basically ($b_d = 1$ if $v_d = 0$), although the actual bias was lower than one in most cases. This problem may be solved if the value of bias is parameterized and estimated when $v_d = 0$.

The estimated results were worse in 2019 than those in the other years. The reason was the difference in the noise structure. As shown in the observed growth curves of the canopy height (Fig. 6-1-2), the growth curves in 2019 fluctuated in every observation, whereas the large bias existed in a few days in 2017 and 2018. In other words, a small bias existed in most of the observations in 2019. It was shown that the bias estimation in such data was difficult using the proposed model.

Although the proposed model included the characteristics of noise and bias of UAV-RS, the estimated growth curves ($\mathbf{\Phi w}^{\mathrm{T}}$, Fig. 6-4 and 6-5) were similar to those estimated by simply fitting splines. Further improvement is required for the proposed model to provide results superior to spline fitting. For example, the degree of freedom of the spline function for the true growth curve can be optimized. The automatic selection of the smoothness enables the closer estimation of true growth curves. Such a method might improve the results in 2017-C, in which the estimated curves were wavy and unreliable. The likelihood or AIC calculated using the estimated expectation of the parameters can be used for the optimization. The smoothing splines may also be useful for implementing the auto-determination of the smoothness since a single continuous smoothing parameter can determine the degree of freedom of the smoothing spline.

Consideration of other growth functions may lead to more robust estimation. For the growth functions, such as the Gompertz curves and the function used in Chapter 5, the growth curve's shape can be determined in advance. This determination would lead to a more robust estimation of the noise and bias against the influence of noise. It is also possible to include the AS model defined in Section 6.2.4. Although it requires a complex algorithm and much computation time, it would provide an advanced method to analyze the UAV-RS model that simultaneously includes the structure of measurement noise and the plant growth daily response.

6.4.2 Modeling of genetic and environmental effects on daily growth

The AS model was also used to visualize the growth responses to soil moisture (Fig 6-13 and 6-14). The result suggests that the AS model can estimate the impact of environmental factors on target traits. The model may also accurately estimate the growth response to soil moisture when multiple treatment levels and sufficient soil moisture data are available.

For the canopy area, changes in the soil moisture's response curve with days after sowing were observed. The result showed that daily growth was proportional to soil moisture around 30 days after sowing, but saturation in the daily growth appeared 40 days after sowing. It was suggested that the increase of soil moisture no longer affected the canopy area when the soil moisture was more than 5%. This increase may be a correct estimation of the growth response curve, but other reasons are also possible; the growth of the canopy area of many genotypes reached the maximum in WW treatments due to the overlap of the leaves or the maturity. Further analysis of growth curves is necessary to validate these reasons.

On the other hand, it was difficult to find common tendencies in other environmental factors. One reason is that the measurements of the weather variables were common for all the plots, unlike the soil moisture. Another reason is using the kernel function of time to estimate the environmental response (k_d in Eq. 6-27). The estimation of the changing environmental response was attempted by weighing the data around the target date, but the lack and bias of the environmental data occurred. This lack and bias in the data were also caused by the seasonal patterns in temperature and the biased distribution of the transpiration where most values were equal to zero. The removal of the kernel function of time may improve the robustness of the estimation for such environmental factors

Consideration of a longitudinal data structure is an important way to extend the AS model. a robust analysis of the environmental responses would be possible by separating the weather data into the seasonal trend and other micro-environmental patterns. Another issue is the autocorrelation of dependent variables, $y_{i,d}$. The Markov chain model should be formally implemented to explain the relationship between $y_{i,d}$ and $y_{i,d+1}$. For the estimation of the growth parameters, the Kalman filter and smoother can also be applied. Several reports using the Kalman filter for data assimilation of RS and CGMs could be found (Jin et al., 2018), suggesting that such a method will help analyze growth data.

The RF model's prediction accuracy exceeded that of the AS model (Fig. 6-11 and 6-12). It was suggested that the daily growth was affected by the interaction terms among the environmental factors, which was not included in the additive model. However, the RF model might overfit the given data due to the small input data size and the seasonal environmental factors pattern. It is necessary to train a prediction model with larger datasets to precisely validate the RF model prediction ability.

The two models proposed in this study took either a statistical or a machine learning approach, and the estimation results were highly dependent on the data in both cases. These approaches allow the flexible estimation of environmental responses. However, their results would be unreliable if enough data is not supplied or the data's characteristics are not taken into account. Crop growth models (CGMs), in which environmental effects on plant growth are modeled based on plant physiological knowledge, might help overcome the problem. For example, the estimation of daily response to the temperature would be robust by replacing the spline of the AS model with the function used in CGMs as the response curve of temperature. Although it requires a complex parameter estimation procedure, effective integration of statistical or machine-learning models and CGMs may enable improved analysis and prediction of the growth process.

7 Discussion

In the chapters up to this point, four research topics were introduced and discussed. In this chapter, the topics will be discussed from two viewpoints: data acquisition and modeling methods.

7.1 Data acquisition

In Chapters 4, 5, and 6, UAV-RS was used to measure the growth process of soybean canopy area and height. One difficulty in tracking plant growth using UAV-RS is the considerable measurement noise. One solution is calibration using manually measured ground-truth data for a limited number of plants. This calibration was shown to be useful for the soybean canopy height in this dissertation (Chapters 4 and 6) and sorghum plant height (Hu et al., 2018). However, this calibration method cannot be applied to traits that are difficult to measure manually, such as the canopy area. Another solution is the estimation of noise using statistical models. The simplest way is to fit curves to the observed data by minimizing the sum of squares of the residuals, but biased results may be obtained if observed data has a specific pattern of noise. In Chapter 6, the noise of the UAV-RS data was estimated with a statistical model, including the estimation of the size and presence of the noise on each day. Although the appropriate model structures may differ for each dataset, the application of statistical models is considered indispensable to estimate precise growth processes from noisy field RS data.

The usefulness of UAV-RS data was demonstrated in Chapters 4, 5, and 6. In Chapter 4, it was shown that growth process data obtained with UAV-RS could improve GP of soybean biomass. Chapters 5 and 6 showed that UAV-RS enabled modeling the influences of genotypic and environmental factors on soybean growth.

Although UAV-RS has high potential in the analysis and modeling of the growth process, the type of traits measured with UAV-RS are limited compared to the ones measured manually. In Chapter 3, the integrated modeling of rice biomass with CGM and GP was proposed. For the modeling, the precise measurement of the leaf age, the number of tillers, the length of leaves, and the heading date were essential for attaining a high accuracy by the prediction model. Development of methods to measure detailed plant traits using RS-based image data will replace manual measurements with field high-throughput phenotyping.

The rapid growth of deep learning might contribute to the precise measurement of traits with UAV-RS. Deep learning is a machine learning method based on neural network models. Deep learning now became a state-of-the-art image analysis method by estimating numerous parameters with large training data. UAV-RS seems to be compatible with deep learning because UAV-RS can supply many images using small cost and labor. Although deep learning requires the time-consuming annotation of images to use them as training data, several reports proposed efficient methods to prepare training data for plant phenotyping (Chandra et al., 2020; Toda et al.,

2020). The development of machine learning methods will increase the range and accuracy of phenotype data acquired with UAV-RS.

7.2 Modeling methods

In this dissertation, three different modeling methods to extend GP to deal with the growth process data are proposed. The first method is the use of simple growth models introduced in Chapters 4 and 5. Although these growth models are simple, their application has contributed to the improvement of GP of the biomass (Chapter 4) and GP of the growth pattern (Chapter 5). Since these simple growth models have high generalization abilities, it is expected that these methods would apply to a wide range of longitudinal data.

The second method is the integration of CGM and GP introduced in Chapter 3. This method was first proposed by Technow et al., 2015. In that paper, CGM parameters were estimated using only target trait data, without growth process data. Although the concept of the integration of CGM and GP was innovative, its prediction accuracy was not higher than the standard GP when validated with real data (Cooper et al., 2016). The study in Chapter 3 can be interpreted as a successful extension of the proposed model to reflect the growth process data and improve prediction accuracy.

One of the advantages of CGM is that it can be extended to take advantage of previous plant physiological knowledge. For example, the model in Chapter 3 considers three environmental factors, solar radiation, temperature, and day length, based on physiological knowledge. In cases where other environmental factors such as drought or nutrient conditions should be considered, these can be easily incorporated into the model by referring to previous research about each factor's effects. Such extensibility is attractive for creating a versatile environmental response model. However, it is essential to note that models with high complexity and too many parameters may degrade generalization performance.

The third method is the modeling of daily response to environmental factors on plant growth. Although many genotypes' daily growth process is hard to measure manually, the widespread use of high-throughput phenotyping technology has made this new method possible. The method is innovative in that it models G × E through the plant growth process. In addition, it was shown that the proposed method is based on statistical or machine-learning models and requires a sufficient amount of data for both the growth process and environmental data. It was found that a sufficient amount of data was needed to estimate the daily environmental data for each plot when applied to field trials in a few years. Continual improvement of modeling methods is essential to expand the use of growth process data in crop breeding. Several improvements can be made in the methods mentioned in this dissertation. For example, in the integrated model of CGM and GP (Chapter 3) and that of the growth model and GP (Chapters 4 and 5), the estimation of CGM or growth model parameters and their GP were executed separately. However, these two procedures can be unified using hierarchical modeling (Das, Li, Wang, et al., 2011; Ma et al., 2002; Technow et al., 2015; Onogi, Watanabe, et al., 2016). Although integrating estimation and prediction

increases computational cost, it is expected to enable parameter estimation not affected by outliers or local optimum solutions. In the growth model, including daily environmental response (Chapter 6), it is necessary to introduce statistical methods for time series analysis, such as linear dynamic systems, to handle the time-series data structure more accurately. Another interesting topic that could not be touched upon in this study is comparing the growth model's performance (Chapters 4 and 5) with a similar method, random regression (Section 2.2.3).

The most important thing for making the best use of growth process data for breeding is to choose an appropriate model for each case according to the characteristics of the data and the purpose of the analysis. In the models used in this dissertation, the CGM should be used when focusing on the physiological mechanisms of plant growth. A simple growth model has a wide range of applicability, which can be used without much knowledge about the growth process's mechanism. If growth process data does not match existing growth models, random regression can be applied. The statistical model of environmental effects on daily growth proposed in Chapter 6 can be used when there are environmental factors of interest and sufficient data. Finally, in all cases, machine learning models are powerful tools in the pursuit of prediction accuracy.

In this dissertation, the relationship of the plant growth process data, genomic data, and environmental data using the CGM, GP, and growth models was modeled. The studies' main novelty is that the growth processes of several genotypes were measured and analyzed simultaneously. In particular, the growth process was frequently measured in Chapters 5 and 6. In other studies using field experiments, we do not find any data where the growth process was measured as frequently as ones in Chapters 5 and 6. As high-throughput phenotyping systems become more widely used in plant breeding, such data will become almost routinely available. This usage is also true for the availability of crop genomic data with the common use of high-throughput genotyping technologies. The use of time-series growth data in crop breeding is still in its infancy. The findings in this dissertation will serve as a basis for future studies on crop growth processes.

8 Acknowledgments

First of all, I appreciate Prof. Hiroyoshi Iwata, Graduate School of Agricultural and Life Sciences, The University of Tokyo, for the guidance in my studies and in writing this thesis. He taught me for six years in total, and many discussions with him created the foundation of my research background. Also, I appreciate Prof. Hirohisa Kishino and Prof. Hiroshi Ohmori for their guidance and advice, and through them, I acquired knowledge about statistics, machine learning, and quantitative genetics, which became the basis of this thesis.

Prof. Takeshi Izawa, Graduate School of Agricultural and Life Sciences, The University of Tokyo, Dr. Toshihiro Hasegawa, Tohoku Agricultural Research Center, National Agriculture and Food Research Organization (NARO), and Dr. Akito Kaga, Institute of Crop Science, NARO, were sub-chief examiner of this thesis. I appreciate their useful advice. Also, Dr. Toshihiro Hasegawa provided phenotype data of the detailed growth process of the RIL population of rice, used in chapter 3. Dr. Akito Kaga provided seeds of the soybean core collection used in chapters 4, 5, and 6 and gave me advice for field experiment designs and management. This thesis cannot be completed without their professional knowledge about each crop. I appreciate their help.

Dr. Hiromi Kajiya-Kanegae, Research Center for Agricultural Information Technology, NARO, provided marker genotype data of the RIL population of rice used in chapter 3 and whole-genome marker data of soybean used in chapters 4, 5, and 6. Also, she taught me the basis of genetic analysis when she worked as an assistant professor in my laboratory. I appreciate her altruistic works.

For chapter 3, Dr. Hitomi Wakatsuki, Institute for Agro-Environmental Sciences, NARO, and Dr. Kaworu Ebana, Genetic Resources Center, NARO, measured and curated phenotypic data of the RIL population of rice. Mr. Hironori Wakabayashi and Mr. Koji Watanabe managed the field experiments. Mr. Takashi Harigae, Ms. Teruyo Omura, Ms. Miyuki Ishibashi, Ms. Noriko Kimoto, and Ms. Chie Muto assisted the field measurements. Also, Prof. Masanori Yamasaki, Food Resources Education and Research Center, Graduate School of Agricultural Science, Kobe University, provided the multi-environment trials dataset of heading date. Mr. Takuma Yoshioka, a graduated student of the same laboratory, acquired and provided marker data of heading-date-related genes. Mr. Maya Watanabe and Mr. Toru Aoike, graduated students of Graduate School of Agricultural and Life Sciences, The University of Tokyo, curated and gave ideas on the heading date modeling. Dr. Takeshi Hayashi, Institute of Crop Science, NARO, and Dr. Hiroshi Nakagawa, Research Center for Agricultural Information Technology, NARO, gave me advice for the analysis. I appreciate all of them for their kind help.

In chapters 4, 5, and 6, data of the field experiments conducted in Arid Land Research Center, Tottori University were used. I appreciate Prof. Hisashi Tsujimoto for his acceptance and advice for the field experiments. Dr. Yuji Yamasaki, Ms. Izumi Higashida, Ms. Kumi Inagaki, and the technical staff helped me in the field experiments. Especially, Ms. Izumi Higashida measured soil moisture data used in chapter 6. Prof. Masanori Okamoto, Utsunomiya University, Center for

Bioscience Research and Education, also helped me when he was an assistant professor in the center. I appreciate all of them for their kindness.

I received many supports also in UAV-RS. Mr. Tomohiro Hattori and Mr. Shuhei Yamaoka, graduated students of Graduate School of Agricultural and Life Sciences, The University of Tokyo, conducted part of the UAV-RS in 2016 used in chapter 4. Mr. Goshi Sasaki, a graduated student of the same laboratory, and Mr. Chen Tai-Shen, a student of the same laboratory, helped the UAV-RS in 2017 used in chapters 5 and 6. I appreciate their supports.

Many other members supported the soybean field experiments. The main members were Prof. Yoshihiro Ohmori, Prof Hideki Takanashi, and Prof. Toru Fujiwara, Graduate School of Agricultural and Life Sciences, The University of Tokyo, Prof. Hirokazu Takahashi and Prof. Mikio Nakazono, Graduate School of Bioagriculture Science, Nagoya University, Prof. Mai Tsuda, T-PIRC, University of Tsukuba, Dr. Yuji Sawada, Center for Sustainable Resource Science, RIKEN, and many people involved in the experiment. The discussions with them sometimes gave me stimulating opinions. I appreciate all the members involved in the experiment.

Dr. Frederic Baret and Dr. Lopez-Lozano Raul, Mediterranean Environment and Modelling of Agroecosystems, National Institute for Agriculture, Food, and Environment, France, not only gave me critical advice on the growth models used in chapter 5 but also supported my three-month stay in France. I appreciate them and other members in the laboratory for their kind help and inspiring discussions.

Finally, I want to appreciate the members of the laboratory of biometry and bioinformatics. Seminars, reading clubs, and other discussions with the members improved my knowledge. Ms. Sawako Maruyama curated the phenotype data of the soybean field experiment. Ms. Sasaki Mieko supported my office procedures. I do not list their names here, but I appreciate all of the members involved in my studies.

9 Abstract

The growth process of crops is an important research topic in breeding science because traits, such as yield and quality, observed at harvest time are determined in a series of growth processes. By dissecting the formation process of a target trait, we can obtain a deeper understanding of its mechanism, which will improve the efficiency of genetic improvement of the target traits. The growth process, however, has been largely neglected in crop breeding research because obtaining data on the growth process requires time-series observations of plants. Since a large number of genotypes are tested in a breeding program, their time-series measurement is time-consuming and labor-intensive.

Recent developments in sensing technology have made high-throughput phenotyping possible. The application of sensing devices, such as unmanned aerial vehicles (UAVs), can reduce the time and labor required for the measurement of plant traits. It is expected that observation of the growth process of many genotypes will become more accessible for crop breeding. The development and application of appropriate modeling methods are necessary to maximize the benefits derived from such growth process data. In this dissertation, I proposed several modeling methods to evaluate the relationships among crop growth, genomic, and environmental data, and validated whether the models are useful for crop breeding.

1. Predicting biomass of rice with intermediate traits: Modeling method combining crop growth models and genomic prediction models

Genomic prediction (GP) is a method that uses genome-wide marker data and statistical models to predict genotypic values of a target trait. Although GP is a useful method for predicting the phenotype of a target trait, the standard GP model cannot take into account the effect of genotype-by-environment interaction on the trait. In this study, I developed new prediction models of rice biomass that take into account manually measured growth process data.

‘Kinmaze’, ‘Koshihikari’, and their 123 recombinant inbred lines (RILs) were used as plant materials. The manually measured phenotypic data from field experiments conducted in Tsukuba, Japan in 2014–2015 were provided and used in the study. The phenotypic data included time-series data of leaf age (a continuous index of the emergence of ith leaf number), the number of tillers, heading date, leaf length, and biomass. RILs were genotyped with 362 SNPs.

Simple growth models were applied to time-series growth data, and the estimated parameters were used as growth-related traits along with the other traits. Two-step models were developed to predict biomass. In the first step, growth-related traits except heading date were predicted by simple GP. The heading date was predicted with a model that takes into account the effect of the heading-date-related genes and a developmental rate model. In the second step, the predicted growth-related traits were used to predict the biomass using a crop growth model (CGM) or machine learning (ML). In CGM, the biomass was described as the summation of daily growth calculated from growth-related traits and environmental information. In ML, the relationship

between the biomass and the growth-related traits was modeled using linear regression or Random forest. As a result, the CGM based model worked better than the standard GP in both known and unknown environments. It was concluded that the efficient use of growth process data could increase the accuracy and robustness of genomic prediction in yield-related traits that are affected by environmental variations.

2. Genomic prediction modeling of soybean biomass using UAV-based remote sensing and longitudinal model parameters

The application of remote sensing (RS) to crop breeding can provide a wealth of information on plant growth processes in field trials. The inclusion of RS data in multivariate GP (MGP) models has been shown to improve the prediction accuracy of target traits, indicating that traits measured by RS were also beneficial for GP. However, the current MGP model cannot incorporate high-dimensional RS data due to the difficulty in estimating the covariance matrix among variables. In this study, I applied growth models to the time-series RS data and used several parameters to represent the growth pattern, and investigated whether the MGP model with these parameters could improve the prediction accuracy of soybean biomass.

In 2016, 198 genotypes of soybean germplasm were grown in experimental fields in Tottori, Japan. Unmanned aerial vehicle (UAV)-RS was used to measure longitudinal changes in canopy height and area continuously. Growth parameters were estimated by applying simple growth models and incorporated into the GP of biomass. The assessment of genomic heritability and correlation of the parameters indicated that the estimated growth parameters adequately represented the observed growth curves. The incorporation of these growth parameters into the MGP model partially contributed to the accuracy of biomass prediction. It was concluded that these growth models could represent genetic variations in the growth pattern as a function of multiple growth parameters measured by RS. These dimensionally reduced growth models are essential for extracting useful information from RS data and using it for GP and plant breeding.

3. Longitudinal growth analysis of soybean using UAV-based remote sensing and its application on genomic prediction

In this study, I developed models to predict the growth process. With the widespread use of high-throughput phenotyping systems, it is expected that the acquisition of the growth process data will become much easily available. Therefore, by applying GP models to growth data, it is expected to be able to predict the growth of untested genotypes. The growth prediction will be useful for crop breeding because it is essential for cultivation to take into account the growth process. In this study, I implemented several prediction models combining GP and a growth model. The accuracy of the models was validated using growth process data measured with UAV-RS.

In 2017–2019, 198 genotypes of soybean germplasm were grown in experimental fields in Tottori, Japan. The longitudinal changes in their canopy area were measured using UAV-RS. A growth model was applied to the canopy area, with growth expressed as a logistic function and senescence expressed as an exponential function. Next, I developed a two-step GP (TGP) model

and tested whether the growth model contributed to the improvement of growth prediction. In the TGP model, the growth process was predicted by first predicting the parameters of the growth model with GP and then substituting the predicted values for the parameters of the growth model. The prediction accuracy was compared with GP and MGP under three prediction schemes. As a result, TGP showed higher prediction accuracy than the other models in the scheme of future growth prediction, in which the second half of the growth period was predicted from the data of the first half of the growth period. It was concluded that the TGP was useful for future growth prediction using data from the early growth period. This prediction method could be applied to the selection at an early growth stage in crop breeding, and could reduce the cost and time of field trials.

4. Prediction of soybean growth curves by modeling genetic and environmental effects on daily growth

It is well known that genotype-by-environment (G×E) interaction has non-negligible effects on crop traits. Compared to G×E on harvest traits, G×E for growth process has been less discussed due to the difficulty of measuring growth. The analysis of G×E on growth process is essential to clarify and understand the mechanism of G×E. In this study, I developed a statistical model to describe G×E on the growth process measured by UAV-RS.

Phenotypic data of canopy area and height measured using UAV-RS in the same field trials were used. Prior to the modeling of G×E, it is necessary to estimate daily phenotypic values of the canopy area and height. Because the measurement data of UAV-RS were affected by certain noises, I develop a model to distinguish between growth and noise. The model was validated to be proper by comparing estimates of canopy height with the plant height data measured manually. Next, I developed models of the environmental response of daily growth using statistical and machine learning methods. In the machine learning model, environmental data and marker genotype data were included as inputs. On the other hand, in the statistical model, marker genotype data were included as input to account for the similarity among genotypes in the shape of environmental response curves. As a result, the estimated daily response of the soil moisture explained the drought stress on the growth well, although no specific tendency was observed for the other environmental factors. The ability to adequately estimate the effect of soil moisture was supported by the experimental design using the multi-level watering treatment and abundant soil moisture data. By comparing the prediction accuracy, the machine learning model outperformed the statistical model in predicting the growth process, although the possibility of overfitting could not be ignored. The results indicated that attention should be paid to data structure when building statistical and machine learning models for G×E of crop growth.

In this dissertation, I proposed several modeling methods to include the growth process data in the prediction of harvest traits or to analyze the growth process. In these models, the relationships among the plant growth process data, genomic data, and environmental data were modeled using CGM, GP, and growth models. The main novelty of these studies was that the

simultaneous measurement and modeling of the growth processes of a large number of genotypes. As high-throughput phenotyping systems become more widely used for crop breeding, such data will become more readily available. Further improvements in methods for modeling growth process data will be essential to extract useful insights and generate benefits for crop breeding.

10 References

Araus, J., & Cairns, J. E. (2014). Field high-throughput phenotyping: the new crop breeding frontier. *Trends in Plant Science*, *19*(1), 52–61. https://doi.org/10.1016/j.tplants.2013.09.008

Asoro, Franco G., Newell, Mark A., Beavis, William D., Scott, M. Paul, Tinker, Nicholas A., & Jannink, Jean-Luc. (2013). Genomic, Marker-Assisted, and Pedigree-BLUP Selection Methods for β-Glucan Concentration in Elite Oat. *Crop Science*, *53*(5), 1894–1906. https://doi.org/10.2135/cropsci2012.09.0526

Bendig, J., Yu, K., Aasen, H., Bolten, A., Bennertz, S., Broscheit, J., Gnyp, M. L., & Bareth, G. (2015). Combining UAV-based plant height from crop surface models, visible, and near infrared vegetation indices for biomass monitoring in barley. *International Journal of Applied Earth Observation and Geoinformation*, *39*, 79–87. https://doi.org/10.1016/j.jag.2015.02.012

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag.

Blancon, J., Dutartre, D., Tixier, M.-H., Weiss, M., Comar, A., Praud, S., & Baret, F. (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. https://doi.org/10.3389/fpls.2019.00685

Bouman, B. A. M., Keulen, H. van, Laar, H. H. van, & Rabbinge, R. (1996). The 'School of de Wit' crop growth simulation models: A pedigree and historical overview. *Agricultural Systems*, *52*(2–3), 171–198. https://doi.org/10.1016/0308-521X(96)00011-X

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Broman, K. W., Wu, H., Sen, Ś., & Churchill, G. A. (2003). R/qtl: QTL mapping in experimental crosses. *Bioinformatics*, *19*(7), 889–890. https://doi.org/10.1093/bioinformatics/btg112

Burgueño, J., Campos, G. de los, Weigel, K., & Crossa, J. (2012). Genomic Prediction of Breeding Values when Modeling Genotype × Environment Interaction using Pedigree and Dense Molecular Markers. *Crop Science*, *52*(2), 707. https://doi.org/10.2135/cropsci2011.06.0299

Cabrera-Bosquet, L., Crossa, J., Zitzewitz, J., Serret, M., & Araus, J. (2012). High-throughput Phenotyping and Genomic Selection: The Frontiers of Crop Breeding Converge. *Journal of Integrative Plant Biology*, *54*(5). https://doi.org/10.1111/j.1744-7909.2012.01116.x

Calus, M. P., & Veerkamp, R. F. (2011). Accuracy of multi-trait genomic selection using different methods. *Genetics Selection Evolution*, *43*(1), 26. https://doi.org/10.1186/1297-9686-43-26

Campbell, M., Walia, H., & Morota, G. (2018). Utilizing random regression models for genomic prediction of a longitudinal trait derived from high-throughput phenotyping. *Plant Direct*, *2*(9), e00080. https://doi.org/10.1002/pld3.80

Chandra, A. L., Desai, S. V., Balasubramanian, V. N., Ninomiya, S., & Guo, W. (2020). Active learning with point supervision for cost-effective panicle detection in cereal crops. *Plant Methods*, *16*(1), 34. https://doi.org/10.1186/s13007-020-00575-8

Cooper, M., Technow, F., Messina, C., Gho, C., & Totir, L. R. (2016). Use of Crop Growth Models with Whole-Genome Prediction: Application to a Maize Multienvironment Trial. *Crop Science*, *56*(5), 2141. https://doi.org/10.2135/cropsci2015.08.0512

Córcoles, J. I., Ortega, J. F., Hernández, D., & Moreno, M. A. (2013). Estimation of leaf area index in onion (Allium cepa L.) using an unmanned aerial vehicle. *Biosystems Engineering*, *115*(1), 31–42. https://doi.org/10.1016/j.biosystemseng.2013.02.002

Crain, J., Mondal, S., Rutkoski, J., Singh, R. P., & Poland, J. (2018). Combining High-Throughput Phenotyping and Genomic Information to Increase Prediction and Selection Accuracy in Wheat Breeding. *The Plant Genome*, *11*(1). https://doi.org/10.3835/plantgenome2017.05.0043

Crispim, A. C., Kelly, M. J., Guimarães, S. E. F., Silva, F. F. e, Fortes, M. R. S., Wenceslau, R. R., & Moore, S. (2015). Multi-Trait GWAS and New Candidate Genes Annotation for Growth Curve Parameters in Brahman Cattle. *PLOS ONE*, *10*(10), e0139906. https://doi.org/10.1371/journal.pone.0139906

Daetwyler, H. D., Villanueva, B., & Woolliams, J. A. (2008). Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PloS One*, *3*(10), e3395. https://doi.org/10.1371/journal.pone.0003395

Das, K., Li, J., Fu, G., Wang, Z., & Wu, R. (2011). Genome-Wide Association Studies for Bivariate Sparse Longitudinal Data. *Human Heredity*, *72*(2), 110–120. https://doi.org/10.1159/000330781

Das, K., Li, J., Wang, Z., Tong, C., Fu, G., Li, Y., Xu, M., Ahn, K., Mauger, D., Li, R., & Wu, R. (2011). A dynamic model for genome-wide association studies. *Human Genetics*, *129*(6), 629–639. https://doi.org/10.1007/s00439-011-0960-6

Duan, S.-B., Li, Z.-L., Wu, H., Tang, B.-H., Ma, L., Zhao, E., & Li, C. (2014). Inversion of the PROSAIL model to estimate leaf area index of maize, potato, and sunflower fields from unmanned aerial vehicle hyperspectral data. *International Journal of Applied Earth Observation and Geoinformation*, *26*, 12–20. https://doi.org/10.1016/j.jag.2013.05.007

Duan, T., Chapman, S. C., Guo, Y., & Zheng, B. (2017). Dynamic monitoring of NDVI in wheat agronomy and breeding trials using an unmanned aerial vehicle. *Field Crops Research*, *210*(Remote Sens. Environ. 11 1981), 71–80. https://doi.org/10.1016/j.fcr.2017.05.025

Eberhart, R., & Kennedy, J. (1995). A new optimizer using particle swarm theory. *MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science*, 39–43. https://doi.org/10.1109/MHS.1995.494215

Endelman, J. B. (2011). Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome Journal*, *4*(3), 250. https://doi.org/10.3835/plantgenome2011.08.0024

Endelman, J. B., & Jannink, J.-L. (2012). Shrinkage Estimation of the Realized Relationship Matrix. *G3: Genes|Genomes|Genetics*, *2*(11), 1405–1413. https://doi.org/10.1534/g3.112.004259

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, *33*(1), 1–22.

Furbank, R. T., & Tester, M. (2011). Phenomics – technologies to relieve the phenotyping bottleneck. *Trends in Plant Science*, *16*(12), 635–644. https://doi.org/10.1016/j.tplants.2011.09.005

García-Ruiz, A., Cole, J. B., VanRaden, P. M., Wiggans, G. R., Ruiz-López, F. J., & Tassell, C. P. V. (2016). Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. *Proceedings of the National Academy of Sciences*, *113*(28), E3995–E4004. https://doi.org/10.1073/pnas.1519061113

Gianola, D. (2013). Priors in Whole-Genome Regression: The Bayesian Alphabet Returns. *Genetics*, *194*(3), 573–596. https://doi.org/10.1534/genetics.113.151753

Golzarian, M. R., Frick, R. A., Rajendran, K., Berger, B., Roy, S., Tester, M., & Lun, D. S. (2011). Accurate inference of shoot biomass from high-throughput images of cereal plants. *Plant Methods*, *7*(1), 2. https://doi.org/10.1186/1746-4811-7-2

González-Recio, O., Rosa, G. J. M., & Gianola, D. (2014). Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. *Livestock Science*, *166*(Mach. Learn. 24 1996), 217–231. https://doi.org/10.1016/j.livsci.2014.05.036

Grenier, C., Cao, T.-V., Ospina, Y., Quintero, C., Châtel, M. H., Tohme, J., Courtois, B., & Ahmadi, N. (2015). Accuracy of Genomic Selection in a Rice Synthetic Population Developed for Recurrent Selection Breeding. *PLOS ONE*, *10*(8), e0136594. https://doi.org/10.1371/journal.pone.0136594

Habier, D., Fernando, R. L., Kizilkaya, K., & Garrick, D. J. (2011). Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics*, *12*(1), 186. https://doi.org/10.1186/1471-2105-12-186

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning, Data Mining, Inference, and Prediction* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Heffner, E. L., Sorrells, M. E., & Jannink, J.-L. (2009). Genomic Selection for Crop Improvement. *Crop Science*, *49*(1), 1. https://doi.org/10.2135/cropsci2008.08.0512

Heslot, N., Akdemir, D., Sorrells, M. E., & Jannink, J.-L. (2014). Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. *Theoretical and Applied Genetics*, *127*(2), 463–480. https://doi.org/10.1007/s00122-013-2231-5

Holzworth, D. P., Huth, N. I., deVoil, P. G., Zurcher, E. J., Herrmann, N. I., McLean, G., Chenu, K., Oosterom, E. J. van, Snow, V., Murphy, C., Moore, A. D., Brown, H., Whish, J. P. M., Verrall, S., Fainges, J., Bell, L. W., Peake, A. S., Poulton, P. L., Hochman, Z., … Keating, B. A. (2014). APSIM – Evolution towards a new generation of agricultural systems simulation. *Environmental Modelling & Software*, *62*, 327–350. https://doi.org/10.1016/j.envsoft.2014.07.009

Hori, K., Ogiso-Tanaka, E., Matsubara, K., Yamanouchi, U., Ebana, K., & Yano, M. (2013). Hd16, a gene for casein kinase I, is involved in the control of rice flowering time by modulating the day-length response. *The Plant Journal*, *76*(1), 36–46. https://doi.org/10.1111/tpj.12268

Horie, T. (1987). A model for evaluating climatic productivity and water balance of irrigated rice and its application to Southeast Asia. *Southeast Asian Studies*, *25*, 762–774. https://doi.org/10.20495/tak.25.1_62

Hu, P., Chapman, S. C., Wang, X., Potgieter, A., Duan, T., Jordan, D., Guo, Y., & Zheng, B. (2018). Estimation of plant height using a high throughput phenotyping platform based on

unmanned aerial vehicle and self-calibration: Example for sorghum breeding. *European Journal of Agronomy*, *95*, 24–32. https://doi.org/10.1016/j.eja.2018.02.004

Huang, G.-B., Zhu, Q.-Y., & Siew, C.-K. (2006). Extreme learning machine: Theory and applications. *Neurocomputing*, *70*(1–3), 489–501. https://doi.org/10.1016/j.neucom.2005.12.126

Iizumi, T., Yokozawa, M., & Nishimori, M. (2009). Parameter estimation and uncertainty analysis of a large-scale crop model for paddy rice: Application of a Bayesian approach. *Agricultural and Forest Meteorology*, *149*(2), 333–348. https://doi.org/10.1016/j.agrformet.2008.08.015

Jahn, C. E., Mckay, J. K., Mauleon, R., Stephens, J., McNally, K. L., Bush, D. R., Leung, H., & Leach, J. E. (2010). Genetic variation in biomass traits among 20 diverse rice varieties. *Plant Physiology*, *155*(1), 157–168. https://doi.org/10.1104/pp.110.165654

Jamrozik, J., & Schaeffer, L. R. (1997). Estimates of Genetic Parameters for a Test Day Model with Random Regressions for Yield Traits of First Lactation Holsteins. *Journal of Dairy Science*, *80*(4), 762–770. https://doi.org/10.3168/jds.s0022-0302(97)75996-4

Jarquín, D., Crossa, J., Lacaze, X., Cheyron, P. D., Daucourt, J., Lorgeou, J., Piraux, F., Guerreiro, L., Pérez, P., Calus, M., Burgueño, J., & Campos, G. de los. (2014). A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theoretical and Applied Genetics*, *127*(3), 595–607. https://doi.org/10.1007/s00122-013-2243-1

Jarquin, D., Howard, R., Xavier, A., & Choudhury, S. D. (2018). Increasing Predictive Ability by Modeling Interactions between Environments, Genotype and Canopy Coverage Image Data for Soybeans. *Agronomy*, *8*(4), 51. https://doi.org/10.3390/agronomy8040051

Jia, Y., & Jannink, J.-L. (2012). Multiple-Trait Genomic Selection Methods Increase Genetic Value Prediction Accuracy. *Genetics*, *192*(4), 1513–1522. https://doi.org/10.1534/genetics.112.144246

Jin, X., Kumar, L., Li, Z., Feng, H., Xu, X., Yang, G., & Wang, J. (2018). A review of data assimilation of remote sensing and crop models. *European Journal of Agronomy*, *92*, 141–152. https://doi.org/10.1016/j.eja.2017.11.002

Kajiya-Kanegae, H., Nagasaki, H., Kaga, A., Hirano, K., Ogiso-Tanaka, E., Matsuoka, M., Ishimori, M., Ishimoto, M., Hashiguchi, M., Tanaka, H., Akashi, R., Isobe, S., & Iwata, H. Whole-genome sequence diversity and association analysis of 198 soybean accessions in mini-core collections. Manuscript submitted for publication.

Kang, H., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., & Eskin, E. (2008). Efficient Control of Population Structure in Model Organism Association Mapping. *Genetics*, *178*(3), 1709–1723. https://doi.org/10.1534/genetics.107.080101

Kang, M. S. (2001). Genotype-environment interaction: progress and prospect. In M. S. Kang (Ed.), *Quantitative Genetics, Genomics and Plant Breeding*. CABI Publishing.

Kasampalis, D. A., Alexandridis, T. K., Deva, C., Challinor, A., Moshou, D., & Zalidis, G. (2018). Contribution of Remote Sensing on Crop Models: A Review. *Journal of Imaging*, *4*(4), 52. https://doi.org/10.3390/jimaging4040052

Khush, G. S. (2013). Strategies for increasing the yield potential of cereals: case of rice as an example. *Plant Breeding*, *132*(5), 433–436. https://doi.org/10.1111/pbr.1991

Kirkpatrick, M, Lofsvold, D., & Bulmer, M. (1990). Analysis of the inheritance, selection and evolution of growth trajectories. *Genetics*, *124*(4), 979–993.

Kirkpatrick, Mark, & Heckman, N. (1989). A quantitative genetic model for growth, shape, reaction norms, and other infinite-dimensional characters. *Journal of Mathematical Biology*, *27*(4), 429–450. https://doi.org/10.1007/BF00290638

Koetz, B., Baret, F., Poilvé, H., & Hill, J. (2005). Use of coupled canopy structure dynamic and radiative transfer models to estimate biophysical canopy characteristics. *Remote Sensing of Environment*, *95*(1), 115–124. https://doi.org/10.1016/j.rse.2004.11.017

Kojima, S., Takahashi, Y., Kobayashi, Y., Monna, L., Sasaki, T., Araki, T., & Yano, M. (2002). Hd3a, a Rice Ortholog of the Arabidopsis FT Gene, Promotes Transition to Flowering Downstream of Hd1 under Short-Day Conditions. *Plant and Cell Physiology*, *43*(10), 1096–1105. https://doi.org/10.1093/pcp/pcf156

Liaw, A., & Wiener, M. (2002). Classification and Regression by RandomForest. *R News*, *2*(3), 18–22.

Ma, C.-X., Casella, G., & Wu, R. (2002). Functional mapping of quantitative trait loci underlying the character process: a theoretical framework. *Genetics*, *161*(4), 1751–1762.

Matsubara, K., Ogiso-Tanaka, E., Hori, K., Ebana, K., Ando, T., & Yano, M. (2012). Natural Variation in Hd17, a Homolog of Arabidopsis ELF3 That is Involved in Rice Photoperiodic Flowering. *Plant and Cell Physiology*, *53*(4), 709–716. https://doi.org/10.1093/pcp/pcs028

Meuwissen, T. H. E. (2009). Accuracy of breeding values of "unrelated" individuals predicted by dense SNP genotyping. *Genetics, Selection, Evolution : GSE*, *41*(1), 35. https://doi.org/10.1186/1297-9686-41-35

Meuwissen, T. H., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819–1829.

Meyer, K. (2002, August 19). *RRGIBBS - A program for simple random regression analyses via Gibbs sampling*. 7th World Congress on Genetics Applied to Livestock Production.

Meyer, K. (2007). WOMBAT—A tool for mixed model analyses in quantitative genetics by restricted maximum likelihood (REML). *Journal of Zhejiang University SCIENCE B*, *8*(11), 815–821. https://doi.org/10.1631/jzus.2007.b0815

Meyer, K., & Hill, W. G. (1997). Estimation of genetic and phenotypic covariance functions for longitudinal or 'repeated' records by restricted maximum likelihood. *Livestock Production Science*, *47*(3), 185–200. https://doi.org/10.1016/S0301-6226(96)01414-5

Montes, J. M., Technow, F., Dhillon, B. S., Mauch, F., & Melchinger, A. E. (2011). High-throughput non-destructive biomass determination during early plant development in maize under field conditions. *Field Crops Research*, *121*(2), 268–273. https://doi.org/10.1016/j.fcr.2010.12.017

Moser, G., Tier, B., Crump, R. E., Khatkar, M. S., & Raadsma, H. W. (2009). A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. *Genetics Selection Evolution*, *41*(1), 56. https://doi.org/10.1186/1297-9686-41-56

Nagasaki, H., Ebana, K., Shibaya, T., Yonemaru, J., & Yano, M. (2010). Core single-nucleotide polymorphisms—a tool for genetic analysis of the Japanese rice population. *Breeding Science*, *60*(5), 648–655. https://doi.org/10.1270/jsbbs.60.648

Nakagawa, H., Yamagishi, J., Miyamoto, N., Motoyama, M., Yano, M., & Nemoto, K. (2005). Flowering response of rice to photoperiod and temperature: a QTL analysis using a phenological model. *Theoretical and Applied Genetics*, *110*(4), 778–786. https://doi.org/10.1007/s00122-004-1905-4

Nelder, J. A. (1961). The Fitting of a Generalization of the Logistic Curve. *Biometrics*, *17*(1), 89. https://doi.org/10.2307/2527498

Okada, S., Suehiro, M., Ebana, K., Hori, K., Onogi, A., Iwata, H., & Yamasaki, M. (2017). Genetic dissection of grain traits in Yamadanishiki, an excellent sake-brewing rice cultivar. *Theoretical and Applied Genetics*, *130*(12), 2567–2585. https://doi.org/10.1007/s00122-017-2977-2

Onogi, A., Ideta, O., Inoshita, Y., Ebana, K., Yoshioka, T., Yamasaki, M., & Iwata, H. (2015). Exploring the areas of applicability of whole-genome prediction methods for Asian rice (Oryza sativa L.). *Theoretical and Applied Genetics*, *128*(1), 41–53. https://doi.org/10.1007/s00122-014-2411-y

Onogi, A., Ideta, O., Yoshioka, T., Ebana, K., Yamasaki, M., & Iwata, H. (2016). Uncovering a Nuisance Influence of a Phenological Trait of Plants Using a Nonlinear Structural Equation: Application to Days to Heading and Culm Length in Asian Cultivated Rice (Oryza Sativa L.). *PLOS ONE*, *11*(2), e0148609. https://doi.org/10.1371/journal.pone.0148609

Onogi, A., Ogino, A., Sato, A., Kurogi, K., Yasumori, T., & Togashi, K. (2019). Development of a structural growth curve model that considers the causal effect of initial phenotypes. *Genetics Selection Evolution*, *51*(1), 19. https://doi.org/10.1186/s12711-019-0461-y

Onogi, A., Watanabe, M., Mochizuki, T., Hayashi, T., Nakagawa, H., Hasegawa, T., & Iwata, H. (2016). Toward integration of genomic selection with crop modelling: the development of an integrated approach to predicting rice heading dates. *Theoretical and Applied Genetics*, *129*(4), 805–817. https://doi.org/10.1007/s00122-016-2667-5

Oraby, H., Venkatesh, B., Dale, B., Ahmad, R., Ransom, C., Oehmke, J., & Sticklen, M. (2007). Enhanced conversion of plant biomass into glucose using transgenic rice-produced endoglucanase for cellulosic ethanol. *Transgenic Research*, *16*(6), 739–749. https://doi.org/10.1007/s11248-006-9064-9

Pierre, C. S., Burgueño, J., Crossa, J., Dávila, G. F., López, P. F., Moya, E. S., Moreno, J. I., Muela, V. M. H., Villa, V. M. Z., Vikram, P., Mathews, K., Sansaloni, C., Sehgal, D., Jarquin, D., Wenzl, P., & Singh, S. (2016). Genomic prediction models for grain yield of spring bread wheat in diverse agro-ecological zones. *Scientific Reports*, *6*(1), srep27312. https://doi.org/10.1038/srep27312

Piles, M., Gianola, D., Varona, L., & Blasco, A. (2003). Bayesian inference about parameters of a longitudinal trajectory when selection operates on a correlated trait. *Journal of Animal Science*, *81*(11), 2714–2724. https://doi.org/10.2527/2003.81112714x

Pinnschmidt, H. O., Batchelor, W. D., & Teng, P. S. (1995). Simulation of multiple species pest damage in rice using CERES-rice. *Agricultural Systems*, *48*(2), 193–222. https://doi.org/10.1016/0308-521X(94)00012-G

Pletcher, S. D., & Geyer, C. J. (1999). The genetic analysis of age-dependent traits: modeling the character process. *Genetics*, *153*(2), 825–835.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Ramirez-Villegas, J., Watson, J., & Challinor, A. J. (2015). Identifying traits for genotypic adaptation using crop models. *Journal of Experimental Botany*, *66*(12), 3451–3462. https://doi.org/10.1093/jxb/erv014

Richards, F. J. (1959). A Flexible Growth Function for Empirical Use. *Journal of Experimental Botany*, *10*(2), 290–301. https://doi.org/10.1093/jxb/10.2.290

Ritchie, Alocilja, Singh, & Uehara, &. (1987). IBSNAT and the CERES-Rice model. *Agrotechnology Transfer*, *3*, 1–5.

Rutkoski, J, Singh, R. P., Huerta-Espino, J., Bhavani, S., Poland, J., Jannink, J. L., & Sorrells, M. E. (2015). Genetic Gain from Phenotypic and Genomic Selection for Quantitative Resistance to Stem Rust of Wheat. *The Plant Genome*, *8*(2), 0. https://doi.org/10.3835/plantgenome2014.10.0074

Rutkoski, Jessica, Poland, J., Mondal, S., Autrique, E., Pérez, L., Crossa, J., Reynolds, M., & Singh, R. (2016). Canopy Temperature and Vegetation Indices from High-Throughput Phenotyping Improve Accuracy of Pedigree and Genomic Selection for Grain Yield in Wheat. *G3: Genes|Genomes|Genetics*, *6*(9), 2799–2808. https://doi.org/10.1534/g3.116.032888

Schulz-Streeck, T., Ogutu, J. O., Gordillo, A., Karaman, Z., Knaak, C., & Piepho, H. (2013). Genomic selection allowing for marker-by-environment interaction. *Plant Breeding*, *132*(6), 532–538. https://doi.org/10.1111/pbr.12105

Soltani, A., & Sinclair, T. (2012). *Modeling Physiology of Crop Development, Growth and Yield*. CABI.

Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., Redoña, E., Atlin, G., Jannink, J.-L., & McCouch, S. R. (2015). Genomic Selection and Association Mapping in Rice (Oryza sativa): Effect of Trait Genetic Architecture, Training Population Composition, Marker Number and Statistical Model on Accuracy of Rice Genomic Selection in Elite, Tropical Rice Breeding Lines. *PLOS Genetics*, *11*(2), e1004982. https://doi.org/10.1371/journal.pgen.1004982

Spindel, J. E., Begum, H., Akdemir, D., Collard, B., Redoña, E., Jannink, J.-L., & McCouch, S. (2016). Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. *Heredity*, *116*(4), 395. https://doi.org/10.1038/hdy.2015.113

Sun, J., Rutkoski, J. E., Poland, J. A., Crossa, J., Jannink, J.-L., & 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). https://doi.org/10.3835/plantgenome2016.11.0111

Takahashi, Y., Shomura, A., Sasaki, T., & Yano, M. (2001). Hd6, a rice quantitative trait locus involved in photoperiod sensitivity, encodes the α subunit of protein kinase CK2. *Proceedings of the National Academy of Sciences*, *98*(14), 7922–7927. https://doi.org/10.1073/pnas.111136798

Tattaris, M., Reynolds, M. P., & Chapman, S. C. (2016). A Direct Comparison of Remote Sensing Approaches for High-Throughput Phenotyping in Plant Breeding. *Frontiers in Plant Science*, *7*, 1131. https://doi.org/10.3389/fpls.2016.01131

Technow, F., Messina, C. D., Totir, L. R., & Cooper, M. (2015). Integrating Crop Growth Models with Whole Genome Prediction through Approximate Bayesian Computation. *PLOS ONE*, *10*(6), e0130855. https://doi.org/10.1371/journal.pone.0130855

Timsina, J., & Humphreys, E. (2006). Performance of CERES-Rice and CERES-Wheat models in rice–wheat systems: A review. *Agricultural Systems*, *90*(1–3), 5–31. https://doi.org/10.1016/j.agsy.2005.11.007

Toda, Y., Okura, F., Ito, J., Okada, S., Kinoshita, T., Tsuji, H., & Saisho, D. (2020). Training instance segmentation neural network with synthetic datasets for crop seed phenotyping. *Communications Biology*, *3*(1), 173. https://doi.org/10.1038/s42003-020-0905-5

Wang, X., Li, L., Yang, Z., Zheng, X., Yu, S., Xu, C., & Hu, Z. (2016). Predicting rice hybrid performance using univariate and multivariate GBLUP models based on North Carolina mating design II. *Heredity*, *118*(3), 302–310. https://doi.org/10.1038/hdy.2016.87

Watanabe, K., Guo, W., Arai, K., Takanashi, H., Kajiya-Kanegae, H., Kobayashi, M., Yano, K., Tokunaga, T., Fujiwara, T., Tsutsumi, N., & Iwata, H. (2017). High-Throughput Phenotyping of Sorghum Plant Height Using an Unmanned Aerial Vehicle and Its Application to Genomic Prediction Modeling. *Frontiers in Plant Science*, *8*, 421. https://doi.org/10.3389/fpls.2017.00421

White, J. W., Andrade-Sanchez, P., Gore, M. A., Bronson, K. F., Coffelt, T. A., Conley, M. M., Feldmann, K. A., French, A. N., Heun, J. T., Hunsaker, D. J., Jenks, M. A., Kimball, B. A., Roth, R. L., Strand, R. J., Thorp, K. R., Wall, G. W., & Wang, G. (2012). Field-based phenomics for plant genetics research. *Field Crops Research*, *133*, 101–112. https://doi.org/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*, *18*(1), 1–8. https://doi.org/10.1073/pnas.18.1.1

Wit, C. T. de. (1965). Photosynthesis of leaf canopies. *Pudoc*, *663*.

Wu, W., Zhou, Y., Li, W., Mao, D., & Chen, Q. (2002). Mapping of quantitative trait loci based on growth models. *TAG Theoretical and Applied Genetics*, *105*(6–7), 1043–1049. https://doi.org/10.1007/s00122-002-1052-8

Xavier, A., Muir, W. M., Craig, B., & Rainey, K. M. (2016). Walking through the statistical black boxes of plant breeding. *Theoretical and Applied Genetics*, *129*(10), 1933–1949. https://doi.org/10.1007/s00122-016-2750-y

Xu, S., Zhu, D., & Zhang, Q. (2014). Predicting hybrid performance in rice using genomic best linear unbiased prediction. *Proceedings of the National Academy of Sciences*, *111*(34), 12456–12461. https://doi.org/10.1073/pnas.1413750111

Xue, W., Xing, Y., Weng, X., Zhao, Y., Tang, W., Wang, L., Zhou, H., Yu, S., Xu, C., Li, X., & Zhang, Q. (2008). Natural variation in Ghd7 is an important regulator of heading date and yield potential in rice. *Nature Genetics*, *40*(6), 761–767. https://doi.org/10.1038/ng.143

Yabe, S., Ohsawa, R., & Iwata, H. (2014). Genomic Selection for the Traits Expressed after Pollination in Allogamous Plants. *Crop Science*, *54*(4), 1448. https://doi.org/10.2135/cropsci2013.05.0319

Yamamoto, T., Nagasaki, H., Yonemaru, J., Ebana, K., Nakajima, M., Shibaya, T., & Yano, M. (2010). Fine definition of the pedigree haplotypes of closely related rice cultivars by means of genome-wide discovery of single-nucleotide polymorphisms. *BMC Genomics*, *11*(1), 1–14. https://doi.org/10.1186/1471-2164-11-267

Yang, G., Liu, J., Zhao, C., Li, Z., Huang, Y., Yu, H., Xu, B., Yang, X., Zhu, D., Zhang, X., Zhang, R., Feng, H., Zhao, X., Li, Z., Li, H., & Yang, H. (2017). Unmanned Aerial Vehicle Remote Sensing for Field-Based Crop Phenotyping: Current Status and Perspectives. *Frontiers in Plant Science*, *8*, 1111. https://doi.org/10.3389/fpls.2017.01111

Yano, M., Katayose, Y., Ashikari, M., Yamanouchi, U., Monna, L., Fuse, T., Baba, T., Yamamoto, K., Umehara, Y., Nagamura, Y., & Sasaki, T. (2000). Hd1, a major photoperiod sensitivity quantitative trait locus in rice, is closely related to the Arabidopsis flowering time gene CONSTANS. *The Plant Cell*, *12*(12), 2473–2484.

Yin, X., Kropff, M. J., Horie, T., Nakagawa, H., Centeno, H. G. S., Zhu, D., & Goudriaan, J. (1997). A model for photothermal responses of flowering in rice I. Model description and parameterization. *Field Crops Research*, *51*(3), 189–200. https://doi.org/10.1016/s0378-4290(96)03456-9

Yin, X., Kropff, M. J., & Stam, P. (1999). The role of ecophysiological models in QTL analysis: the example of specific leaf area in barley. *Heredity*, *82*(4), 6885030. https://doi.org/10.1038/sj.hdy.6885030

Yin, X., Struik, P. C., & Kropff, M. J. (2004). Role of crop physiology in predicting gene-to-phenotype relationships. *Trends in Plant Science*, *9*(9), 426–432. https://doi.org/10.1016/j.tplants.2004.07.007

Zhang, Z.-H., Li, P., Wang, L.-X., Hu, Z.-L., Zhu, L.-H., & Zhu, Y.-G. (2004). Genetic dissection of the relationships of biomass production and partitioning with yield and yield related traits in rice. *Plant Science*, *167*(1), 1–8. https://doi.org/10.1016/j.plantsci.2004.01.007

Zhou, Y., Li, W., Wu, W., Chen, Q., Mao, D., & Worland, A. J. (2001). Genetic dissection of heading time and its components in rice. *TAG Theoretical and Applied Genetics*, *102*(8), 1236–1242. https://doi.org/10.1007/s001220100539