Phenomics data processing: A plot-level model for repeated measurements to extract the timing of key stages and quantities at defined time points

Lukas Roth[a,*], María Xosé Rodríguez-Álvarez[b,c], Fred van Eeuwijk[d], Hans-Peter Piepho[e], Andreas Hund[a]

[a]*ETH Zurich, Institute of Agricultural Sciences, Universitätstrasse 2, 8092 Zurich, Switzerland*
[b]*BCAM – Basque Center for Applied Mathematics, Alameda de Mazarredo, 14. E-48009 Bilbao, Basque Country, Spain*
[c]*IKERBASQUE, Basque Foundation for Science, Bilbao, Spain*
[d]*Wageningen University and Research, Biometris, P.O. Box 16, 6700 AA Wageningen, The Netherlands*
[e]*University of Hohenheim, Institute for Crop Science, Biostatistics Unit, Fruwirthstrasse 23, 70593 Stuttgart, Germany*

Abstract

Decision-making in breeding increasingly depends on the ability to capture and predict crop responses to changing environmental factors. Advances in crop modeling as well as high-throughput field phenotyping (HTFP) hold promise to provide such insights. Processing HTFP data is an interdisciplinary task that requires broad knowledge on experimental design, measurement techniques, feature extraction, dynamic trait modeling, and prediction of genotypic values using statistical models. To get an overview of sources of variations in HTFP, we develop a general plot-level model for repeated measurements. Based on this model, we propose a seamless stage-wise process that allows to carry on estimated means and variances from stage to stage and approximates the gold standard of a single-stage analysis. The process builds on the extraction of three intermediate trait categories; (1) timing of key stages, (2) quantities at defined time points or periods, and (3) dose-response curves. In a first stage, these intermediate traits are extracted from low-level traits' time series (e.g., canopy height) using P-splines and the quarter of maximum elongation rate method (QMER), as well as final height percentiles. In a second and third stage, extracted traits are further processed using a stage-wise linear mixed model analysis. Using a wheat canopy growth simulation to generate canopy height time series, we demonstrate the suitability of the stage-wise process for traits of the first two above-mentioned categories. Results indicate that, for the first stage, the P-spline/QMER method was more robust than the percentile method. In the subsequent two-stage linear mixed model processing, weighting the second and third stage with error variance estimates from the previous stages improved the root mean squared error. We conclude that processing phenomics data in stages represents a feasible approach if using appropriate weighting through all stages. P-splines in combination with the QMER method are suitable tools to extract timing of key stages and quantities at defined time points from HTFP data.

Keywords: High-throughput field phenotyping, dynamic modeling, stage-wise processing, canopy height

*Corresponding author: lukas.roth@usys.ethz.ch

Highlights

- General plot-level model for repeated high-throughput field phenotyping measurements
- Three main intermediate trait categories for dynamic modeling
- Seamless stage-wise process that allows to carry on estimated means and variances
- Phenomics data processing cheatsheet

1. Introduction

Advances in high-throughput field phenotyping (HTFP) allow capturing large data sets with high temporal and spatial resolution (Rebetzke et al., 2019). Summarizing these spatio-temporal data in a meaningful way is essential to support selection and decision-making in breeding. In HTFP the primary data often consists of images, point measurements, orthophotos, or point clouds from which low-level traits (e.g., shoot counts, canopy cover, canopy height, or senescence) are extracted. After feature extraction, these low-level traits may be tracked over time in a subsequent temporal modelling step (van Eeuwijk et al., 2019; Moreira et al., 2020). If monitored across the lifetime of a plant, low-level traits often follow some sort of monotonically increasing function (e.g., canopy height or senescence) or concave functions (e.g., number of growing shoots or canopy cover), which allows estimating a dynamics' characteristics, referred to as intermediate traits.

Estimating such intermediate traits from spatio-temporal measurements implies *a priori* knowledge of growth processes, best summarized in crop growth models. The performance of these crop growth models can only advance if they become validated with field-based data (Ramirez-Villegas et al., 2015). Crop models have rapidly gained in complexity over time, culminating in the description of plants by 3-D functional-structural models (Vos et al., 2010). Indoor platforms have proven useful to characterize the dynamics of such models (Tardieu et al., 2017), but discrepancies between field and indoor experiments raised doubts if results are always directly transferable (Poorter et al., 2016). Field-based phenotyping may help to bridge this gap (Araus et al., 2018).

While under controlled conditions environmental factors may be adequately controlled, the lack of control over meteorological conditions poses a major challenge for field phenotyping. Several additional types of errors need to be considered, which can be classified into those directly affecting the sensor reading, and those affecting the plant development.

In HTFP there are attempts to quantify genotype-specific timing of phenology stages (Hurtado et al., 2012) and response patterns to distinct environmental variables like temperature (Grieder et al., 2015; Kronenberg et al., 2020a). A comparable approach in genomics uses functional mapping of quantitative trait loci (QTLs), e.g., based on logistic growth curves (Ma et al., 2002; Malosetti et al., 2006). Ma et al. proposed to distinguish three biological processes in such models: allometric laws, growth models, and reaction norms. Characterizing crop model dynamics using field data becomes increasingly difficult as models become more complex. A solution

is to model the dynamic process of growth based on traits or scores that lack a clear physiological interpretation. In phenomics, this was demonstrated using serial measurements as predictors for statistical learning (e.g. Ubbens et al., 2020; Maimaitijiang et al., 2020; Herrero-Huerta et al., 2020). In genomics, comparable approaches are based on functional principal component analysis, where curves are specified as linear combinations of basis functions, and the corresponding scores then used as intermediate traits (Kwak et al., 2016; Moreira et al., 2020).

From a plant physiology point of view, such approaches represent a 'black box': Drawing conclusions on the biological importance of the underlying traits is rather difficult. Therefore, we believe that a classical approach to extract traits related to distinct crop ideotypes based on *a priori* knowledge is more suitable (see also van Eeuwijk et al., 2019; Bustos-Korts et al., 2019). This approach may then represent a standard to compare modern learning approaches with.

Based on HTFP literature and the biological processes described in Ma et al. (2002), we identified three main intermediate trait categories which can be related to ideotype concepts:

1. **Timing of key stages**: Turning points in the dynamics of numeric measurements which may be related to phenology; e.g., beginning of stem elongation (Kronenberg et al., 2017), time of canopy closure (Soltani and Galeshi, 2002), time of maximum canopy growth rate (Borra-Serrano et al., 2020), heading and flowering (Sadeghi-Tehran et al., 2017), or onset and end of senescence (Anderegg et al., 2020; Aasen et al., 2020). Genotype-specific responses to environmental covariates and/or indices may help to predict key stages; e.g., flowering time (Millet et al., 2019).
2. **Quantities at defined time points or periods**: Traits based on numeric measurements; either at a steady state; e.g., canopy temperature between flowering and beginning of senescence (Perich et al., 2020), or at well-defined time points; e.g., number of tillers at beginning of stem elongation (Roth et al., 2020) and at harvest (Jin et al., 2019), number of ears at harvest (Fernandez-Gallego et al., 2018), or canopy cover at maximum (Borra-Serrano et al., 2020). Area-under-the-curve traits may represent a special case of this category where one summarizes quantities over a defined range of time points (Blancon et al., 2019).
3. **Dose-response curves**: Traits that describe developmental responses in dependence on covariates between clearly defined boundary key stages. Dose-response experiments are classically conducted under controlled conditions, e.g., by examining the response of leaves to temperature and water deficit (Reymond et al., 2003) and to soil water deficiency and evaporative demand (Welcker et al., 2011) during their linear growth phase. More recently, such experiments were conducted in the field; e.g., in the early, exponential development phase of canopy cover between emergence and tillering (Grieder et al., 2015) or at the linear development phase of canopy height between start and end of stem elongation (Kronenberg et al., 2020a).

Despite the differences in subsequent processing, the extraction of each of the three different trait categories is a highly repetitive task which requires analysis routines with sufficient robustness and generality. While timing

of key stages and quantities belong to growth model processes, dose-response curves relate to reaction norm processes (Via et al., 1995). Arguably, dose-response curves represent the most challenging modelling aspect in field phenotying, as they require quantifying growth and relate it to environmental covariates. We will cover this aspect in a follow-up paper. However, a robust evaluation of such dose-response curves requires to determine the boundaries between which a steady development takes place. Here, we aimed to develop a method to extract such timing of key stages and quantity traits.

We start by developing a plot-level model for repeated measurements, with a focus on the outdoor field phenotyping platform FIP (Kirchgessner et al., 2017). The FIP allows to densely monitor a large set of replicated genotypes ($\geq 2\times300$) over a whole growing season with genotypes being the only treatment. The aim of such experiments is to i) allow developing new traits and phenotyping methodologies; ii) characterize a specific target environment including the targeted ideotypes; and iii) to serve as part of a multi-environment experiment that covers a mega-environment. We propose a possible solution to analyze such experiments based on existing statistical tools such as P-spline fitting and stage-wise linear mixed model analysis. We further evaluate and demonstrate the suitability of the approach to extract the timing of key stages and quantities at defined time points from low-level traits using simulated wheat canopy height data.

2. Materials and Methods

2.1. A plot-level model for repeated measurements

A planned experiment generally includes an experimental design (Figure 1b, green boxes) in which the treatment factors to be tested are randomly assigned to experimental units (usually plots). For the FIP, the only treatment factor are genotypes. The design comprises only one site but multiple years. The data for each year holds a subset of treatment levels (genotypes) together with checks and design factors (blocks) to allow correcting for spatial variability at the site. In the specific case of the FIP, a panel of on average 345 genotypes is replicated twice per year and each replication is planted on a different lot in the FIP area. Each replication is augmented with spatial checks in a 3×3 block arrangement.

Performing measurements includes the application of a sensing device collecting measurements from a plot (Figure 2a-c). This process results in data which either directly represent a trait value (e.g., a point measurement of temperature) or can be translated to one or several low-level traits by means of feature-extraction (Figure 1b, blue boxes). Measurements of the same type on the same set of plots, but at subsequent time points, can be summarized as campaigns (Figure 1b, red boxes). The sensing concept and the degree of automation determine the time intervals at which measurements can be taken.

Nowadays, the measuring intervals of campaigns often last days rather than minutes or hours. Here, we define a campaign time point as the time at which the whole set of plots (usually belonging to an experimental design or year-site) is completely measured. Such a measurement may take seconds to hours depending on the

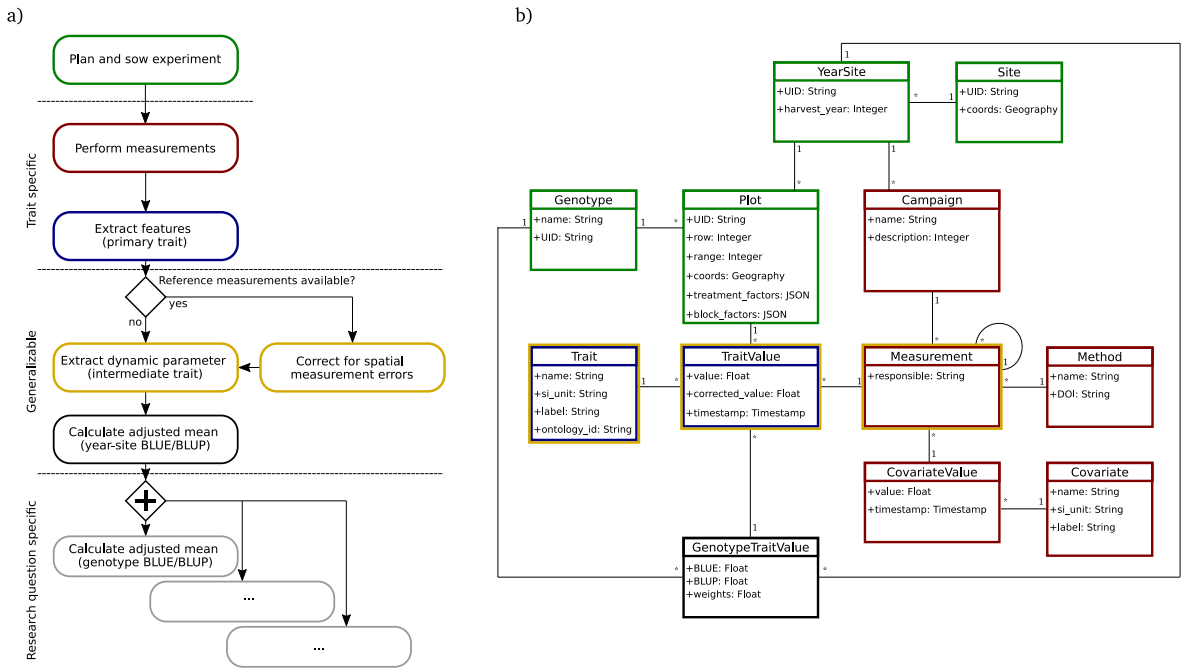

Figure 1: Minimal process-driven model for the FIP: a) Process model, b) Data model.

phenotyping methodology and size of the design. By contrast, a measuring time point (or timestamp) is the exact timing at which an experimental unit (usually a plot) is measured.

The same approach of campaigns and measurements also applies to covariate measurements. Covariates are usually measured at very short time intervals of minutes to hours. In contrast to traits, environmental covariates can be measured at various levels (e.g., year-site, plot, plant, or plant organ). The measurement level determines what is regarded as phenotypic heterogeneity caused by covariate variation. The FIP site comprises a site-specific local weather station, which corresponds to covariate measurements at year-site level (Figure 3b). Therefore, one must consider heterogeneity caused by covariate variations at plot, plant and organ levels and their effects on plant growth (Figure 3a)—namely variations of the timing of key stages (Figure 3c1) resulting in growth period condition variations (Figure 3c2) and consequently variations of quantities at defined time points (Figure 3c3).

In a phenotyping experiment one has to distinguish between nuisance factors affecting the growth and development of the plant (Figure 3a1–3), and measurement errors affecting the precision at which a certain phenotype is measured at a given time (Figure 3a4). The latter factors may affect whole campaign time points (i.e., at the day-to-day level, Figure 3c4, red one-sided arrow) but also individual measuring time points within a day (Figure 3c4, red two-sided arrow). Nuisance factors affecting growth and development are, e.g., soil fertility inhomogeneities, spatial temperature gradients, mices, herbivore damages, and other abiotic and biotic factors varying spatially and temporally in the field. Such factors will in the best case lead to spatio-temporally correlated phenotypic observations.

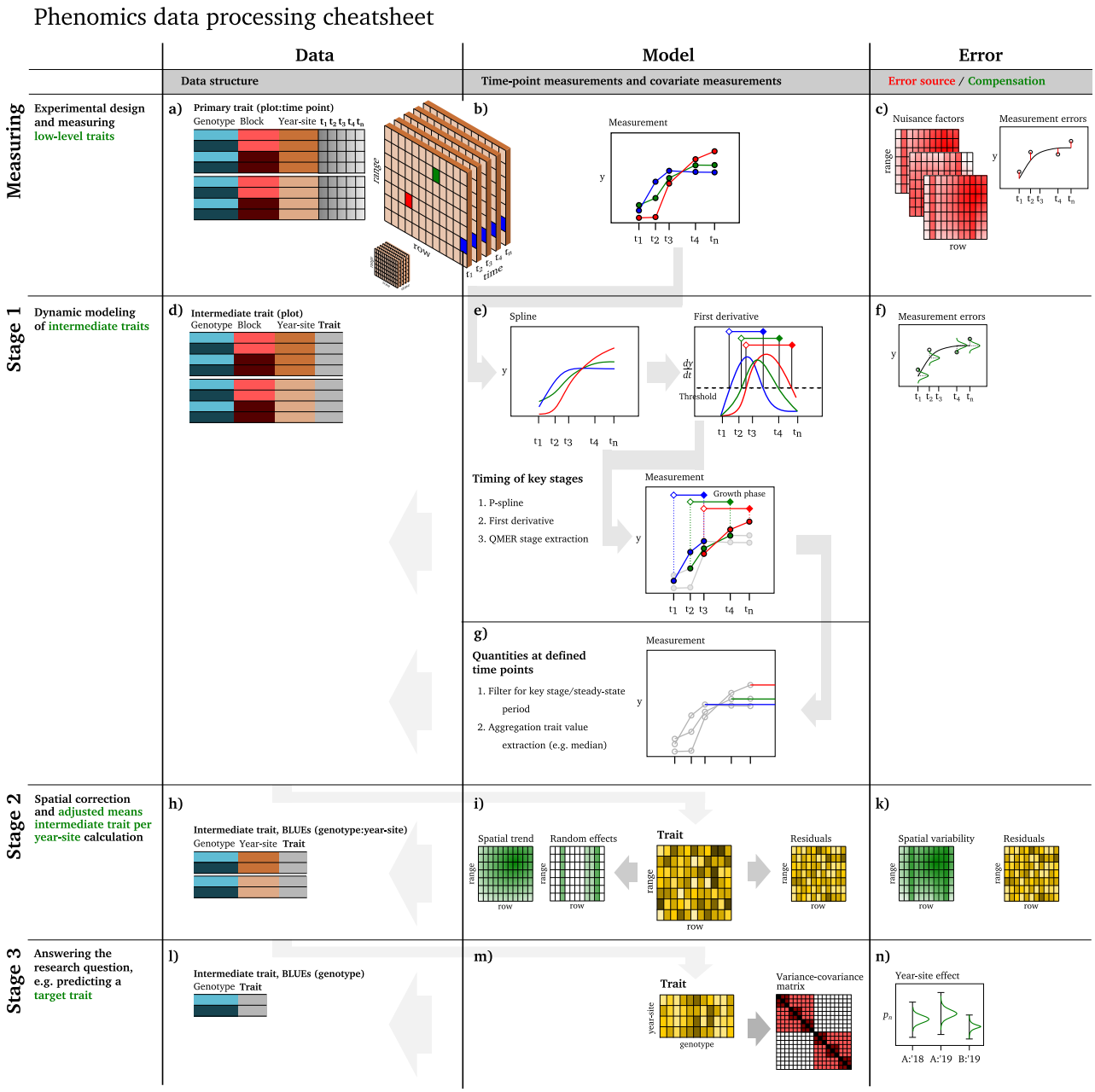

Figure 2: Phenomics data processing cheatsheet: Extraction of timing of key stages and quantities at defined time points from high-throughput field phenotpying data.

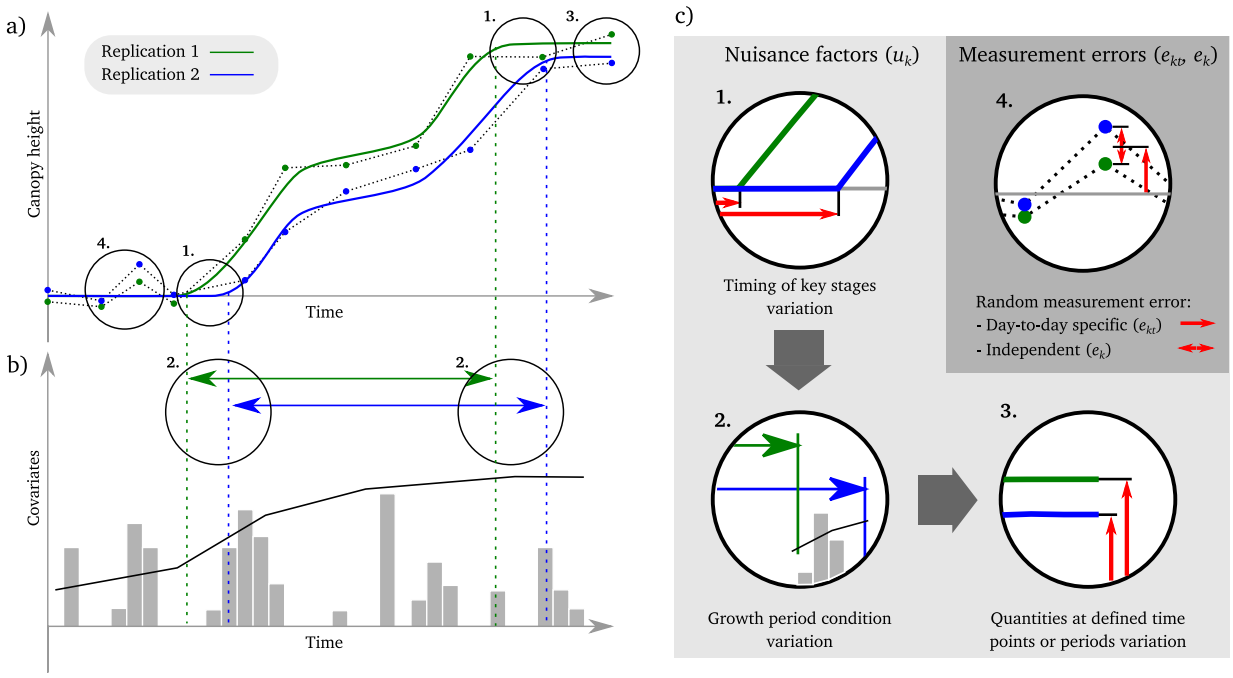

Figure 3: Sources of variation in HTFP on the example of canopy height measurements. (a) Canopy height development of two replications of the same genotype (green and blue lines) and realized measurement time points (green and blue points). (b) Covariate measurements during the growth phase of canopies (e.g., temperature and precipitation). (c) Sources of variation: (1) spatial and crop-husbandry effect leading to different timings of key stages, e.g., start and end of stem elongation; (2) timing of key stage variations leading to variations in the different gradients of environmental covariates, e.g., temperature gradients in the stem elongation phase; (3) spatial and crop-husbandry effects leading to quantitative variations in trait values; e.g., final height at the end of the stem elongation phase; (4) Day-to-day random measuring errors, e.g., related to differing conditions between measurement days; and independent random measuring errors, e.g., related to the measurement precision of the device.

Sources for measurement errors are, e.g., factors differing between campaign time points. These factors may lead to day-specific under- or overestimation of measurements, e.g., due to positioning shifts of the sensor head, re-adjustment of sensor settings between measurement campaigns, changes in canopy characteristics after rain or during hot days, and differing illumination conditions (Figure 3c4, red one-sided arrow). Taking reference measurements (e.g., by the use of calibration targets) allows correcting for some of these errors, but such measures may not always be feasible in crop phenotyping experiments. Apart from the effects related to the whole campaign time point, changing conditions during the measuring sequence may lead to additional, temporally correlated measurement errors among measuring time points. Sources for such errors are, e.g., changing weather conditions during a measurement that takes a considerable amount of time. Such temporal effects may translate into apparent spatial effects within a campaign time point and therefore be confounded with nuisance factors. Finally, random measurement device errors (Figure 3c4, red two-sided arrow) represent another source of variation in HTFP. These errors are usually assumed to be identically and independently normally distributed.

Consequently, we define a HTFP observation y_{kt} for the t-th time point on the k-th plot ($k = 1, \ldots, K$) as the result of a dynamic model g that is a function of time t and of a vector ($\vec{}$) of plot-specific crop growth parameters $\vec{\theta}_{k(i)}$ associated with genotype i modulated (;) by a vector of time-varying covariates $\vec{x}_t$, and of a plot residual e_{kt} that is i.i.d. ($\sim \mathcal{N}(0, \sigma_k^2)$),

$$y_{kt} = g(t, \vec{\theta}_{k(i)}; \vec{x}_t) + e_{kt}\,. \tag{1}$$

While e_{kt} will account for random measurement device errors, we assume here that g will absorb any spatio-temporal correlation among measurements. Dynamic modeling (Equation 1) is done separately for each individual plot-based time series (Stage 1), i.e., $(y_{k1}, \ldots, y_{kT})$ (Figure 2d-g).

Stage 1 therefore estimates plot-specific crop growth model parameters $\vec{\theta}_{k(i)}$. Those crop growth model parameter will become a phenotypic trait when measured / estimated at a set of genotypes. Correcting for spatial correlations is done in a subsequent stage (Stage 2) of a stage-wise approach to get estimates of genotype specific crop growth model parameters $\vec{\theta}_i$ (Figure 2h-k). This estimation step is done separately for each crop growth model parameter in $\vec{\theta}_i$ based on fitting the linear model

$$\widehat{\theta}_{k(i)} = \theta_i + u_k + e_k\,, \tag{2}$$

where $\widehat{\theta}_{k(i)}$ is the crop growth parameter estimate from Stage 1, u_k a spatially correlated random component, and e_k are plot residuals assumed to be normally distributed with zero mean and $\mathrm{var}(e_k) = \sigma^2 w_k^{-1}$, where $w_k = (\text{s.e.})_k^{-2}$ are weights based on the standard error estimates (s.e.) from Stage 1. For a stage-wise approach with weights based on variance estimations, one usually fixes σ^2 to unity. Nevertheless, if expecting proportionality of $\mathrm{var}(e_k)$ to w_k^{-1} only—e.g., when the s.e.'s are derived from a correlated trait—it is required to estimate σ^2. The spatially correlated error term u_k will absorb any spatial correlation caused by random measurement errors and by physical phenotypic differences, and e_k any plot-specific residual.

This approach is not limited to parametric or dimensionality reduction techniques but allows including arbitrary dynamic models g with high complexity based on biologically meaningful traits. Nevertheless, it also obviates modeling a spatio-temporally correlated residual term in its full extent by assuming that all serial correlation is accounted for by the time-dependence of g. In the following, we hypothesize and exemplify with a simulation that our approximation of the spatio-temporal correlation structure is well suited to extract intermediate traits with adequate precision from HTFP data.

2.2. Dynamic modeling of three trait categories

In dynamic modeling, one has to specify a method, based on g of Equation 1, to estimate a vector of meaningful plot-level traits $\vec{\theta}_{k(i)}$ (for brevity we henceforth drop the index i for genotypes, referring to $\vec{\theta}_k$, it being understood that a plot-level parameter is always genotype-specific) based on measured phenotypes y_{kt} and measured covariates x at (potentially differing) time points t. In the following, we will provide theoretical considerations and specific examples for each of the three trait categories defined in the introduction, (1) timing of key stages' traits, (2) quantities at defined time points or periods, and (3) dose-response curve traits.

The first intermediate trait category—timing of key stages—describes growth as a sequence of key stages. Such phenology traits are well-known in agronomy, e.g., the timing of jointing, heading, and flowering in wheat.

The second intermediate trait category—quantities at defined time points or periods' traits—describes phenotypic characteristics at key stages or steady state phases. Hence, such traits include a time point definition, e.g., with traits of the first category. The number of tillers per plant at jointing, the number of ears per square unit at harvest, or the average canopy cover between tillering and jointing are examples of such traits for wheat.

The third intermediate trait category—dose-response curves—describes phenotypes as the result of a dose-response model dependent on a covariate course between key stages. Hence, also these traits require time point definitions, e.g., with traits of the first category. The response of the stem elongation to temperature is an example of such a trait for wheat.

To obtain traits of the first two categories, we favor semi-parametric approaches (e.g., spline fitting) over parametric approaches (e.g., logistic regression) for the dynamic modeling based on the following considerations: Taking the example of early canopy development of winter wheat, where one wants to extract a timing (1) or quantity (2) trait at a specific stage, growth may fluctuate strongly due to the environment, leading to a "stepped" growth curve (Figure 3a). While non-parametric approaches are able to follow such growth curves, parametric approaches would require to modify the timescale to, e.g., thermal time. Despite the fact that thermal time is a widely accepted concept in agriculture (Parent et al., 2019), it is nevertheless based on model assumptions such as the existence of a base temperature, and the linearity of the response. Using such a scale is therefore at odds with the research aim to identify the model behind timing and quantity traits.

When using a semi-parametric approach (e.g., P-splines), one approximates g with a plot-specific model, i.e. a smooth function of time $s(t)$. To extract traits of the first category—timing of key stages—from such a smooth

function, a set of methods q_n $(n = 1, \ldots, N)$ to estimate timing traits $\theta^{T(n)}$ (e.g., to approximate the end of the stem elongation phase) from s has to be defined,

$$g(t, \vec{\theta}_k; x_t) \mathrel{\widehat{=}} s_k(t), \tag{3}$$

$$\theta_k^{T(n)} = q_n(s_k), \tag{4}$$

where $\mathrel{\widehat{=}}$ indicates that s_k estimates g for the k-th plot.

Extracting traits of the second category—quantities at defined time points or periods—builds on the spline function s (Equation 3) and extracted timing of key stages (Equation 4) but inverts the approach of extracting key stages: If $\theta^{T(n)}$ represent timing of key stages (e.g., the end of stem elongation), then quantities at defined time points $\theta^{Q(n)}$ (e.g., canopy cover at the approximated end of stem elongation) may be extracted from the spline s as

$$\theta_k^{Q(n)} = s_k(\theta_k^{T(n)}). \tag{5}$$

It is important to note that the underlying low-level traits for the timing trait $\theta^{T(n)}$ and the spline s in Equation 5 may differ, giving rise to a vast amount of possible trait combinations, e.g., when combining canopy height timing traits with canopy cover quantity traits. While Equation 5 extracts quantities at points in time, extracting aggregated quantities (e.g., normalized area-under-the-curve traits) for a period of time may be of interest as well. If $\theta^{T(a)}$ and $\theta^{T(b)}$ represent two cautiously chosen timings of key stages' traits where $\theta^{T(a)} < \theta^{T(b)}$ (e.g., approximated start and end of flowering), then a quantity at defined time period trait $\theta^{Q(a\ldots b)}$ (e.g., average temperature at approximated flowering) may be extracted from s as

$$\theta_k^{Q(a\ldots b)} = \frac{1}{\theta_k^{T(b)} - \theta_k^{T(a)}} \int_{\theta_k^{T(a)}}^{\theta_k^{T(b)}} s_k(t)\, dt\,. \tag{6}$$

If either $\theta_k^{T(a)}$ or $\theta_k^{T(b)}$ corresponds to a time series boundary (e.g., end of stem elongation to end of time series), the trait may represent an initial or final trait value (e.g., final height).

For the third trait category—dose-response curves—one describes a phenotype as the result of a dose-response model $\dot{g}$ that relates growth rates to a covariate course x_t and a corresponding set of crop growth model parameters $\theta^C = (\theta^{C(1)}, \theta^{C(2)}, \ldots, \theta^{C(L)})$ where L is the total number of parameters of the dose-response curve,

$$g(t, \vec{\theta}_k; x_t) = \int_{\theta_k^{T(a)}}^{\theta_k^{T(b)}} \dot{g}(\theta_k^C, x_t)\, dt\,. \tag{7}$$

Similar to quantities at defined time periods' traits (Equation 6), dose-response curve traits require the definition of a corresponding growth phase, characterized by a start ($\theta^{T(a)}$) and a stop ($\theta^{T(b)}$). Therefore, a preliminary extraction of traits of the category one (Equation 4) is required. Subsequently, θ^C may be estimated.

The striking similarity of Equation 6 and 7 is no coincidence: The area-under-the-curve of a defined growth period can be seen as a direct cause of a response to covariates in this growth phase. The two approaches differ

in how they include covariates: While dose-response curves model an explicit dependency to covariates, an area-under-the-curve quantifies implicitly the result of such a dependency.

An example for a dose-response curve $\dot{g}$ at a defined growth phase is the stem elongation rate of wheat in relation to temperature. Extracting such a dose-response curve implies that one is interested in fitting a specific non-linear function.

2.3. Combining multi-year measurements

HTFP platforms such as the FIP are usually run on a continuous basis, thus increasing the number of year measurements with each year of operation since inauguration. Experimental designs and genotype sets may change to some extent along the years. The question is how to combine such multi-year measurements in a way that one can process years in stages, which is of high benefit for both documentation purpose and processing requirements.

The problem of stage-wise analysis we are addressing here has a long history (Cochran, 1954) and is well known in plant breeding (Smith et al., 2005; Piepho et al., 2012) and also in other contexts, most notably in meta-analysis (Whitehead, 2002; Borenstein et al., 2009). Most commonly, the problem arises in settings were information needs to be combined across several experiments, whereas in the present work we consider the case where different pieces of information need to be combined across units in a single experiment. Despite these differences in scale, the statistical challenges are the same. To illustrate, consider a simple setting in which a set of replicated genotypes is tested for yield at a number of years in a platform. The response of the i-th genotype on the k-th plot at the j-th year can be written as

$$y_{ijk} = \mu + g_i + v_j + (gv)_{ij} + e_{ijk}, \tag{8}$$

where μ is an intercept, g_i is the main effect of the i-th genotype, v_j the main effect of the j-th year, assumed to be normal with zero mean and variance σ^2_v, $(gv)_{ij}$ is the interaction of the i-th genotype and j-th year assumed to be normal with zero mean and variance σ^2_{gv}, and e_{ijk} a residual error assumed to be normal with mean zero and year-specific variance $\sigma^2_{e(j)}$. An objective among others in field phenotyping platforms is to estimate genotype means across years, $\eta_i = \mu + g_i$ and their differences.

This can be done in a single stage by fitting the model (Equation 8) directly to plot data y_{ijk}. Alternatively, we may proceed in two stages and first estimate genotype means per year using sample means $\overline{y}_{ij\cdot}$. These means have variance $\text{var}(\overline{y}_{ij\cdot}) = r_{ij}^{-1}\sigma^2_{e(j)}$, where r_{ij} is the number of replications of the i-th genotype in the j-th location. In the second stage, we can fit the model

$$\overline{y}_{ij\cdot} = \mu + g_i + v_j + (gv)_{ij} + \overline{e}_{ij\cdot}, \tag{9}$$

where $\text{var}(\overline{e}_{ij\cdot}) = r_{ij}^{-1}\sigma^2_{e(j)}$, which is the conditional variance of the genotype means computed in the first stage. The estimates of genotype means, $\eta_i = \mu + g_i$, are identical for single-stage and two-stage analysis, provided the

variance components are known (Piepho et al., 2012). Differences arise in practice because variances need to be estimated. Stage-wise analysis entails an approximation of the gold standard of single-stage analysis because variances $\text{var}(\overline{y}_{ij\cdot}) = r_{ij}^{-1}\sigma^2_{e(j)}$ as estimated in the first stage are treated as known quantities in the second stage, disregarding the degrees of freedom associated with these estimates and their uncertainty. A key feature of stage-wise analysis is that the inverses of these estimated variances act as weights in the second-stage analysis. A major challenge in any stage-wise analysis is how to best determine the weights and how to account for the uncertainties associated with them.

The situation faced in the analysis of HTFP is comparable in that it proceeds in stages with necessity because a single-stage analysis is in conflict with performance and generalization demands (i.e., multi-year HTFP data may comprise a number of differing experimental designs that require individual processing in a first stage) and that the primary interest is the genotype main effect g_i, which equals θ_i in HTFP (Figure 2l-n). The statistical challenges are rather more daunting, however, for several reasons: (i) HTFP involves high-frequency time series in which observations are serially correlated; (ii) summarizing time-series data usually requires nonlinear regression models; (iii) analyses of field trials are often done exploiting spatial correlations among neighboring plots; (iv) remote or proximate sensed data are affected by environmental conditions (wind, illumination) and may change during the course of a measurement; (v) the number of stages required for the full analysis process is much greater than two. These additional features make the determination of appropriate weights to be carried forward from one stage to the next even more challenging than in the simple example given above.

Here, we propose a weighing approach for the intermediate trait category (1) (timing of key stages) and (2) (quantities at defined time points or periods) only for brevity, and illustrate its application using a simulation study described in the following section. Traits of the third category (dose-response curves) will be considered in a follow-up paper.

2.4. Simulation of canopy height data

To demonstrate the extraction of traits of the first two categories (timing of key stages and quantities at defined time point or periods), winter wheat canopy height data were simulated implementing a temperature dose-response curve (trait category three, Equation 7). The temperature response of the stem elongation phase was assumed to follow a dose-response curve with break points (Figure 4),

$$r_{\text{BP}}(T, \theta^C) = \begin{cases} 0, & T < T_{\text{min}} \\ r_{\text{max}}, & T > T_{\text{opt}} \\ r_{\text{max}} \cdot \frac{T - T_{\text{min}}}{T_{\text{opt}} - T_{\text{min}}}, & \text{otherwise}, \end{cases} \tag{10}$$

where T_{min} is the base temperature below which the elongation rate r is zero and T_{opt} the optimum temperature above which the elongation rate reaches the maximum hourly elongation rate r_{max}, while $\theta^C = (r_{\text{max}}, T_{\text{min}}, T_{\text{opt}})$ (Figure 4).

297 As starting point for the simulation, existing experimental designs of three consecutive years at the ETH
298 research station of agricultural sciences in Lindau Eschikon, Switzerland (47.449 N, 8.682 E, 556 m a.s.l.) were
299 used. The experiment consisted of 352 wheat genotypes, replicated twice per year on two spatially separated
300 fields, both augmented with spatial checks in a 3×3 block arrangement.

301 To simulate canopy height time series, existing weather data were used to introduce a close-to-realistic
302 stochastic behavior. Canopy growth was simulated for a measurement interval of one per day and for a pe-
303 riod between first of March and 20th of July ($d = 1, \ldots, 142$) for each of the three simulated years j ($j =$
304 $2016, 2017, 2018$). Growth between daily campaign time points t was modeled as cumulative response to hourly
305 temperature measurements T_{jdh} ($h = 1, \ldots, 24$). The canopy height y_{ijkt} of genotype i ($i = 1, \ldots, 352$) at plot k
306 ($k = 1, \ldots, 704$) in the year j at a specific time point ($t = 1, \ldots, 142$) was then simulated as

$$y_{ijkt} = g_T(t, \theta^C_{ijk}, \theta^T_{ijk}; T_{jdh}) + e_{jkt}\,, \tag{11}$$

307 where g_T depends on r_{BP} in Equation 10 (see below) and simulates growth as a function of temperature T_{jdh},
308 time point t, a vector of plot-specific crop growth model parameters $\theta^C_{ijk} = (r_{\max}, T_{\min}, T_{\text{opt}})$, and a vector of
309 plot-specific timing traits $\theta^T_{ijk} = (t\text{PH}_{\text{start}}, t\text{PH}_{\text{stop}})$. The error term e_{jkt} simulates plot and time point residuals.
310 The growth function g_T was specified as

$$g_T(t, \theta^C, \theta^T; T_{dh}) = \sum_{d=1}^{t} \begin{cases} \sum_{h=1}^{24} r_{\text{BP}}(T_{dh}, \theta^C) & t\text{PH}_{\text{start}} < d < t\text{PH}_{\text{stop}} \\ 0, & \text{otherwise} \end{cases}, \tag{12}$$

311 where r_{BP} represents a dose-response as function of hourly temperatures T_{dh} and a vector of crop growth model
312 parameters θ^C (Equation 10), $t\text{PH}_{\text{start}}$ the time point where canopy growth started, and $t\text{PH}_{\text{stop}}$ the time point
313 where canopy growth stopped.

314 This approach produced realistic-looking canopy growth curves (compare Figure 5 with, e.g., real data in
315 Kronenberg et al. 2017, 2020a) with a characteristic start of growth ($t\text{PH}_{\text{start}}$) and a stop of growth ($t\text{PH}_{\text{stop}}$),
316 corresponding to the first intermediate trait category (timing of key stages). Additionally, growth curves indi-

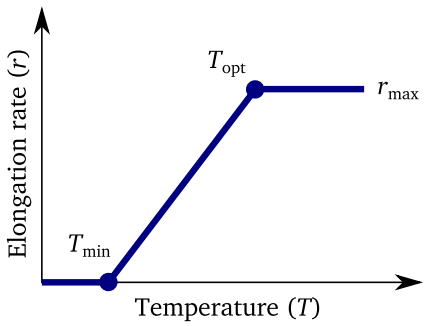

Figure 4: Schematic drawing of the dose-response model ($\dot{g}$ of Equation 7) implemented as break-point model (r_{BP}, Equation 10) used for the simulation of canopy height time series based on temperature courses.

cated a characteristic final height (PH_{max}), corresponding to the second intermediate trait category (quantities at defined time points or periods).

Noise as specified in Section 2.1 was introduced on a genotype, plot and time point level. Genotype-year interactions were not explicitly introduced, as it was assumed that they will emerge as the result of random θ_i^C and θ_i^T combinations applied to year-specific temperature courses.

To add noise to genotype traits, the crop growth model parameters θ_{ijk}^C and the timing traits θ_{ijk}^T were further decomposed in genotypic and spatially correlated parts,

$$\theta_{ijk}^C = \theta_i^C + \theta_{jk}^C \tag{13}$$

$$\theta_{ijk}^T = \theta_i^T + \theta_{jk}^T\,, \tag{14}$$

where θ_i^C and θ_i^T were simulated using normal distributions ($\sim \mathcal{N}(\mu, \sigma^2)$). Trait-specific μ and σ^2 were chosen based on literature if available, and otherwise based on own unpublished field data. θ_{jk}^C and θ_{jk}^T were spatial correlated heterogeneity components for those traits ($AR(1)_x \otimes AR(1)_y$), where AR(1)⊗AR(1) is a two-dimensional first-order autoregressive model in row (x) and range (y) direction, mimicking the influence of other covariates and therefore spatial heterogeneity. Note that a high autocorrelation in row and range direction with $\rho = 0.95$ and half the variance of the corresponding input parameter (Appendix Table 2) was assumed, which appeared reasonable for cereal experiments (e.g. Velazco et al., 2017).

The plot residual e_{jkt} was simulated as sum of three error terms,

$$e_{jkt} = e_{jkt,1} + e_{jk} + e_{jkt,2}\,. \tag{15}$$

The first error term $e_{jkt,1}$ corresponds to the serial correlation of measurement errors ($AR(1)_t$) that g in Equation 1 presumably absorbs. The second error term e_{jk} mimics a systematic spatially correlated measurement error, e.g., after an incomplete correction with reference measurements ($AR(1)_x \otimes AR(1)_y$). We note that adding this error introduces an intentional discrepancy between the analysis model and the simulation: the proposed plot-level model for repeated measurements does not include such a systematic error in the first stage (dynamic modeling). Consequently, estimating the spatial correlation in the second stage will confound measurement errors and nuisance factors, which corresponds to a situation we frequently encounter in HTFP. The third error term $e_{jkt,2}$ corresponds to e_{kt} in Equation 1 and represents a plot-based i.i.d. residual ($\sim \mathcal{N}(0, \sigma^2)$). The first error term was assumed to cause most of the known measurement error, wherefore σ was set accordingly to 0.01 m (Roth et al., 2020), while for the second and third error term σ was significantly reduced. The autocorrelation parameter ρ was arbitrary set to 0.7. All simulation input parameters and sources for the aforementioned assumptions are summarized in the Appendix (Table 2).

A total of 500 simulation runs were performed. These simulated time series with a measurement interval of one day were then further thinned to intervals of three, five, seven and 14 days to study the effect of lower frequencies.

We note that the simulation (Equation 11) comprised θ^T, i.e. traits of the first category, and θ^C, i.e. traits of the third category. The second trait category θ^Q was dependent on the first and third category and year specific temperature courses, and therefore only an indirect input parameter of the simulation. Therefore, the simulation allowed extracting traits of all three categories, and validating traits of category one (θ^T) and three (θ^C) with genotypic input data, and traits of category two (θ^Q) with plot-level (indirect) input data. Here, we illustrate the extraction of θ^T and θ^Q only for brevity. The extraction of θ^C and therefore dose-response curve parameters of a crop growth model will be considered in a follow-up paper.

We further note that all simulation input parameters for a given genotype i (θ_i^T and θ_i^C) were uncorrelated. In reality, genetic effects and artificial selection have certainly resulted in weak to strong correlations for those parameters. Dynamic modeling may introduce new, artificial correlations of parameters. When examining a real-world genotype set, e.g., a breeding population, these effects will be confounded, but using a simulation with uncorrelated input parameters allows quantifying the extraction artifacts.

2.5. Stage 1: Extracting the timing of key stages and quantities at define time points

To extract timing of key stages, a monotonically increasing P-spline was fitted to plot time series using the R package *scam* (Pya, 2019), thus implementing $s_k(t)$ of Equation 3. The package fits shape constrained generalized additive models (GAM) (Pya and Wood, 2015). A Bayesian approach to uncertainty quantification is used to obtain standard errors of predictions.

The number of knots was set proportional to 3/4 of the observations. In a next step, the start and end of stem elongation ($t\mathrm{PH}_{\mathrm{start}}$ and $t\mathrm{PH}_{\mathrm{stop}}$) were extracted based on the quarter of maximum elongation rate (QMER) method, which in brief extracts key time points with elongation rates greater than a threshold of 1/4 of the maximum elongation rate. Thus, the QMER method represents an implementation of $q_n(s_k)$ of Equation 4.

In detail, in a first step spline predictions for canopy heights $\hat{y}_t$ and standard error estimates s.e.($\hat{y}_t$) were calculated separately for each plot at hourly time steps using the prediction function of the *scam* package. Thereafter, hourly growth rates $\hat{r}_t$ were derived from the difference between subsequent predictions, $\hat{r}_t = \hat{y}_t - \hat{y}_{t-1}$ (Figure 2e). Then, the following algorithm was applied to extract intermediate traits and corresponding weights w based on standard errors of spline predictions:

1. Determine maximum elongation rate:
   $\hat{r}_{\max} = \max(\hat{r}_t)$
2. Filter $\hat{r}_t$ for data points with an elongation rate greater than $1/4$ of the maximum elongation rate:
   $\hat{r}_{t,\mathrm{set1}} = \hat{r}_t$ where $\hat{r}_t \geq 1/4 \cdot \hat{r}_{\max}$
3. Define the earliest time points that is left after filtering as the start of growth:
   $t\mathrm{PH}_{\mathrm{start}} = t$ of first($\hat{r}_{t,\mathrm{set1}}$)
   ${w_{t\mathrm{PH}_{\mathrm{start}}}}^{-1/2} = \mathrm{s.e.}(\hat{y}_t)$ where $t = t\mathrm{PH}_{\mathrm{start}}$

4. Filter $\hat{r}_t$ for data points with an elongation rate lower than $1/4$ of the maximum elongation rate and a minimum distance of 40 days to the approximated start of growth:
   $\hat{r}_{t,\text{set2}} = \hat{r}_t$ where $\hat{r}_t \leq 1/4 \cdot \hat{r}_{\max} \wedge t - t\text{PH}_{\text{start}} \geq 40$
5. The earliest value that is left after filtering indicates the approximated end of growth:
   $t\text{PH}_{\text{stop}} = t$ of first$(\hat{r}_{t,\text{set2}})$
   ${w_{t\text{PH}_{\text{stop}}}}^{-1/2} = \text{s.e.}(\hat{y}_t)$ where $t = t\text{PH}_{\text{stop}}$

Note that the weights for timing of key stages' traits in this work were based on the standard errors of spline predictions $\hat{y}$. We will address the conditions that should be met to justify our approach in the following section.

We extracted the growth stages start and end of stem elongation ($t\text{PH}_{\text{start}}$ and $t\text{PH}_{\text{stop}}$) and corresponding standard error estimates based on the quarter of maximum elongation rate (QMER) method. To compare the QMER method with the approach taken by Kronenberg et al. (2017), we additionally determined the time points where 15% ($t\text{PH}_{15}$) and 95% ($t\text{PH}_{95}$) of final height was reached (for details, see Kronenberg et al., 2017). In Figure 2e, we depict only the QMER method.

The quantity at a defined time point final height ($\text{PH}_{\max}$) was calculated as the median of 24 hourly spline predictions after the estimated stop of growth:

1. Filter $\hat{y}_t$ for data points after reaching final height:
   $\hat{y}_{t,\text{final}} = \hat{y}_t$ where $t\text{PH}_{\text{stop}} \leq t \leq t\text{PH}_{\text{stop}} + 24$ h
2. Aggregate data points:
   $\text{PH}_{\max} = \text{median}(\hat{y}_{t,\text{final}})$
   ${w_{\text{PH}_{\max}}}^{-1/2} = \text{s.e.}(\hat{y}_t)$ where $t = t\text{PH}_{\text{stop}}$

2.6. Weigting based on estimated standard errors

The chosen implementation of the QMER method did not provide standard errors for the derived growth rate ($\hat{r}$) and time points (t). Therefore, weighting for further processing after the dynamic modeling was based on standard errors of spline-based predictions of the response (s.e.$(\hat{y}_t)$).

Using weights based on the standard errors of spline predictions is intuitive for quantities at defined time points or periods' traits (e.g., $\text{PH}_{\max}$), as both s.e.$(\hat{y}_t)$ and $\hat{y}_t$ share the same unit. Nevertheless, for timing of key stages (e.g., $t\text{PH}_{\text{start}}$ and $t\text{PH}_{\text{stop}}$), such a weighting approach requires a positive and high association between the true weights for t and y for a given (to be determined) time point. Alternatively, one could use an inverse regression approach (e.g., the Fieller's theorem (Seber, 2003) or the delta method (Johnson et al., 1993)) to determine weights for two means with different units. Such an inverse regression approach becomes non-trivial when involving a combination of statistical tools—e.g., P-splines and the QMER method. Therefore, using an inverse regression approach may contradict the requirement to provide a seamless workflow to integrate arbitrary complex dynamic models g (Equation 2).

Consequently, we decided to assume proportionality of weights for standard errors of spline predictions and timing of key stage estimations. The factor of proportionality was estimated via the residual variance (σ^2), which was estimated in each analysis, rather than fixed at unity, as is customary in standard weighted analysis, where the inverse weights are taken to be the known residual variances (Piepho et al., 2012). Our assumption is based on considerations on a concrete example (see Appendix). In addition, standard errors of spline predictions suppose that observations of plot-based time series are independent. As this is—at least for the simulation—not true (see Section 2.1), the calculated standard errors of the estimates will be biased. To test whether weighting was advantageous, despite possible bias in the weights and imperfect proportionality for timing of key stage traits, we optionally provided the weights in the next processing step.

2.7. Stage 2: Calculating adjusted genotype means per year

The extraction of dynamics characteristics resulted in measurement time point independent trait values at a plot level (Stage 1). These plot values were subsequently processed in a two-stage linear mixed model analysis (Stage 2 and 3), where the second stage averaged over within-year effects (e.g., spatial heterogeneity) and the third stage over between-year effects.

We used SpATS (Rodríguez-Álvarez et al., 2018) to fit a model with a smooth bivariate surface defined over spatial coordinates of plot centers ($f(x(jk), y(jk))$) and added fixed genotype effects (θ_{ij}) and random effects of plot rows and ranges ($p_{r(jk)}$ and $p_{c(jk)}$),

$$\hat{\theta}_{jk} = \theta_{ij} + f(x(jk), y(jk)) + p_{r(jk)} + p_{c(jk)} + e_{jk}\,. \tag{16}$$

Model parameters are listed and explained in Table 1 (Stage 2). Stage 2a and 2b are two nested models; Stage 2b corresponds to Stage 2a but additionally includes weights. Equation 16 was applied to all intermediate traits to calculate BLUEs of genotype means per year.

2.8. Stage 3: Genotypic marginal means calculation

The second stage already covered aspects such as spatial heterogeneity and design-specific characteristics such as row and range arrangements, and allowed obtaining adjusted year genotype means $\hat{\theta}_{ij}$ (BLUEs). In the third stage, those means were further processed with a model based on Equation 9,

$$\hat{\theta}_{ij} = \mu + v_j + \theta_i + (\theta v)_{ij} + e_{ij}\,. \tag{17}$$

The model assumes that genotype-environment effects can be partitioned into genotype response effects (θ_i) and genotype-year interaction effects ($(\theta v)_{ij}$) (Piepho et al., 2012) while the residual errors (e_{ij}) are assumed to be identically and independently normally distributed. Model parameters are listed and explained in Table 1 (Stage 3). Stage 3a and 3b are two nested models; Stage 3b corresponds to Stage 3a but additionally includes weights. Models were fitted using the R package R-asreml (Butler, 2018).

Table 1: Model parameters for the second and third stage of the stage-wise linear mixed model analysis. k denotes the k-th plot, j the j-th year, and i the i-th genotype.

Stage	Term	Description	Part
2)	$\hat{\theta}_{jk}$	Plot response based on dynamic modeling	Response
	θ_{ij}	Year genotype response	Fixed
	$p_{c(jk)}$	Range effect on field (main working direction, e.g., for sowing)	Random
	$p_{r(jk)}$	Row effect on field (orthogonal to main working direction)	Random
	$f(x(jk), y(jk))$	Smooth bivariate surface in spatial x and y coordinates (mapping real distances on the field) consisting of a bivariate polynomial and a smooth part (for details see Rodríguez-Álvarez et al., 2018)	Spatial
a)	e_{jk}	Residuals with $\text{var}(e) = \sigma^2$	Residual
b)	e_{jk}	Residuals with $\text{var}(e) = \sigma^2 w^{-1}$, where w are weights based on the standard error estimations from the previous dynamic modeling step (Stage 1), and σ^2 the residual variance parameter	Weights
3)	$\hat{\theta}_{ij}$	Adjusted year genotype mean (BLUE) from Stage 2	Response
	μ	Global intercept	Fixed
	v_j	Year effect	Random
	θ_i	Genotype response	Fixed
	$(\theta v)_{ij}$	Genotype year interaction	Residual
a)	e_{ij}	Residuals with $\text{var}(e) = \sigma^2$	Residual
b)	e_{ij}	Residuals with $\text{var}(e) = \sigma^2 w^{-1}$, where w are weights based on the square rooted diagonal of the variance-covariance matrix from Stage 2, and σ^2 the residual variance parameter	Weights

Separating the dynamic modeling step from further processing steps prevents implementing the gold standard of a one-stage analysis. Nevertheless, subsequent processing stages can be summarized in one stage, hence resulting in a two-stage approach. To allow comparing such an approach with a three-stage approach, the estimated intermediate traits from Stage 1 were additionally processed using a two-stage model,

$$\hat{\theta}_{jk} = \mu + v_j + \theta_i + (\theta v)_{ij} + p_{r(jk)} + p_{c(jk)} + f(r(jk), c(jk)) + e_{jk}\,, \tag{18}$$

where μ is a global intercept, v_j a year intercept, θ_i the genotype response, $(\theta v)_{ij}$ genotype year interactions, $p_{r(jk)}$ and $p_{c(jk)}$ range and row effects, $f(r(jk), c(jk))$ year specific AR(1) $\otimes$ AR(1) interactions based on ranges ($c()$) and rows ($r()$) of plots, and e_{jk} plot residuals with year-specific variances.

2.9. Simulation validation

Bias, variance, root-mean squared error (RMSE) and Pearson's correlation were calculated both after dynamic modeling (Stage 1) and after the stage-wise linear mixed model analysis (Stage 2 and 3) separately for each simulation run.

3. Results

A total of 176,000 genotypes replicated on 1,056,000 plots (500 runs × 3 years × 2 replications × 352 genotypes) containing 149,952,000 data points (number of plots × 142 measurement days) were simulated. In the following, we give insights on the precision of extracted traits influenced by the choice of method, weighting, and measurement interval.

3.1. Dynamic modeling

P-splines model fits converged for all simulated plot time series and produces smooth-looking growth curves (Figure 5). Start and end of stem elongation estimations were successfully extracted using the QMER method as well as the final height percentile method.

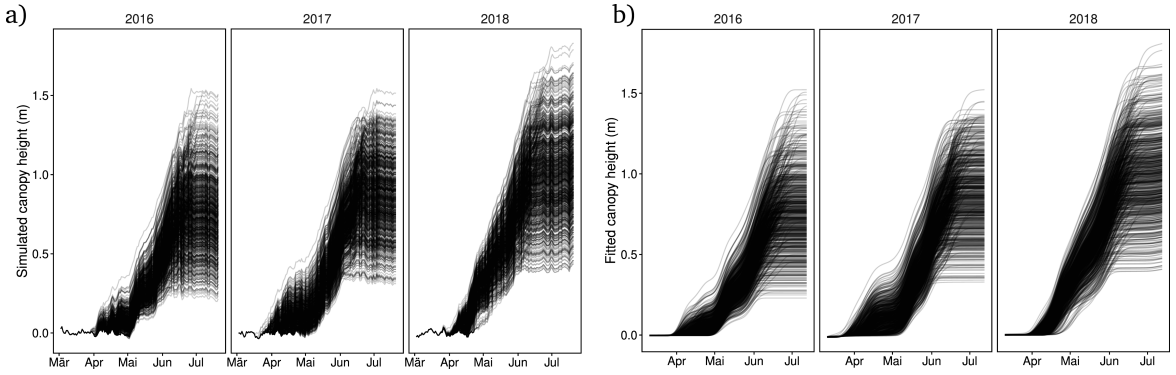

Figure 5: Simulated canopy heights (a) and fitted canopy height splines (b) for one simulation run with 352 genotypes, two replications per year, and three years.

The timing of the key stage trait $t\text{PH}_{\text{start}}$ was better estimated by the P-spline/QMER method with a lower median RMSE and lower median bias (Figure 6). Nevertheless, in comparison to the final height percentile method, the median variance was higher, and larger outliers for RMSE and variance were found. The trait $t\text{PH}_{\text{stop}}$ was better estimated by the final height percentiles method with lower median bias, median RMSE and median variance than by the P-spline/QMER method, but the percentiles method also produced larger outliers for variance and RMSE than the P-spline/QMER method.

Both the P-spline/QMER and final height percentile methods performed comparably and were able to predict $t\text{PH}_{\text{start}}$ with a strong and $t\text{PH}_{\text{stop}}$ with a very strong correlation to input values (Figure 7), but also for both methods, the estimated start of stem elongation ($t\text{PH}_{\text{start}}$) was weakly biased by the input trait base temperature. Nevertheless, the correlation between the extracted start and end of stem elongation—an artifact of the method, as the simulation input was uncorrelated—was much higher for the Percentile method than for the P-spline/QMER method. Based on these findings, the P-spline/QMER model was selected for further processing in the stage-wise analysis.

3.2. Required measurement intervals

Estimating $t\text{PH}_{\text{stop}}$ and PH_{max} using the P-spline/QMER or Percentile method was not affected by increased or reduced measurement intervals unless reduced from 7 to 14 days, where the correlation for both $t\text{PH}_{\text{start}}$ and $t\text{PH}_{\text{stop}}$ dropped (Figure 8). The estimation of $t\text{PH}_{\text{start}}$ was, in contrast to the two other traits, sensitive to reduced measurement intervals above five days for the P-spline/QMER method. The prediction of final height was not affected by increased measurement intervals.

3.3. Stage-wise linear mixed model analysis

For both traits $t\text{PH}_{\text{start}}$ and $t\text{PH}_{\text{stop}}$, calculating overall adjusted genotype means reduced the median variance and median bias if compared to plot-based values for the P-spline/QMER method (Figure 6) and improved the median RMSE for $t\text{PH}_{\text{start}}$ but not for $t\text{PH}_{\text{stop}}$ (Figure 9, Appendix Table 3). Based on bias and variance, for the three-stage model (dynamic modeling followed by a two-stage linear mixed model analysis), weighting Stage 2 with errors of the prediction from dynamic modeling (Stage 1) was of advantage for $t\text{PH}_{\text{start}}$. The lowest median bias was found for the combination of weighting Stage 2 as well as Stage 3 and the lowest median variance for the combination of weighting Stage 3 but not Stage 2 (Figure 9, Appendix Table 3). For $t\text{PH}_{\text{stop}}$, median differences between weighting combinations for Stage 2 and 3 were overall very small, but weighting stage 2 reduced outliers for bias and RMSE.

When compared to a two-stage model (dynamic modeling followed by a one-stage linear mixed model analysis), using a three-stage model was of disadvantage for $t\text{PH}_{\text{start}}$, indicated by a lower median RMSE and larger outliers (Figure 9). For $t\text{PH}_{\text{stop}}$, the median RMSE was slightly higher for the two-stage model than for the three-stage model, but outliers were more frequent for the three-stage model.

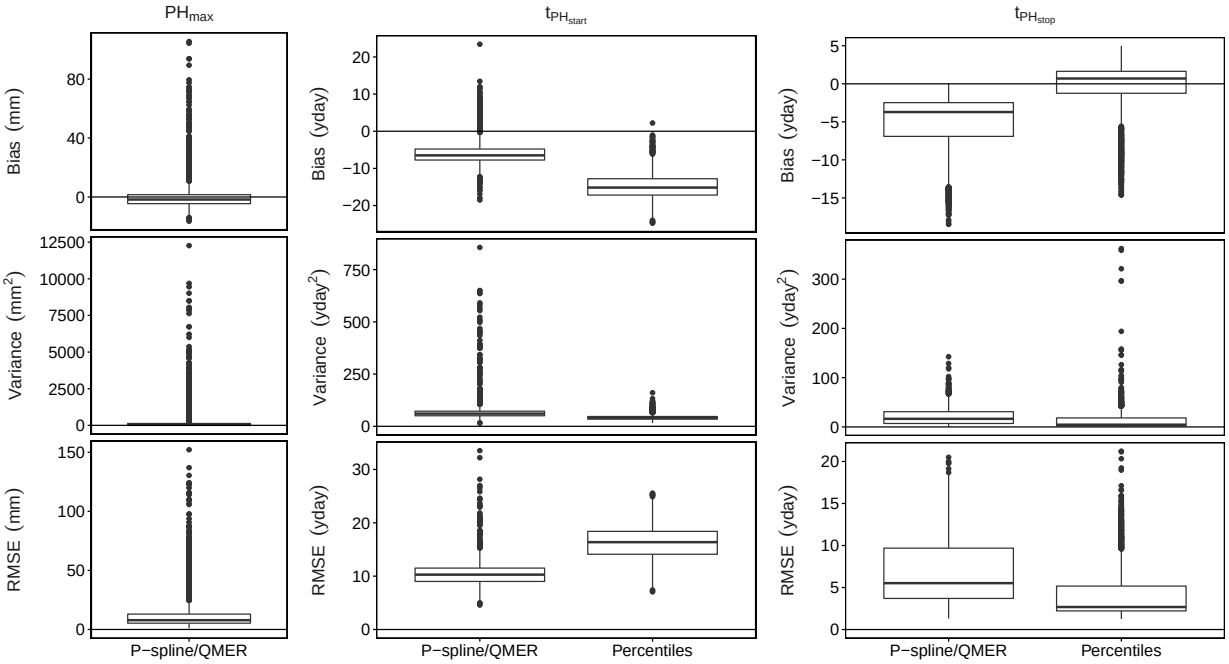

Figure 6: Box plots for the 500 simulated datasets of plot-based bias, variance and root-mean squared error (RMSE) of two timing of key stages models (P-spline/QMER model and final height percentiles).

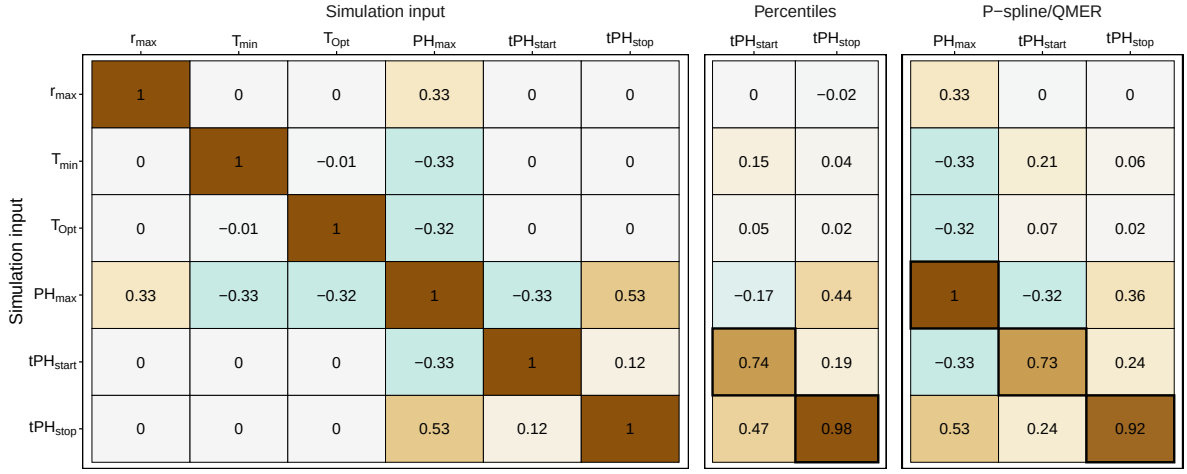

Figure 7: Pearson's correlations of plot time series traits. Provided are simulated input parameters and extracted timing of key stages' traits for the P-spline/QMER and final height percentile model. Black bold boxes indicate correlations between predicted and true values for identical traits.

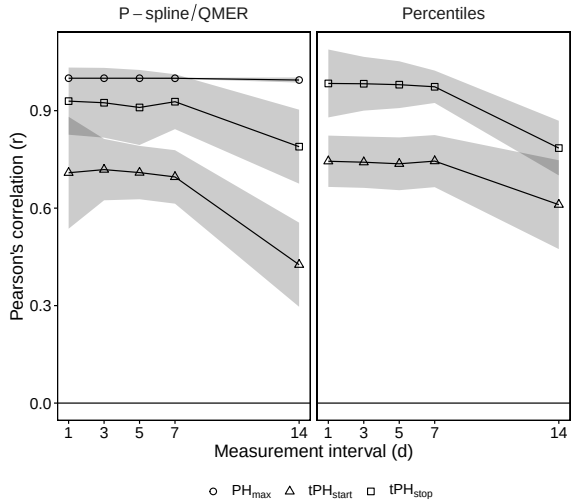

Figure 8: Pearson's correlations for differing measurement intervals for the timing of key stages based on splines (P-spline/QMER method) and final height percentiles (Percentiles method).

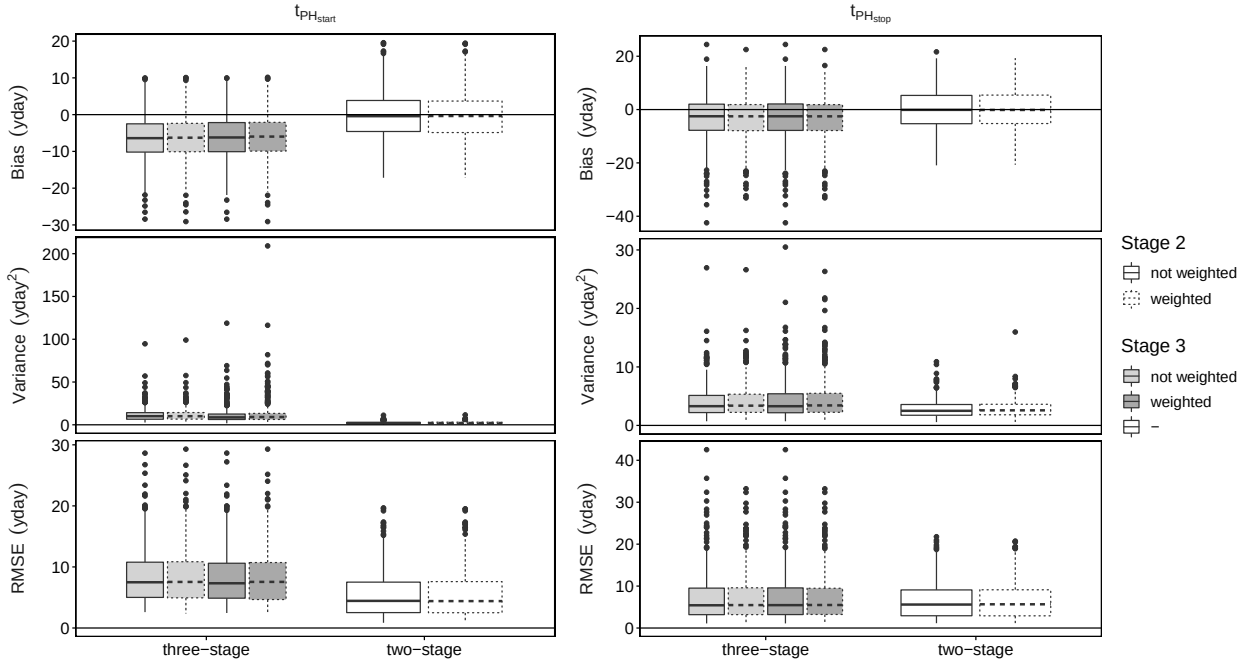

Figure 9: Box plots for the 500 simulated datasets of genotype based bias, variance, and root-mean squared error (RMSE) for the key stages P-spline/QMER model for the three-stage model and the two-stage model.

4. Discussion

4.1. Data processing in stages

The overall workflow of HTFP requires a joint effort of disciplines (Cobb et al., 2013; Araus and Cairns, 2014) which may be separated into three main domains: (1) automation and sensing including feature extraction from sensor readouts, (2) applied phenotyping including dynamic modeling and trait extraction from sensor-derived features, and (3) analysis of designed agricultural experiments or breeding experiments. Plant phenomics must bridge these three disciplines with the overall aim to characterize phenotypes as the result of genotype, environment and management. A plot-level model for repeated measurements may help to link the highly specific domains of sensing and the analysis of experiments. The link to genomic information in breeding and quantitative genetics further increases the complexity of the topic, but is only marginally addressed in this study.

Here, we presented a strategy to process HTFP data. Based on the evaluated sources of variation, we decided to process in stages, starting with dynamic modeling, followed by two stages of a linear mixed model analysis. This approach is to some extent the reverse of van Eeuwijk et al. (2019) who suggested correcting time point measurements in a first stage of a stage-wise linear mixed model analysis, followed by dynamic modeling and modeling of environmental dependencies, and a second stage of a stage-wise linear mixed model analysis to calculate adjusted means across years. Both options—correcting for spatial or temporal correlations first—represent valid alternatives. In the present case, we decided not to correct for spatial gradients before dynamic modeling for two reasons:

(1) Calculating adjusted genotype means in a first stage will correct for spatially correlated measurement errors, but also for effects caused by start and lag phase variations, quantitative trait variations, and environment variations due to start and lag phase variations (Figure 3). While correcting for measurement errors is a desired effect, correcting for other effects will bias the result of dynamic modeling by altering the time point variances of low-level traits. Using measurement references and correcting for day-to-day random measurement errors outside the framework of the stage-wise processing may therefore be of advantage. (2) In a linear mixed model analysis interlaced by a dynamic modeling part, weighting becomes a daunting task. In opposite, weighting is an integral part of the stage-wise analysis strategy presented in this study.

For the P-spline/QMER method, processing multiple years using a linear mixed model analysis reduced the variance and bias of predictions while slightly increasing the RMSE. Weighting the first stage recovered the RMSE. For the second stage, using weights based on estimated variances to approximate the gold standard of a single-stage analysis proved to be of advantage for all traits if using meaningful weights for the first stage as well. These findings indicate that our assumption about dynamic modeling was justified: the spatio-temporal correlation caused by unconsidered covariates yields spatial correlated intermediate traits $\vec{\theta}_{ijk}$. Nevertheless, using a one-stage linear mixed model with an $AR(1) \otimes AR(1)$ autocorrelation structure outperformed the stage-wise approach for $t\mathrm{PH}_{\mathrm{start}}$ and to some extent for $t\mathrm{PH}_{\mathrm{stop}}$.

Providing robust and reusable analysis routines represented an essential objective of the proposed approach. The resulting generalization requirements may be in conflict with well-established analysis principles. This conflict became well visible when formulating a linear mixed model for Stage 2: The philosophy "analyse-as-randomise" would require to include all randomization factors—e.g., incomplete blocks—in the analysis. A generalized model as used in this work that includes besides a smooth bivariate surface just row and range effects is certainly less efficient, but may nevertheless be suitable to draw correct conclusions on the outcome of the experiment. Proposing a robust and reusable processing workflow therefore always represents a trade-off between generalization and most efficient modeling.

4.2. Intermediate trait categories

In this study, we proposed three different trait categories: (1) Timing of key stages, (2) quantities at defined time points or periods, and (3) dose-response curves. A fundamental difference between traits of the first two categories and dose-response curve traits is how they include covariate dependencies. Dose-response curve traits describe an explicit dependency on covariates. In contrast, timing of key stages' traits include the effects of covariates implicitly through the dependency on the timescale: Favorable conditions in spring may for example accelerate the development of plants and therefore early key stages. Quantities at defined time points or periods' traits may show a similar behavior, but here the directions are less clear: Early jointing in cereals due to favorable conditions in spring may for example reduce the early canopy cover in the corresponding phase because of a reduced growing time span. Nevertheless, one may also argue that favorable conditions in this reduced time span may increase canopy cover. Both categories have in common that they describe an implicit reaction to a set of covariate courses.

Consequently, to analyze traits of the first two categories, one reduces growing seasons with their characteristic covariate courses to environments (E) and quantifies the influence of genotypes (G) and environments on measured traits in a subsequent G×E analysis (for an overview see van Eeuwijk et al., 2016). In contrast, dose-response curve traits are less affected by—but rather drivers of—G×E. This difference may require differing processing steps. We will cover dose-response curves in a follow-up paper.

4.3. Limitations of dynamic modeling

Clear limitations of the proposed approach became visible: Although all input parameters of the simulation were completely uncorrelated, the extracted traits were to varying extents correlated. The simulation consisted of 500 independent simulation runs, and correlations were aggregated over all runs. Therefore, the observed effects are a systematic result of the extraction methods and should be seen as corresponding limitations. When using P-splines to extract key points of the stem elongation, the estimated start of the stem elongation may be biased by the base temperature of growth. Nevertheless, this effect presumably applies to any other method including the Percentile method, as both early start and low base temperature may result in a comparable phenotype in early stages.

An increased length of the measurement interval may save considerable time and labor costs which may be invested in larger number of tested genotypes. If aiming to extract timing of key stages, high frequencies are to some extent superfluous if using P-splines, as the spline approach is presumably able to interpolate critical measurement time points. Therefore, one to two measurements a week are sufficient, providing that the total number of measurements does not drop below eight data points (as fitting a shape constrained P-spline using the *scam* package to a time series with less than eight data points becomes to our experience challenging).

4.4. Limitations of processing in stages

A salient feature of our suggested approach is to proceed in several stages, starting with an analysis of time series per plot. Because of this feature, our approach does not explicitly account for gross day-dependent errors operating across all plots, although such errors represent an issue in real field data (Kronenberg et al., 2020b). Explicitly accounting for such errors while also modelling the temporal trajectory would require joint spatio-temporal modelling of the time series across all plots simultaneously. There are several approaches for spatio-temporal modelling of environmental data that could be used here. As we are using splines for modelling both the temporal and the spatial dimension, the most immediate option would be to use three-dimensional tensor spline smoothing (Wood, 2017; Verbyla et al., 2018; Pérez et al., 2020). However, most of these are rather more complex and computationally demanding and as such less suited for a seamless implementation for routine analysis.

5. Conclusion

Processing repeated plot-level measurements using a well-defined process and data model revealed insights on best practice in phenomics data handling. The results confirmed that HTFP measurements allow extracting genotype specific timing of key stages and quantities at defined time points. P-splines combined with the QMER method allowed extracting the timing of key stages and quantities at define time points with a precision that is suitable for, e.g., plant breeding purposes.

Weighting turned out to be essential if processing HTFP data in stages, and linear mixed model analysis was suitable to account for heterogeneity introduced by not considered covariates. Clear restrictions of the proposed data processing strategy became obvious: Correlations between extracted traits cannot only arise from data, but also from the extraction method itself. Therefore, care has to be taken when interpreting such correlations.

Yet, overall, the scientific community dealing with crop phenotyping has not come up with generally accepted procedures how to organize the workflow from raw data generation to extraction of physiologically meaningful results. Hopefully, the herein introduced modeling framework can contribute to achieving this aim; not only for the merit of increased scholarly knowledge generation, but in the interest of a more efficient workflow for crop breeding to improve global nutrition aspects in times of climate change.

Appendix

5.1. Table: Simulation input parameters

Table 2: Model input parameters for the simulation

	Distribution	Values	Sources
θ_i^C	$\mathcal{N}(\mu, \sigma^2)$	T_{min}: $\mu = 8$, $\sigma = 2$	Kemp and Blacklow (1982)
		T_{opt}: $\mu = 18$, $\sigma = 2$	Kemp and Blacklow (1982)
		r_{max}: $\mu = 0.9$, $\sigma = 0.2$	Own data
θ_{jk}^C	$\text{AR(1)}_{\text{row}} \otimes \text{AR(1)}_{\text{range}}$	$\rho = 0.95$, $\sigma_T = \frac{\sigma}{2\sqrt{2}}$	Velazco et al. (2017)
θ_i^T	$\mathcal{N}(\mu, \sigma^2)$	2016: $\mu_{t\text{PH}_{\text{start}}} = 108$, $\sigma_{t\text{PH}_{\text{start}}} = 2.8$	Kronenberg et al. (2020a)
		2017: $\mu_{t\text{PH}_{\text{start}}} = 103$, $\sigma_{t\text{PH}_{\text{start}}} = 3.0$	Kronenberg et al. (2020a)
		2018: $\mu_{t\text{PH}_{\text{start}}} = 101$, $\sigma_{t\text{PH}_{\text{start}}} = 3.1$	Own data
		2016: $\mu_{t\text{PH}_{\text{stop}}} = 165$, $\sigma_{t\text{PH}_{\text{stop}}} = 2.5$	Kronenberg et al. (2020a)
		2017: $\mu_{t\text{PH}_{\text{stop}}} = 162$, $\sigma_{t\text{PH}_{\text{stop}}} = 3.5$	Kronenberg et al. (2020a)
		2018: $\mu_{t\text{PH}_{\text{start}}} = 158$, $\sigma_{t\text{PH}_{\text{stop}}} = 4.0$	Own data
θ_{jk}^T	$\text{AR(1)}_{\text{row}} \otimes \text{AR(1)}_{\text{range}}$	$\rho = 0.95$, $\sigma = \frac{\sigma_{\text{PH}}}{2\sqrt{2}}$	Velazco et al. (2017)
$e_{jkt,1}$	AR(1)_{t}	$\rho = 0.7$, $\sigma_m = 0.01$	Roth et al. (2020)
e_{jk}	$\text{AR(1)}_{\text{row}} \otimes \text{AR(1)}_{\text{range}}$	$\rho = 0.7$, $\sigma = \frac{\sigma_m}{50}$	Assumption
$e_{jkt,2}$	$\mathcal{N}(\mu, \sigma^2)$	$\mu = 0$, $\sigma = \frac{\sigma_m}{100}$	Assumption

5.2. Table: Median bias, variance and root-mean squared errors for the P-spline/QMER method

Table 3: Genotype based bias, variance, and root-mean squared error (RMSE) for the key stages obtained using the P-spline/QMER method, with weighting as option for the second and third stage of the three-stage model, and weighting as option for the second stage of the two-stage model. Results report the median values over the 500 simulated datasets. For sake of completeness, plot-based median values for the P-spline/QMER method are reported as well.

Trait	Model	Weighted? Stage 2	Stage 3	Bias (yday)	Variance (yday^2)	RMSE (yday)
$t\text{PH}_{\text{start}}$	Plot-based			-6.8	57.6	10.3
	Three-stage	no	no	-6.4	10.1	7.49
		no	yes	-6.21	8.68	7.32
		yes	no	-6.27	10	7.54
		yes	yes	-5.97	9.09	7.55
	Two-stage	no	-	-0.417	2.01	4.44
		yes	-	-0.37	2.07	4.41
$t\text{PH}_{\text{stop}}$	Plot-based			-2.85	11.5	4.43
	Three-stage	no	no	-2.52	3.29	5.42
		no	yes	-2.52	3.28	5.46
		yes	no	-2.53	3.38	5.47
		yes	yes	-2.54	3.42	5.49
	Two-stage	no	-	-0.11	2.5	5.6
		yes	-	-0.139	2.58	5.66

5.3. A thought on weighting for traits of the second category (timing of key stages)

Splines can be thought of as polynomials, or other functions that are linear in the regression parameters, pieced together at the knots. Thus, to gain some insight, we here consider a quadratic polynomial as a simple concrete example: $f(t) = \mu + \beta_1 t + \beta_2 t^2$. We observe data $y_i(t) = f(t) + e_i$ $(i = 1, \ldots, n)$, where $e_i \sim \mathcal{N}(0, \sigma^2)$. The model is linear and can be written in general for as $y = X\beta + e$, where $e \sim \mathcal{MVN}(0, I_n\sigma^2)$. Parameters can be estimated by ordinary least squares using $\hat{\beta} = (X^T X)^{-1} X^T y$ with

$$\mathrm{var}(\hat{\beta}) = (X^T X)^{-1}\sigma^2 . \tag{19}$$

A prediction at a particular value of t is obtained from $\hat{y}(t) = \hat{f}(t) = k^T\hat{\beta}$ with $k^T = (1 \quad t \quad t^2)$, and this has variance

$$\mathrm{var}(k^T\hat{\beta}) = k^T(X^T X)^{-1}k\sigma^2 . \tag{20}$$

Now assume that the aim is to find the value of t at which the response $f(t)$ is maximized. For simplicity, we take for granted that a maximum indeed occurs in the relevant range for t. At the maximum, the slope of the curve, i.e. the first derivative equals zero. This can be used to determine the optimal input level: $\frac{\partial f(t)}{\partial t} = \beta_1 + 2\beta_2 t = 0 \Leftrightarrow t_{\mathrm{opt}} = -\frac{\beta_1}{2\beta_2}$. This can be estimated by $\hat{t}_{\mathrm{opt}} = -\frac{\hat{\beta}_1}{2\hat{\beta}_2}$.

Now what can be said about the variance of this estimator, which would be needed for weighting? Here, we may use the delta method (Johnson et al., 1993) to find

$$\mathrm{var}(\hat{t}_{\mathrm{opt}}) \approx \left(\frac{\partial t_{\mathrm{opt}}}{\partial \beta_1}\right)^2_{\beta_1=\hat{\beta}_2} \mathrm{var}(\hat{\beta}_1) + \left(\frac{\partial t_{\mathrm{opt}}}{\partial \beta_2}\right)^2_{\beta_2=\hat{\beta}_2} \mathrm{var}(\hat{\beta}_2) + 2\left(\frac{\partial t_{\mathrm{opt}}}{\partial \beta_1}\right)_{\beta_1=\hat{\beta}_1}\left(\frac{\partial t_{\mathrm{opt}}}{\partial \beta_2}\right)_{\beta_2=\hat{\beta}_2} \mathrm{cov}(\hat{\beta}_1, \hat{\beta}_2). \tag{21}$$

From Equation 19, this is a linear function of σ^2 . Now Equation 20 is also linear in σ^2. This suggests that the weights for $\hat{t}_{\mathrm{opt}}$ will be positively associated with those for $\hat{y}(t_{\mathrm{opt}})$. Exact proportionality cannot be expected, however, because whereas $k^T(X^T X)^{-1}k$ in Equation 20 is constant across plots, the variance in Equation 21 depends on regression parameters that are plot-specific. However, so long as these parameters are not very variable between plots, the association between weights for $\hat{t}_{\mathrm{opt}}$ and $\hat{y}(t_{\mathrm{opt}})$ may be expected to be positive and high.

Acknowledgement

We like to acknowledge Helge Aasen, Lukas Kronenberg and Norbert Kirchgessner (ETH Zurich) for feedback on an early version of the manuscript.

Funding

LR received funding from Innosuisse (http://www.innosuisse.ch) in the framework for the project "Trait spotting" (grant number: KTI P-Nr 27059.2 PFLS-LS). MXRA was supported by project MTM2017-82379-R (AEI/FEDER, UE), by the Basque Government through the BERC 2018-2021 program, and by the Spanish

Ministry of Science, Innovation, and Universities (BCAM Severo Ochoa accreditation SEV-2017-0718). HPP was supported by DFG grant PI 377/24-1.

Declaration of Competing Interest

The authors declare no conflict of interest.

CRediT authorship contribution statement

Lukas Roth: Conceptualization, Methodology, Software, Formal analysis, Visualization, Writing - Original Draft. **María Xosé Rodríguez-Álvarez**: Software, Writing - Review & Editing **Fred van Eeuwijk**: Writing - Review & Editing. **Hans-Peter Piepho**: Conceptualization, Methodology, Writing - Original Draft, Review & Editing. **Andreas Hund**: Conceptualization, Supervision, Project administration, Funding acquisition, Writing - Review & Editing.

References

H. Aasen, N. Kirchgessner, A. Walter, and F. Liebisch. PhenoCams for Field Phenotyping: Using Very High Temporal Resolution Digital Repeated Photography to Investigate Interactions of Growth, Phenology, and Harvest Traits. *Frontiers in Plant Science*, 11(593), 2020. ISSN 1664462X. doi: 10.3389/fpls.2020.00593.

J. Anderegg, K. Yu, H. Aasen, A. Walter, F. Liebisch, and A. Hund. Spectral Vegetation Indices to Track Senescence Dynamics in Diverse Wheat Germplasm. *Frontiers in Plant Science*, 10(1749), 2020. ISSN 1664462X. doi: 10.3389/fpls.2019.01749.

J. L. Araus and J. E. Cairns. Field high-throughput phenotyping: The new crop breeding frontier. *Trends in Plant Science*, 19(1):52–61, 2014. ISSN 13601385. doi: 10.1016/j.tplants.2013.09.008.

J. L. Araus, S. C. Kefauver, M. Zaman-Allah, M. S. Olsen, and J. E. Cairns. Translating High-Throughput Phenotyping into Genetic Gain. *Trends in Plant Science*, 23(5):451–466, 2018. ISSN 13601385. doi: 10.1016/j.tplants.2018.02.001.

J. Blancon, D. Dutartre, M. H. Tixier, M. Weiss, A. Comar, S. Praud, and F. Baret. 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), 2019. ISSN 1664462X. doi: 10.3389/fpls.2019.00685.

M. Borenstein, L. V. Hedges, J. P. Higgings, and H. R. Rothstein. *Introduction to Meta-Analysis*. 2009. ISBN 978-0-470-05724-7.

I. Borra-Serrano, T. De Swaef, P. Quataert, J. Aper, A. Saleem, W. Saeys, B. Somers, I. Roldán-Ruiz, and P. Lootens. Closing the Phenotyping Gap: High Resolution UAV Time Series for Soybean Growth Analysis Provides Objective Data from Field Trials. *Remote Sensing*, 12(1644), 2020. ISSN 20724292. doi: 10.3390/rs12101644.

D. Bustos-Korts, M. P. Boer, M. Malosetti, S. Chapman, K. Chenu, B. Zheng, and F. A. van Eeuwijk. Combining Crop Growth Modeling and Statistical Genetic Modeling to Evaluate Phenotyping Strategies. *Frontiers in Plant Science*, 10(1491), 2019. ISSN 1664462X. doi: 10.3389/fpls.2019.01491.

D. Butler. *asreml: Fits the Linear Mixed Model*, 2018. URL www.vsni.co.uk. R package version 4.1.0.93.

J. N. Cobb, G. DeClerck, A. Greenberg, R. Clark, and S. McCouch. Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype-phenotype relationships and its relevance to crop improvement. *Theoretical and Applied Genetics*, 126: 867–887, 2013. ISSN 00405752. doi: 10.1007/s00122-013-2066-0.

W. G. Cochran. The Combination of Estimates from Different Experiments. *Biometrics*, 10(1):101–129, 1954.

J. A. Fernandez-Gallego, S. C. Kefauver, N. A. Gutiérrez, M. T. Nieto-Taladriz, and J. L. Araus. Wheat ear counting in-field conditions: High throughput and low-cost approach using RGB images. *Plant Methods*, 14(22), 2018. ISSN 17464811. doi: 10.1186/s13007-018-0289-4.

C. Grieder, A. Hund, and A. Walter. Image based phenotyping during winter: a powerful tool to assess wheat genetic variation in growth response to temperature. *Functional Plant Biology*, 42:387–396, 2015. ISSN 1445-4408. doi: 10.1071/fp14226.

M. Herrero-Huerta, P. Rodriguez-Gonzalvez, and K. M. Rainey. Yield prediction by machine learning from UAS-based mulit-sensor data fusion in soybean. *Plant Methods*, 16(78), 2020. ISSN 17464811. doi: 10.1186/s13007-020-00620-6.

P. X. Hurtado, S. K. Schnabel, A. Zaban, M. Veteläinen, E. Virtanen, P. H. Eilers, F. A. van Eeuwijk, R. G. Visser, and C. Maliepaard. Dynamics of senescence-related QTLs in potato. *Euphytica*, 183(3):289–302, 2012. ISSN 00142336. doi: 10.1007/s10681-011-0464-4.

X. Jin, S. Madec, D. Dutartre, B. de Solan, A. Comar, and F. Baret. High-Throughput Measurements of Stem Characteristics to Estimate Ear Density and Above-Ground Biomass. *Plant Phenomics*, 2019(4820305), 2019. doi: 10.34133/2019/4820305.

N. L. Johnson, S. Kotz, and A. Kemp. *Univariate discrete distributions*. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. J. Wiley, New York, 2nd edition, 1993. ISBN 0471548979.

D. R. Kemp and W. M. Blacklow. The responsiveness to temperature of the extension rates of leaves of wheat growing in the field under different levels of nitrogen fertilizer. *Journal of Experimental Botany*, 33(132):29–36, 1982. ISSN 00220957. doi: 10.1093/jxb/33.1.29.

N. Kirchgessner, F. Liebisch, K. Yu, J. Pfeifer, M. Friedli, A. Hund, and A. Walter. The ETH field phenotyping platform FIP: A cable-suspended multi-sensor system. *Functional Plant Biology*, 44:154–168, 2017. doi: 10.1071/FP16165.

L. Kronenberg, K. Yu, A. Walter, and A. Hund. Monitoring the dynamics of wheat stem elongation: genotypes differ at critical stages. *Euphytica*, 213(157), 2017. doi: 10.1007/s10681-017-1940-2.

L. Kronenberg, S. Yates, M. P. Boer, N. Kirchgessner, A. Walter, and A. Hund. Temperature response of wheat affects final height and the timing of stem elongation under field conditions. *Journal of Experimental Botany*, 2020a. doi: 10.1093/jxb/eraa471.

L. Kronenberg, S. Yates, S. Ghiasi, L. Roth, M. Friedli, M. E. Ruckle, R. A. Werner, F. Tschurr, M. Binggeli, N. Buchmann, B. Studer, and A. Walter. Rethinking temperature effects on leaf growth, gene expression and metabolism: Diel variation matters. *Plant, Cell and Environment*, pages 1–15, 2020b. ISSN 0140-7791. doi: 10.1111/pce.13958.

I.-Y. Kwak, C. R. Moore, E. P. Spalding, and K. W. Broman. Mapping quantitative trait loci underlying function-valued traits using functional principal component analysis and multi-trait mapping. *G3: Genes, Genomes, Genetics*, 6:79–86, 2016. ISSN 21601836. doi: 10.1534/g3.115.024133.

C.-X. Ma, C. George, and W. Rongling. Functional mapping of quantitative trait loci underlying the character process: A theoretical framework. *Genetics*, 161:1751–1762, 2002. ISSN 00166731.

M. Maimaitijiang, V. Sagan, P. Sidike, S. Hartling, F. Esposito, and F. B. Fritschi. Soybean yield prediction from UAV using multimodal data fusion and deep learning. *Remote Sensing of Environment*, 237(111599), 2020. ISSN 00344257. doi: 10.1016/j.rse.2019.111599.

M. Malosetti, R. G. Visser, C. Celis-Gamboa, and F. A. van Eeuwijk. QTL methodology for response curves on the basis of non-linear mixed models, with an illustration to senescence in potato. *Theoretical and Applied Genetics*, 113(2):288–300, 2006. ISSN 00405752. doi: 10.1007/s00122-006-0294-2.

E. J. Millet, W. Kruijer, A. Coupel-Ledru, S. Alvarez Prado, L. Cabrera-Bosquet, S. Lacube, A. Charcosset, C. Welcker, F. van Eeuwijk, and F. Tardieu. Genomic prediction of maize yield across European environmental conditions. *Nature Genetics*, 51(6):952–956, 2019. ISSN 15461718. doi: 10.1038/s41588-019-0414-y.

F. F. Moreira, H. R. Oliveira, J. J. Volenec, K. M. Rainey, and L. F. Brito. Integrating High-Throughput Phenotyping and Statistical Genomic Methods to Genetically Improve Longitudinal Traits in Crops. *Frontiers in Plant Science*, 11(681), 2020. ISSN 1664462X. doi: 10.3389/fpls.2020.00681.

B. Parent, E. J. Millet, and F. Tardieu. The use of thermal time in plant studies has a sound theoretical basis provided that confounding effects are avoided. *Journal of Experimental Botany*, 70(9):2359–2370, 2019. ISSN 14602431. doi: 10.1093/jxb/ery402.

D. M. Pérez, M. X. Rodríguez-Álvarez, M. P. Boer, E. J. Millet, and F. A. van Eeuwijk. Spatio-temporal and hierarchical modelling of high-throughput phenotypic data. In *Proceedings of the 35th International Workshop on Statistical Modelling : July 20- 24, 2020 Bilbao, Basque Country, Spain*, pages 394–397, Bilbao, 2020. URL http://hdl.handle.net/10810/45863.

G. Perich, A. Hund, J. Anderegg, L. Roth, M. P. Boer, A. Walter, F. Liebisch, and H. Aasen. Assessment of Multi-Image Unmanned Aerial Vehicle Based High-Throughput Field Phenotyping of Canopy Temperature. *Frontiers in Plant Science*, 11(150), 2020. ISSN 1664-462X. doi: 10.3389/fpls.2020.00150.

H. P. Piepho, J. Möhring, T. Schulz-Streeck, and J. O. Ogutu. A stage-wise approach for the analysis of multi-environment trials. *Biometrical Journal*, 54(6):844–860, 2012. ISSN 15214036. doi: 10.1002/bimj.201100219.

H. Poorter, F. Fiorani, R. Pieruschka, T. Wojciechowski, W. H. van der Putten, M. Kleyer, U. Schurr, and J. Postma. Pampered inside, pestered outside? Differences and similarities between plants growing in controlled conditions and in the field. *New Phytologist*, 212:838–855, 2016. ISSN 14698137. doi: 10.1111/nph.14243.

N. Pya. *scam: Shape Constrained Additive Models*, 2019. URL https://CRAN.R-project.org/package=scam. R package version 1.2-5.

N. Pya and S. N. Wood. Shape constrained additive models. *Statistics and Computing*, 25:543–559, 2015. ISSN 15731375. doi: 10.1007/s11222-013-9448-7.

J. Ramirez-Villegas, J. Watson, and A. J. Challinor. Identifying traits for genotypic adaptation using crop models. *Journal of Experimental Botany*, 66(12):3451–3462, 2015. ISSN 14602431. doi: 10.1093/jxb/erv014.

G. J. Rebetzke, J. Jimenez-Berni, R. A. Fischer, D. M. Deery, and D. J. Smith. Review: High-throughput phenotyping to enhance the use of crop genetic resources. *Plant Science*, 282:40–48, 2019. doi: 10.1016/j.plantsci.2018.06.017.

M. Reymond, B. Muller, A. Leonardi, A. Charcosset, and F. Tardieu. Combining Quantitative Trait Loci Analysis and an Ecophysiological Model to Analyze the Genetic Variability of the Responses of Maize Leaf Growth to Temperature and Water Deficit. *Plant Physiology*, 131:664–675, 2003. ISSN 0032-0889. doi: 10.1104/pp.013839.

M. X. Rodríguez-Álvarez, M. P. Boer, F. A. van Eeuwijk, and P. H. Eilers. Correcting for spatial heterogeneity in plant breeding experiments with P-splines. *Spatial Statistics*, 23:52–71, 2018. ISSN 22116753. doi: 10.1016/j.spasta.2017.10.003.

L. Roth, M. Camenzind, H. Aasen, L. Kronenberg, C. Barendregt, K.-H. Camp, A. Walter, N. Kirchgessner, and A. Hund. Repeated Multiview Imaging for Estimating Seedling Tiller Counts of Wheat Genotypes Using Drones. *Plant Phenomics*, 2020(3729715), 2020. doi: 10.34133/2020/3729715.

P. Sadeghi-Tehran, K. Sabermanesh, N. Virlet, and M. J. Hawkesford. Automated Method to Determine Two Critical Growth Stages of Wheat: Heading and Flowering. *Frontiers in Plant Science*, 8(252), 2017. doi: 10.3389/fpls.2017.00252.

G. A. F. G. A. F. Seber. *Linear regression analysis*. Wiley series in probability and statistics. Wiley-Interscience, Hoboken, N.J., 2nd ed. edition, 2003. ISBN 1-280-58916-7.

A. B. Smith, B. R. Cullis, and R. Thompson. The analysis of crop cultivar breeding and evaluation trials: An overview of current mixed model approaches. *Journal of Agricultural Science*, 143:449–462, 2005. ISSN 00218596. doi: 10.1017/S0021859605005587.

A. Soltani and S. Galeshi. Importance of rapid canopy closure for wheat production in a temperate sub-humid environment: Experimentation and simulation. *Field Crops Research*, 77:17–30, 2002. ISSN 03784290. doi: 10.1016/S0378-4290(02)00045-X.

F. Tardieu, L. Cabrera-Bosquet, T. Pridmore, and M. Bennett. Plant Phenomics, From Sensors to Knowledge. *Current Biology*, 27:R770–R783, 2017. ISSN 09609822. doi: 10.1016/j.cub.2017.05.055.

J. Ubbens, M. Cieslak, P. Prusinkiewicz, and I. Stavness. Latent Space Phenotyping: Automatic Image-Based Phenotyping for Treatment Studies. *Plant Phenomics*, 2020(5801869), 2020. doi: 10.1101/557678.

F. A. van Eeuwijk, D. Bustos-Korts, and M. Malosetti. What should students in plant breeding know about the statistical aspects of genotype x Environment interactions? *Crop Science*, 56(5):2119–2140, 2016. ISSN 14350653. doi: 10.2135/cropsci2015.06.0375.

F. A. van Eeuwijk, D. Bustos-Korts, E. J. Millet, M. P. Boer, W. Kruijer, A. Thompson, M. Malosetti, H. Iwata, R. Quiroz, C. Kuppe, O. Muller, K. N. Blazakis, K. Yu, F. Tardieu, and S. C. Chapman. Modelling strategies for assessing and increasing the effectiveness of new phenotyping techniques in plant breeding. *Plant Science*, 282:23–39, 2019. ISSN 18732259. doi: 10.1016/j.plantsci.2018.06.018.

J. G. Velazco, M. X. Rodríguez-Álvarez, M. P. Boer, D. R. Jordan, P. H. Eilers, M. Malosetti, and F. A. van Eeuwijk. Modelling spatial trends in sorghum breeding field trials using a two-dimensional P-spline mixed model. *Theoretical and Applied Genetics*, 130:1375–1392, 2017. ISSN 00405752. doi: 10.1007/s00122-017-2894-4.

A. P. Verbyla, J. De Faveri, J. D. Wilkie, and T. Lewis. Tensor Cubic Smoothing Splines in Designed Experiments Requiring Residual Modelling. *Journal of Agricultural, Biological, and Environmental Statistics*, 23(4):478–508, 2018. ISSN 15372693. doi: 10.1007/s13253-018-0334-9.

S. Via, R. Gomulkiewicz, G. De Jong, S. M. Scheiner, C. D. Schlichting, and P. H. Van Tienderen. Adaptive phenotypic plasticity: consensus and controversy. *Trends in Ecology & Evolution*, 10(5):212–217, 1995. ISSN 01695347. doi: 10.1016/S0169-5347(00)89061-8.

J. Vos, J. B. Evers, G. H. Buck-Sorlin, B. Andrieu, M. Chelle, and P. H. De Visser. Functional-structural plant modelling: A new versatile tool in crop science. *Journal of Experimental Botany*, 61(8):2101–2115, 2010. ISSN 00220957. doi: 10.1093/jxb/erp345.

C. Welcker, W. Sadok, G. Dignat, M. Renault, S. Salvi, A. Charcosset, and F. Tardieu. A common genetic determinism for sensitivities to soil water deficit and evaporative demand: Meta-analysis of quantitative trait loci and introgression lines of maize. *Plant Physiology*, 157(2):718–729, 2011. ISSN 00320889. doi: 10.1104/pp.111.176479.

A. Whitehead. *Meta-Analysis of Controlled Clinical Trials*. John Wiley & Sons, Ltd, Chichester, 2002. ISBN 0-471-98370-5.

S. N. Wood. *Generalized additive models an introduction with R*. Chapman & Hall/CRC texts in statistical science. CRC Press/Taylor & Francis Group, Boca Raton, second edi edition, 2017. ISBN 1498728332.