Full Bayesian inference in hidden Markov models of plant growth

Gautier Viaud, Yuting Chen, Paul-Henry P.-H. Cournède

Submitted to the Annals of Applied Statistics

FULL BAYESIAN INFERENCE IN HIDDEN MARKOV MODELS OF PLANT GROWTH

By Gautier Viaud[1], Yuting Chen[2], and Paul-Henry Cournède[1],

[1] *Université Paris-Saclay, CentraleSupélec, Mathématiques et Informatique pour la Complexité et les Systèmes, 91190, Gif-sur-Yvette, France*
[2] *Energy Analysis and Environmental Impacts Division, Lawrence Berkeley National Laboratory, United States*

Accurately modeling the growth process of plants in interaction with their environment is important for predicting their biophysical characteristics, referred to as phenotype prediction. Most models are described by discrete dynamic systems in general state-space representation, with important domain-specific characteristics: First, plant model parameters have usually clear functional meanings and may be of genetic origins, thus necessitating a precise estimation. Second, critical growth variables, specifically biomass production and dynamic allocation to organs, are hidden variables, not accessible to measure. Finally, the difficulty to assess the local plant environment may imply the introduction of process noises in models. Therefore, a precise understanding of the system's behavior requires the joint estimation of functional parameters, hidden states, and noise parameters. In this paper, we describe how a full Bayesian method of estimation can accurately estimate all these key model variables using Markov chain Monte Carlo (MCMC) techniques. In the presence of both process and observation noises, it requires to use adequate Particle MCMC (PMCMC) algorithms to efficiently sample the hidden states, which consequently allows for a precise estimation of all noise parameters involved. Thanks to the Bayesian framework, appropriate choices of prior distributions for the noise parameters have enabled analytical posterior distributions and only simple updates are required. Furthermore, this estimation strategy can be easily generalized and adapted to different types of plant growth models, such as organ-scale or compartmental, provided that they are formulated as hidden Markov models. Our estimation method improves on those classically used in plant growth modeling in several aspects: First, by building upon a general probabilistic framework, the estimation results allow proper statistical analyses. It is useful in prediction, for uncertainty and risk analysis (for example for crop yield prediction), but also to analyze the results of experimental trials, for example to compare genotypes in breeding. Moreover, the care taken in the estimation of hidden variables opens new perspectives in the understanding of inner growth processes, notably the balance and in-

MSC 2010 subject classifications: Bayesian inference, hidden Markov model, plant growth, particle MCMC

teraction between biomass production and allocation (referred to as source-sink dynamics). Applications of this estimation procedure are demonstrated on the GreenLab model for *Arabidopsis thaliana* and the Log-Normal Allocation and Senescence (LNAS) model for sugar beet, on both synthetic and real data.

1. Introduction. The ecophysiological processes underlying plant growth are well understood from a biological point of view. Over the last decades, many plant growth models have been proposed in order to predict plant's phenotypes, that is to say their biophysical characteristics under a wide range of varying environments, by describing the main physical processes of the plant's growth (DeJong et al., 2011). Plant growth models find applications in agronomy, horticulture, forestry or ecology, they are used for plant yield prediction (see for example the crop models APSIM (Keating et al., 2003) or DSSAT (Jones et al., 2003)), for the optimization of crop management, notably for irrigation (Wu et al., 2012; Pluchinotta et al., 2018) or fertilization (Lehmann et al., 2013), for the optimization of forestry systems (Pretzsch, Biber and Ďurskỳ, 2002; Fransson et al., 2020), or finally in breeding to characterize genotypes (Yin, Struik and Kropff, 2004; Hammer et al., 2006) or determine ideotypes (Qi et al., 2010; Quilot-Turion et al., 2012).

Most of these models are described by discrete dynamic systems in general state-space representation, with various time scales: usually daily for crops like wheat or corn, but also yearly for trees or hourly for horticultural crops in greenhouses. Mathematically speaking, the plant's phenotypic traits of interest are contained in model outputs

Plant growth models have important domain-specific characteristics: First, model parameters have usually clear functional meanings and may be of genetic origins, thus necessitating a precise estimation. Second, critical growth variables, specifically biomass production and dynamic allocation to organs, are hidden variables, not accessible to measure. Finally, the difficulty to assess the local plant environment may imply the introduction of process noises in models. Therefore, a precise understanding of the system's behavior requires the joint estimation of functional parameters, hidden states, and noise parameters.

However, in plant growth modelling, the methods generally used rely on a classical (generalized) least-square estimation, and it is generally done without a proper modelling of observation and process noises and without a proper handling of estimation uncertainty. Only recent attempts at Bayesian estimation have been made, with crude versions of approximate Bayesian

computation (Jones et al., 2015).

In this paper we present a generic method for the Bayesian inference of plant growth models formulated in a general hidden Markov model framework. Bayesian estimation appears more adapted to the biological case studies under consideration. First, it is more adapted to cases where there are a limited number of observations that are particularly noisy (as is often the case in plant growth applications). Second, models of biological systems usually describe biophysical processes for which parameters have a clear interpretation and prior knowledge can be used to specify prior distributions (Illian, Møller and Waagepetersen, 2009).

Bayesian system identification amounts to estimate certain functional parameters jointly with some hidden states, and possibly the parameters related to the process and observation noises. Hence, several approaches to modelling the system under consideration can be undertaken, and the estimation problem changes accordingly. Our method relies on using particle MCMC (PMCMC) algorithms for a precise joint estimation of functional parameters and hidden states and, subsequently, noise parameters.

We start by introducing in Section 2 the mathematical framework of state space models. In Sections 3 and 4, we introduce the GreenLab model for *Arabidopsis thaliana* and the Log-Normal Allocation and Senescene (LNAS) model for sugar beet as hidden Markov models. In Section 5, we describe how an accurate joint estimation of functional parameters and hidden states can be performed using PMCMC algorithms. Section 6 is devoted to the estimation of observation and process noise parameters and summarizes the whole Bayesian inference procedure for all statistical variables involved. Finally, in Section 7, the overall method is illustrated on the two models considered. First, on a real data set of *A. thaliana* growth using the GreenLab model, which contains observation noises only. Since a wide range of models in plant science do not comprise process noises, it is crucial to ensure that this method is adapted to this subclass of models. Second, the LNAS model for sugar beet is used for the complete estimation of functional parameters, hidden states and noise parameters on synthetic and real data. All the notations used throughout this article are summarized in Section 1of the Supplementary Material (**?**).

2. General state space models. General state space models (GSSMs) were introduced by Kalman (1960) and have been widely used for model calibration and prediction in various fields (Doucet, De Freitas and Gordon, 2001). They describe the time evolution of a dynamic system at each time $t_n = n\Delta t$, between $t_0 = 0$ and $t_N = N\Delta t = T$. Values are indexed with the

time step index n. At each time step, the system is described by a set of state variables that can be considered as a vector of real values $x_n \in \mathbb{R}^{d_x}$. In the terminology of hidden Markov models (HMMs) (Rabiner, 1989), state variables are also known as hidden states, since they are a priori not observable. The initial state of the system is denoted by x_0 and may include variables such as biomasses or surface areas. At time step $n+1$, the state variables of the system x_{n+1} are updated by modelling biological processes at work in the system. This is done by using the state variables $x_n \in \mathbb{R}^{d_x}$ of the previous time step n, environmental variables $u_n \in \mathbb{R}^{d_u}$ and a set of parameters θ intervening in the equations modelling the evolution of the system. Like state variables, environmental variables and parameters can be considered real-valued vectors. Environmental variables play a significant role on the evolution of the system in plant science: temperature, humidity, or radiation heavily influence biological processes such as photosynthesis, evapotranspiration, and so on. Last but not least, process noises can be introduced in order to account for model limitations or imperfections. They are stochastic factors represented at each time step by the realization of a random vector $\eta_n \in \mathbb{R}^{d_\eta}$. The transition from one time step to another can therefore be synthesized in the most generic manner as $x_{n+1} = f_n(x_n, u_n, \theta, \eta_n)$, where f_n is the transition function summarizing all the equations of the system evolution at time step n.

In most real life applications, particularly when dealing with continuous variables, measurements on the system under consideration are not performed exactly. Observation noises are therefore introduced, once again under the form of stochastic factors represented at each time step by the realization of a random vector $\xi_n \in \mathbb{R}^{d_\xi}$. The observations on the system at time step n can hence be summarized as $y_n = g_n(x_n, \theta, \xi_n)$, where g_n is the observation function specifying what variables are observed and how.

Plant growth models are considered within this generic GSSM framework, summarized by two equations, one for the system's transition and the other for its observation. At each time step n, y_n is a vector of observations, that are not necessarily the same at different times.

In their stochastic formulation with random vectors defining the process and observation noises, SSMs are equivalent to HMMs where x_n represents the hidden states, y_n the observations. Thus, $x_0 \sim p(x_0)$ is the initial distribution, $x_{n+1} \sim p(x_{n+1}|\theta, x_n)$ is the transition distribution, and $y_n \sim p(y_n|\theta, x_n)$ is the observation distribution. In the rest of this paper, we consider the most important and general case where both types of noises

are present, i.e.:

$$\begin{cases} x_{n+1} & \sim & p(x_{n+1}|\theta, x_n), \\ y_n & \sim & p(y_n|\theta, x_n). \end{cases}$$

It must be mentioned, however, that models with only observation noises and no process noise (deterministic transition from x_n to x_{n+1}) can be handled in the same generic framework, with the deterministic transition distribution a Dirac delta function (that is to say $p(x_{n+1}|\theta, x_n) = 1$ when x_{n+1} is given by the deterministic function and 0 elsewhere). It is an important case to consider since most classical plant growth models are built this way, from the most classical crop models (Keating et al., 2003; Jones et al., 2003) to functional-structural models like GreenLab (de Reffye et al., 2020). In the next sections we present two examples of plant growth models with the objective to illustrate how they can be written in the HMM framework: the first one, GreenLab, without process noises; the second one, LNAS, with process noises.

3. The GreenLab model for *A. thaliana*. In the context of this work, our interest for *A. thaliana* stems from the large amount of data that high-throughput phenotyping platforms provide. This is the case of the Phenoscope (Tisné et al., 2013) where many plants are grown in a controlled environment (temperature, radiation, hygrometry, etc.). The GreenLab model (de Reffye et al., 2020) is a functional-structural model: it combines the description of plant architectural development and ecophysiological functioning. It has been widely used in the last two decades and calibrated for large varieties of plant species, however mostly with generalized least-square estimators (Cournède et al., 2011). We adapt here a version for *A. thaliana*. More details can be found in (Viaud, Loudet and Cournède, 2017).

Leaves of *A. thaliana* appear in a predefined order, called rank. The leaf of rank v appears at time t^v. The 1st and 2nd leaves first jointly appear at time $t^{12} = t^1 = t^2$. Then the 3rd and 4th leaves jointly appear at time $t^{34} = t^3 = t^4$. For $v \geq 5, t^v = t^{v-1} + \phi$, where ϕ (h) is the phyllochron, i.e. the time-lapse between the appearance of two successive leaves (Wilhelm and McMaster, 1995). The initial biomass for the first two leaves is q_0 and the plant is assumed to grow only during the day, which lasts $n_s = 8$h in our study. The biomass produced at time step n is given by the Beer–Lambert law (Marcelis, Heuvelink and Goudriaan, 1998):

$$q_n = \mu\, r_n\, s \left[1 - \exp\left(-\frac{k}{s\, e} \sum_{v \in [\![1, \nu_n]\!]} q_n^v\right)\right] \tag{1}$$

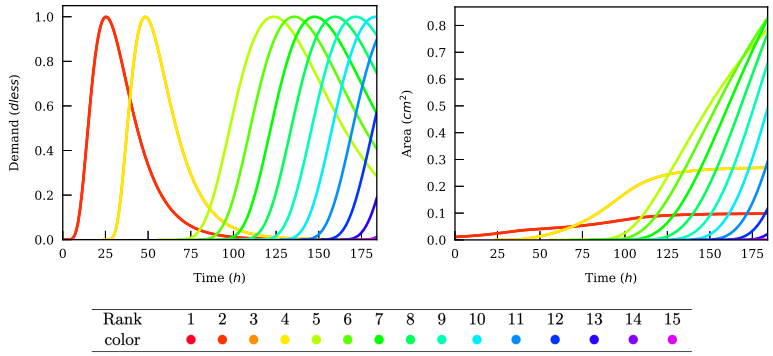

FIG 1. *Left: normalized demand curves d_v for the different leaves. Right: leaf area a_v (cm^2) for the different leaves. Evolution during daily hours (the n-th day corresponds to the $(n_s \times n)$-th hour on the graphs).*

where μ ($\mathrm{g\,MJ}^{-1}$) is the radiation use efficiency, r_n ($\mathrm{MJ\,cm}^{-2}$) the photosynthetically active radiation, s (cm^2) is related to the projected area of the plant, k is the Beer–Lambert law coefficient of light extinction, e ($\mathrm{g\,cm}^{-2}$) the leaf mass per area, ν_n the number of leaves of the plant and q_n^v (g) the biomass of the v-th leaf. The term $r_n\, s \left[1 - \exp\left(-\frac{k}{s\,e} \sum_{v \in [\![1,\nu_n]\!]} q_n^v\right)\right]$ represents the absorbed radiation, it increases with the leaf surface area, but a saturation effect occurs as soon as leaves start to superimpose, as higher leaves cast shade on lower ones.

The biomasses allocated to the different leaves are proportional to their respective demands, which are functions of their thermal time since appearance. An index $k(v) = 1 + \mathbb{1}(v > 4)$ indicates whether a leaf belongs to the first 4 leaves, as leaves of rank $v \leq 4$ and those with $v > 4$ exhibit different dynamics. The demand of the v-th leaf at time step n is:

$$d_n^v = f_{\mathcal{G}}\left(\tau_n^v / \tau_{\mathrm{exp}}^v; \alpha_{k(v)}, \beta_{k(v)}\right). \tag{2}$$

Here, $f_{\mathcal{G}}(x; \alpha, \beta) = \beta^\alpha \Gamma(\alpha)^{-1} x^{\alpha-1} \exp(-\beta x)$ is the pdf of a Gamma distribution parameterized by its shape α and rate β, τ_n^v (°C h) is the accumulated thermal time of the v-th leaf since its emergence and τ_{exp}^v (°C h) is a characteristic thermal time of expansion for the v-th leaf. The 1st and 2nd leaves (resp. the 3rd and 4th leaves) share the same thermal time of expansion τ_{exp}^{12} (resp. τ_{exp}^{34}) and $\tau_{\mathrm{exp}}^v = \tau_{\mathrm{exp}}^5$ for $v \geq 5$. (α_1, β_1) and (α_2, β_2) are the parameters of the Gamma distributions for the preformed leaves and those with rank higher than 5 respectively. The biomass allocated to a leaf is then $\delta q_n^v = d_n^v / (\sum_{w \in [\![1,\nu_n]\!]} d_n^w)^{-1}\, q_n$, which allows to compute the accumulated biomass of each leaf $q_n^v = q_{n-1}^v + \delta q_n^v$ and leaf area $a_n^v = e^{-1} q_n^v$. The observa-

tions associated to the whole growth cycle of a plant is a sequence of vectors of leaf areas. If $V_n \subset [\![1, \nu_n]\!]$ is the set of ranks for which observations on leaf areas are available at time step n, then the transition equation is:

$$x_{n+1} = \begin{bmatrix} q_{n+1} \\ \left[d^v_{n+1}\right]_{v\in[\![1,\nu_{n+1}]\!]} \\ \left[\delta q^v_{n+1}\right]_{v\in[\![1,\nu_{n+1}]\!]} \\ \left[q^v_{n+1}\right]_{v\in[\![1,\nu_{n+1}]\!]} \\ \left[a^v_{n+1}\right]_{v\in[\![1,\nu_{n+1}]\!]} \end{bmatrix} = \begin{bmatrix} r_n\,\mu\,s\left[1-\exp\left(-\frac{k}{s\,e}\sum q^v_n\right)\right] \\ \left[f_{\mathcal{G}}\left(\tau^v_n/\tau^v_{\text{exp}};\alpha_{k(v)},\beta_{k(v)}\right)\right]_{v\in[\![1,\nu_{n+1}]\!]} \\ \left[\left(\sum d^w_{n+1}\right)^{-1} d^v_{n+1} q_{n+1}\right]_{v\in[\![1,\nu_{n+1}]\!]} \\ \left[q^v_n + \delta q^v_{n+1}\right]_{v\in[\![1,\nu_{n+1}]\!]} \\ \left[e^{-1} q^v_{n+1}\right]_{v\in[\![1,\nu_{n+1}]\!]} \end{bmatrix} = f_n(x_n, u_n, \theta)$$

where $x_n = (q_n, [d^v_n]_{1:\nu_n}, [\delta q^v_n]_{1:\nu_n}, [q^v_n]_{1:\nu_n}, [a^v_n]_{1:\nu_n})$ represents the hidden state, $u_n = (r_n, \tau_n)$ the environmental variables and $\theta = (\phi, q_0, \mu, s, e, k, \alpha_1, \beta_1, \alpha_2, \beta_2, t^{12}, t^{34}, [\tau^v_{\text{exp}}]_{1:\nu_N})$ the parameters. The observation equation is:

$$y_n = (\tilde{a}^v_n)_{v\in[\![1,\nu_n]\!]} = (a^v_n \cdot (1+\xi_{v,n}))_{v\in[\![1,\nu_n]\!]} = g_n(x_n, \theta, \xi_n) \tag{3}$$

where $\xi_n \sim \mathcal{N}(0, \sigma^2 I_{\nu_n})$ represents the observation noises, and thus:

$$p(y_{1:N}|\theta, x_{1:N}) = \prod_{n=1}^{N} \prod_{v\in V_n} f_{\mathcal{N}}(\tilde{a}^v_n; a^v_n, \sigma a^v_n). \tag{4}$$

Here, $f_{\mathcal{N}}(x;\mu,\sigma) = (2\pi\sigma^2)^{-1/2}\exp(-(x-\mu)^2/(2\sigma^2))$ is the pdf of a normal distribution. Demands and areas for each leaf are shown in Figure 1.

4. The LNAS model for sugar beet. Although GreenLab belongs to an important family of models in plant science comprising only observation noises, others also include process noises for an increased flexibility. This is the case of the LNAS model for sugar beet (Cournède et al., 2013). Plant organs are not considered individually, as in GreenLab, but as compartments. Two compartments are considered: leaves and roots. On day n, the biomass of the roots is denoted q^r_n and that of the foliage is $q^\ell_n = q^g_n + q^s_n$, where q^g_n is the biomass of green leaves and q^s_n is the biomass of senescent leaves. All biomasses are expressed in $\mathrm{g\,m^{-2}}$. The environmental variables of day n, such as the temperature t_n (°C) and the photosynthetically active radiation r_n ($\mathrm{MJ\,m^{-2}}$), are daily averages.

Plant growth is driven by thermal time $\tau_n = \sum_{i=1}^{n} (t_i - t_b)^+$. At each time step i, if the average temperature t_i is above the base temperature ($t_b = 0\,°\mathrm{C}$ in the case of sugar beet), their difference is accumulated in the thermal time. When τ_n becomes greater than an initiation thermal time τ_{init}, the

plant emerges, starts to intercept light, and produces biomass through photosynthesis. The production equation is again based on the Beer–Lambert law and the deterministic produced biomass on day n is:

$$q_n^{\text{det}} = \mu\, r_n\, (1 - \exp(-k\, q_n^g/e)) \tag{5}$$

where μ ($\mathrm{g\,MJ^{-1}}$) is a radiation use efficiency coefficient, k the Beer–Lambert extinction coefficient, and e ($\mathrm{m^2\,g^{-1}}$) the leaf mass per area (such that q_n^g/e is the leaf area index). There is however an important difference with the GreenLab model: to account for some inaccuracies of the Beer–Lambert law and the difficulty of assessing the local plant environment, the biomass production is rendered stochastic. It is done by multiplying its deterministic value with a multiplicative normal noise such that $q_n^{\text{sto}} = q_n^{\text{det}} \cdot (1 + \eta_n^q)$, where $\eta_n^q \sim \mathcal{N}\left(0, (\sigma_q)^2\right)$. To emphasize the variables on which are set process noises, deterministic variables are denoted with a superscript det whereas their stochastic equivalent are denoted with a superscript sto. The introduction of such stochastic values is of interest to express the transition pdf and compute its value in a generic manner (Viaud, 2018, Chapter 5).

The biomass produced on day n is distributed between the foliage and root system compartments according to an empirical function γ whose deterministic value is given by $\gamma_n^{\text{det}} = \gamma_0 + (\gamma_\ell - \gamma_0)\, F_{\log\mathcal{N}}(\tau_n; \mu_a, \sigma_a)$ where $\gamma^0, \gamma^\ell \in [0, 1]$ are respectively the initial and final proportions of biomass allocated to the leaves, and $F_{\log\mathcal{N}}$ is the cdf of a log-normal distribution parameterized by its median μ and its standard deviation σ:

$$F_{\log\mathcal{N}}(\tau; \mu, \sigma) = \frac{1}{2}\left(1 + \operatorname{erf}\left[\frac{\log(\tau/\mu)}{\sigma\sqrt{2}}\right]\right) \mathbb{1}(\tau \geq 0)\,. \tag{6}$$

A process noise for the allocation is introduced, since this allocation strategy heavily depends on environmental conditions and is known to be rather plastic. Once again, a multiplicative normal noise is chosen and $\gamma_n^{\text{sto}} = \gamma_n^{\text{det}} \cdot (1 + \eta_n^\gamma)$ with $\eta_n^\gamma \sim \mathcal{N}\left(0, (\sigma_\gamma)^2\right)$. The biomass of the whole foliage increases every day: it receives the proportion γ_n^{sto} of the biomass produced, $q_n^\ell = q_{n-1}^\ell + \gamma_n^{\text{sto}}\, q_n^{\text{sto}}$. The biomass of senescent leaves is calculated as a proportion $\rho^s = F_{\log\mathcal{N}}(\tau_n - \tau_s; \mu_s, \sigma_s)$ of the foliage biomass, $q_n^s = \rho_n^s\, q_n^\ell$ This process begins with some delay, once the thermal time has reached a certain threshold τ_s. The biomass of green leaves is therefore $q_n^g = q_n^\ell - q_n^s = (1 - \rho_n^s)\, q_n^\ell$. Finally, the biomass of the roots is increased by what is not allocated to the foliage $q_n^r = q_{n-1}^r + (1 - \gamma_n^{\text{sto}})\, q_n^{\text{sto}}$. It can be noted that both the allocation and the senescence processes are driven by thermal time, hence temperature. The standard formulation of the LNAS transition function eventually reads:

$$x_{n+1} = \begin{bmatrix} q^{\text{det}}_{n+1} \\ q^{\text{sto}}_{n+1} \\ \gamma^{\text{det}}_{n+1} \\ \gamma^{\text{sto}}_{n+1} \\ q^{\ell}_{n+1} \\ q^{g}_{n+1} \\ q^{r}_{n+1} \end{bmatrix} = \begin{bmatrix} r_n\, \mu\, \left(1 - \exp(k\, q^g_n / e)\right) \\ q^{\text{det}}_{n+1} \cdot (1 + \eta^q_n) \\ \left(\gamma^0 + (\gamma^\ell - \gamma^0)\, F_{\log\mathcal{N}}(\tau_n; \mu^a, \sigma^a)\right) \\ \gamma^{\text{det}}_{n+1} \cdot (1 + \eta^\gamma_n) \\ q^\ell_n + \gamma_{n+1}\, q_{n+1} \\ (1 - F_{\log\mathcal{N}}(\tau_n - \tau^s; \mu^s, \sigma^s)) q^\ell_n \\ q^r_n + (1 - \gamma_{n+1})\, q_{n+1} \end{bmatrix} = f_n(x_n, u_n, \theta, \eta_n)$$

where $x_n = (q^{\text{det}}_n, q^{\text{sto}}_n, \gamma^{\text{det}}_n, \gamma^{\text{sto}}_n, q^\ell_n, q^g_n, q^r_n)$ represents the hidden state, $u_n = (r_n, \tau_n)$ the environmental variables, $\theta = (\mu, k, e, \gamma^0, \gamma^\ell, \mu^a, \sigma^a, \tau^s, \mu^s, \sigma^s)$ the functional parameters and $\eta_n = (\eta^q_n, \eta^\gamma_n)$ the process noises parameterized by $(\sigma_q, \sigma_\gamma)$. The transition pdf can be written as:

$$p(x_{n+1}|\theta, x_n) = f_{\mathcal{N}}(q^{\text{sto}}_{n+1}; q^{\text{det}}_{n+1}, \sigma_q q^{\text{det}}_{n+1}) \cdot f_{\mathcal{N}}(\gamma^{\text{sto}}_{n+1}; \gamma^{\text{det}}_{n+1}, \sigma_\gamma \gamma^{\text{det}}_{n+1}). \tag{7}$$

It is assumed that we observe biomasses $\tilde{q}^\Diamond = q^\Diamond_n \cdot (1 + \xi^\Diamond_n)$ at time steps $\mathcal{T}_\Diamond \subset [\![1, T]\!]$, with $\xi^\Diamond_n \sim \mathcal{N}(0, (\sigma_\Diamond)^2)$ for $\Diamond \in \{g, r\}$. The observation pdf is:

$$p(y_n|\theta, x_n) = f_{\mathcal{N}}(\tilde{q}^g_n; q^g_n, \sigma_g q^g_n) \cdot f_{\mathcal{N}}(\tilde{q}^r_n; q^r_n, \sigma_r q^r_n) \tag{8}$$

Figure 2 displays the dynamics of the main variables of this model.

5. Joint estimation of functional parameters and hidden states.

5.1. *Problem description.* Some key variables to understand plant functioning are not easily accessible to experimental measure and are represented

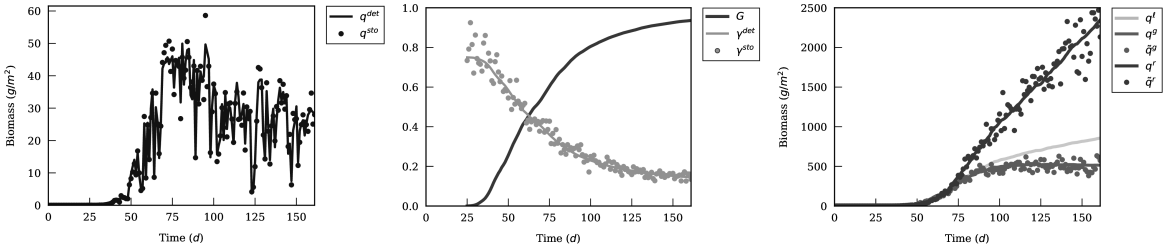

FIG 2. *Left: production of biomass, the continuous line represents the deterministic values predicted by the Beer–Lambert law, and the filled circles the corresponding noised values. Center: allocation of the produced biomass to the different compartments; log-normal cdf* $F_{\log\mathcal{N}(\mu^a, \sigma^a)}$ *(black), allocation variables* γ^{det} *and* γ^{sto} *(gray). Right: biomasses of the different compartments* q^ℓ *(light gray),* q^g *(medium gray),* q^ℓ *(dark gray); hidden states as lines and observed values as filled circles.*

as hidden variables in the state-space representation of plant growth models. It is typically the case for biomass production and biomass partitioning among the different organs which drive the source-sink dynamics, known to be critical for plant performance (White et al., 2016; Smith, Rao and Merchant, 2018). Statistical inference of these variables is thus a critical issue.

In the presence of process noises, the objective is thus to jointly estimate functional parameters θ and hidden states $x_{1:N}$. It is classically done within an MCMC algorithm with target distribution $p(\theta, x_{1:N}|y_{1:N})$ and proposal distribution $q(\theta^*, x^*_{1:N}|\theta^t, x^t_{1:N})$, where variables with a superscript t denote accepted values at the end of MCMC iteration t and variables with a superscript $*$ denote candidate values at iteration $t+1$.

A natural approach is to update the parameters conditionally to the current value of the hidden states and reciprocally, with a Gibbs algorithm for example. However, the efficiency of such a scheme is compromised as soon as there exists a strong correlation between parameters and hidden states as shown by Liu, Wong and Kong (1994) and Roberts and Sahu (1997), which is most often the case in plant growth models. To overcome this issue, Fearnhead (2011) proposed the joint update of the parameters and hidden states with proposal distribution for the candidates $q(\theta^*, x^*_{1:N}|\theta^t, x^t_{1:N}) = q(\theta^*|\theta^t)\, q(x^*_{1:N}|\theta^*)$. Hence, the changes in the parameters can easily be controlled via $q(\theta^*|\theta^t)$, and the candidate hidden states $x^*_{1:N}$ will be consistent with the candidate parameters θ^* thanks to the proposal $q(x^*_{1:N}|\theta^*)$. A classical choice for the proposal distribution is $p(x^*_{1:N}|\theta^*)$, i.e. performing a model simulation. It fails, however, to take into account observations to optimize state space exploration and sample from $p(x^*_{1:N}|\theta^*, y_{1:N})$. PMCMC methods (Andrieu, Doucet and Holenstein, 2010) aim at overcoming this issue. If we managed to sample from $p(x^*_{1:N}|\theta^*, y_{1:N})$, the acceptance probability would be $\alpha = 1 \wedge \frac{p(\theta^*|y_{1:N})}{p(\theta^t|y_{1:N})} \frac{q(\theta^t|\theta^*)}{q(\theta^*|\theta^t)}$, (see Lemma in Web Appendix 5). This ultimately shows that such an MCMC scheme would essentially target the marginal density $p(\theta|y_{1:N})$ which stems from its marginal Metropolis–Hastings (MMH) sampler, as already exploited in (Beaumont, 2003) and (Andrieu and Roberts, 2009). A reasonable approximation to obtaining samples from $p(x_{1:N}|\theta, y_{1:N})$ is to use a sequential Monte Carlo (SMC) algorithm within this MMH sampler, whence the appellation particle marginal Metropolis–Hastings (PMMH).

It must be stressed that this estimation procedure is adapted to models with and without process noise: in the absence of process noise, we can simply revert to the decomposition $q(\theta^*, x^*_{1:N}|\theta^t, x^t_{1:N}) = q(\theta^*|\theta^t)\, p(x^*_{1:N}|\theta^*)$

without the need to use an SMC method, and fall back to a standard MCMC algorithm. It it therefore appropriate for the most common types of plant growth models.

5.2. *Choice of the SMC algorithm.* Within the MMH sampler, the choice of the SMC algorithm is crucial: it must be sufficiently accurate to provide reliable samples from $p(x_{1:N}|\theta, y_{1:N})$ without being too much time-consuming. The SMC algorithm is indeed run at every MCMC iteration, and the whole procedure can become very expensive in terms of computing time and memory, notably with complicated models like plant growth models that usually require a considerable time to simulate.

In order to identify the best compromise in terms of accuracy and numerical efficiency, we tested different SMC algorithms within PMMH, namely the unscented Kalman filter (UKF) (Julier and Uhlmann, 1997), the ensemble Kalman filters (EnKF) (Evensen, 1994), and the regularized particle filters (RPF) (Chen and Cournède, 2014). The different methods are detailed in Web Appendix 6. Despite a very low number of particles for UKF of $2(d_\theta + d_x) + 1$ (i.e. $2{\cdot}(3+2)+1 = 11$ in our benchmarks on the LNAS model), PMMH-UKF manages to obtain very good estimates for both parameters and hidden states and outclassed both PMMH-EnKF and PMMH-RPF (tested with 100 and 1000 particles each) for two parameters out of three and one hidden state out of two. Because of its low computing time and the good estimates it provides, it represents an excellent choice for the joint estimation of parameters and hidden states within a PMMH sampler. This echoes the results of Sherlock, Thiery and Lee (2017) who investigated the performance of PMMH samplers and found that if the computational cost of the algorithm is proportional to the number of particles N of the SMC algorithm, it is often better to set N as low as possible. Doucet et al. (2015) on the other hand provide guidelines to select the optimal number of samples in the SMC step in some particular cases, suggesting that the number of samples could be adapted at each PMMH iteration.

5.3. *Adaptive scheme.* For a better state space exploration and convergence, we use an adaptive scheme derived from (Haario, Saksman and Tamminen, 2001) and described in (Andrieu and Thoms, 2008). At iteration $t+1$, a new candidate is sampled as $\theta^* \sim \mathcal{N}(\theta^t, \lambda^t \Sigma^t)$, where:

$$
(9) \qquad \begin{cases} \mu^t &= \mu^{t-1} + \gamma^t(\theta^t - \mu^{t-1}) \\ \Sigma^t &= \Sigma^{t-1} + \gamma^t\left[(\theta^t - \mu^t)(\theta^t - \mu^t)^T - \Sigma^{t-1}\right] \\ \lambda^t &= \lambda^{t-1}\,\exp\left(\gamma^t\,(\alpha - \alpha^*)\right) \\ \gamma^t &= 1/(t+1) \end{cases}
$$

Algorithm 1 PMMH for the joint estimation of $(\theta, x_{1:N})$

Choose prior and sample $\theta^0 \sim p(\theta)$
Initialize $\gamma^0 = 1$, $\mu^0 = \mathbb{E}[p(\theta)]$, $\Sigma^0 = \mathrm{Cov}[p(\theta)]$, $\lambda^0 = 2.38^2/d_\theta$
Run an SMC algorithm targeting $p(x_{1:N}|\theta^0, y_{1:N})$ and sample $x_{1:N}^0 \sim \hat{p}(x_{1:N}|\theta^0, y_{1:N})$
while $t < M$ and **convergence is not reached do**
  Sample $\theta^* \sim \mathcal{N}(\theta^t, \lambda^t \Sigma^t)$
  Run an SMC algorithm targeting $p(x_{1:N}|\theta^*, y_{1:N})$, sample $x_{1:N}^* \sim \hat{p}(x_{1:N}|\theta^*, y_{1:N})$
  Set $\left(\theta^{t+1}, x_{1:N}^{t+1}\right) = (\theta^*, x_{1:N}^*)$ with probability $\alpha = 1 \wedge \frac{p(\theta^*|y_{1:N})}{p(\theta^t|y_{1:N})} \frac{q(\theta^t|\theta^*)}{q(\theta^*|\theta^t)}$
  Update adaptive scheme variables γ^{t+1}, μ^{t+1}, Σ^{t+1}, λ^{t+1}
end while
Compute $\hat{\theta}$, $\hat{\Sigma}^\theta$ and $\hat{x}_{1:N}$, $\hat{\Sigma}^x$ with burn-in period L

where α^* is an optimal acceptance rate. Convergence is ensured as long as the parameters of the adaptive schemes stay away from poor values (Andrieu and Thoms, 2008). The whole algorithm is detailed in Algorithm 1.

6. Estimation of noise parameters. So far, the observation and process noise parameters were assumed to be known, which is usually not the case. However, the joint estimation of the deterministic and stochastic values for a single state variable at different time steps can be used to estimate the parameters underlying their distributions. We focus on the LNAS model, comprising both observation and process noises, although the following procedure can be easily adapted to models with observation noises only.

State variables carrying process noises are the produced biomass $q_{n+1}^{\mathrm{sto}} = q_{n+1}^{\mathrm{det}} \cdot (1 + \eta_n^q)$, with $\eta_n^q \sim \mathcal{N}\left(0, (\sigma_q)^2\right)$, and the biomass allocation coefficient $\gamma_{n+1}^{\mathrm{sto}} = \gamma_{n+1}^{\mathrm{det}} \cdot (1 + \eta_n^\gamma)$, with $\eta_n^\gamma \sim \mathcal{N}\left(0, (\sigma_\gamma)^2\right)$. State variables affected by observation noises are the biomass of green leaves $\tilde{q}^g = q_n^g \cdot (1 + \xi_n^g)$, with $\xi_n^g \sim \mathcal{N}\left(0, (\sigma_g)^2\right)$, and that of roots $\tilde{q}_n^r = q_n^r \cdot (1 + \xi_n^r)$, with $\xi_n^r \sim \mathcal{N}\left(0, (\sigma_r)^2\right)$. Defining $\boldsymbol{\sigma}^2 = \{\sigma_g^2, \sigma_r^2, \sigma_q^2, \sigma_\gamma^2\}$, the full likelihood is:

$$\begin{aligned}\ell(x_{1:N}, y_{1:N}|\theta, \boldsymbol{\sigma}^2) \quad \propto \quad & \prod_{n \in \mathcal{T}_g} f_{\mathcal{N}}(\tilde{q}_n^g; q_n^g, \sigma_g q_n^g) \prod_{n \in \mathcal{T}_r} f_{\mathcal{N}}(\tilde{q}_n^r; q_n^r, \sigma_r q_n^r) \\ & \prod_{n=1}^{N} f_{\mathcal{N}}(q_n^{\mathrm{sto}}; q_n^{\mathrm{det}}, \sigma_q q_n^{\mathrm{det}}) \prod_{n=1}^{N} f_{\mathcal{N}}(\gamma_n^{\mathrm{sto}}; \gamma_n^{\mathrm{det}}, \sigma_\gamma \gamma_n^{\mathrm{det}})\end{aligned}$$

and the posterior distribution is therefore:

$$\begin{aligned}p(\theta, \boldsymbol{\sigma}^2|x_{1:N}, y_{1:N}) \quad & \propto \quad \ell(x_{1:N}, y_{1:N}|\theta, \boldsymbol{\sigma}^2)\, p(\theta, \boldsymbol{\sigma}^2) \\ & \propto \quad \ell(x_{1:N}, y_{1:N}|\theta, \boldsymbol{\sigma}^2)\, p(\theta)\, p(\sigma_g^2)\, p(\sigma_r^2)\, p(\sigma_q^2)\, p(\sigma_\gamma^2).\end{aligned}$$

If the prior distributions for noise parameters are appropriately chosen, their full conditional distributions can be analytically derived. More precisely, if $\sigma_\Diamond^2 \sim \mathcal{IG}(\alpha_\Diamond, \beta_\Diamond)$ then $\sigma_\Diamond^2|\cdots \sim \mathcal{IG}(\hat{\alpha}_\Diamond, \hat{\beta}_\Diamond)$ for $\Diamond \in \{g, r, q, \gamma\}$, where:

$$
\begin{aligned}
\hat{\alpha}_g &= \alpha_g + \frac{|\mathcal{T}_g|}{2}, & \hat{\beta}_g &= \beta_g + \frac{1}{2}\sum_{n\in\mathcal{T}_g} \frac{(\tilde{q}_n^g - q_n^g)^2}{(q_n^g)^2},\\
\hat{\alpha}_r &= \alpha_r + \frac{|\mathcal{T}_r|}{2}, & \hat{\beta}_r &= \beta_r + \frac{1}{2}\sum_{n\in\mathcal{T}_r} \frac{(\tilde{q}_n^r - q_n^r)^2}{(q_n^r)^2},\\
\hat{\alpha}_q &= \alpha_q + \frac{N}{2}, & \hat{\beta}_q &= \beta_q + \frac{1}{2}\sum_{n=1}^{N} \frac{\left(q_n^{\text{sto}} - q_n^{\text{det}}\right)^2}{(q_n^{\text{det}})^2},\\
\hat{\alpha}_\gamma &= \alpha_\gamma + \frac{N}{2}, & \hat{\beta}_\gamma &= \beta_\gamma + \frac{1}{2}\sum_{n=1}^{N} \frac{\left(\gamma_n^{\text{sto}} - \gamma_n^{\text{det}}\right)^2}{(\gamma_n^{\text{det}})^2}.
\end{aligned}
\tag{10}
$$

The derivation of these formulas are given in Web Appendix 4. Equivalent formulas can be obtained for other types of noise models (additive normal for instance). A strategy for a full Bayesian estimation is therefore to iteratively estimate parameters and hidden states using a PMMH sampler, and update the values of noise parameters based on these estimates as described in Algorithm 2. It covers the generic case of a model with N_o observation noises and N_p process noises parameterized with $[\sigma_{o_i}]_{i\in[\![1,N_o]\!]}$ and $[\sigma_{p_j}]_{j\in[\![1,N_p]\!]}$ respectively (such that, for GreenLab, $N_p = 0$). A first step consists of jointly estimating θ and $x_{1:N}$, after which observation parameters can be estimated. The latter are then used to refine estimates of θ and $x_{1:N}$, and finally, process noise parameters are updated. The whole procedure is done in this order since observation noises have variances typically much greater than those of process noises.

7. Applications. We now illustrate the whole estimation method, first using the GreenLab model for *A. thaliana* (without process noise) on a real data set in Section 7.1 (a case study on synthetic data is also described in Web Appendix 3). We then focus on the LNAS model for sugar beet (with process noise) to investigate the overall performance of our estimation strategy on synthetic data in Section 7.2 before turning to a real case scenario in Section 7.3.

7.1. *GreenLab: real data.* Images of *A. thaliana* were acquired using the Phenoscope (Tisné et al., 2013) which allows to grow many individual plants in a controlled environment. It outputs zenithal images of each plant every day. The data set considered here consists of a series of 21 images for one plant of each of the 4 genotypes Burren (Bur), Columbia (Col), Shahdara (Sha) and Tsushima (Tsu) for a total of 1,043 observations. A manual segmentation performed using a graphics editor (Viaud, Loudet and Cournède, 2017) allowed us to know the area of each leaf every day with a high preci-

Algorithm 2 Overall Bayesian inference

Context: Model with N_o observation noises and N_p process noises parameterized with $[\sigma_{o_i}]_{i\in[\![1,N_o]\!]}$ and $[\sigma_{p_j}]_{j\in[\![1,N_p]\!]}$ respectively

1. Choose priors for functional parameters $\theta^0 \sim p(\theta)$
2. Choose priors for obs. noise parameters $\left(\sigma^0_{o_i}\right)^2 \sim \mathcal{IG}(\alpha_{o_i}, \beta_{o_i})$ for $i \in [\![1, N_o]\!]$

If $N_p \geq 1$:

2'. Choose priors for prc. noise parameters $\left(\sigma^0_{p_j}\right)^2 \sim \mathcal{IG}(\alpha_{p_j}, \beta_{p_j})$ for $j \in [\![1, N_p]\!]$

3. Define $\boldsymbol{\sigma} = \left\{ \left[\sigma^0_{o_i}\right]_{i\in[\![1,N_o]\!]}, \left[\sigma^0_{p_j}\right]_{j\in[\![1,N_p]\!]} \right\}$
4. Estimate $\theta, x_{1:N} | y_{1:N}, \boldsymbol{\sigma}$ using a PMMH-UKF sampler
5. Update $\sigma_{o_i} | \cdots \sim \mathcal{IG}(\hat{\alpha}_{o_i}, \hat{\beta}_{o_i})$ for $i \in [\![1, N_o]\!]$ (using Equation 10 or equivalent)
6. Update $\boldsymbol{\sigma} = \left\{ \left[\hat{\sigma}^0_{o_i}\right]_{i\in[\![1,N_o]\!]}, \left[\sigma^0_{p_j}\right]_{j\in[\![1,N_p]\!]} \right\}$ accordingly

If $N_p \geq 1$:

7. Estimate $\theta, x_{1:N} | y_{1:N}, \boldsymbol{\sigma}$ using a PMMH-UKF sampler
8. Update $\sigma_{p_j} | \cdots \sim \mathcal{IG}(\hat{\alpha}_{p_j}, \hat{\beta}_{p_j})$ for $j \in [\![1, N_p]\!]$ (using Equation 10 or equivalent)

Return: Posterior distributions for parameters θ and states $x_{1:N}$ (obtained at 4. if $N_p \geq 1$ else 7.), obs. noise parameters $[\sigma_{o_i}]_{i\in[\![1,N_o]\!]}$ (obtained at 5.) and possibly prc. noise parameters $[\sigma_{p_j}]_{j\in[\![1,N_p]\!]}$ (obtained at 8.)

sion. For each individual, we jointly estimate functional and noise parameters $\theta = (e, \mu, \phi, \alpha_1, \beta_1, \alpha_2, \beta_2)$ and σ using Algorithm 2 with a simple model simulation instead of a UKF (at Step 4) and $M = 20,000$ MCMC iterations. The formula for the update of σ in this model is given in Web Appendix 2. The mean and standard deviation of the posterior distributions of estimated functional parameters are displayed in Table 1. We used the modelling efficiency (Wallach, 2006) $\mathrm{EF}(x, y) = 1 - (\sum_i (y_i - x_i)^2)/(\sum_i (y_i - \bar{y})^2)$ as a criterion to compare hidden states and observations, i.e. $\mathrm{EF}(\hat{x}_{1:N}, y_{1:N})$. The results are displayed in Table 2. Finally, Figure 3 shows the graphs of the observations $y_{1:N}$ and the estimated hidden states $\hat{x}_{1:N}$ for each individual.

Great precision is achieved in the estimation of leaf areas, as can be seen from Figure 3. Only the 5th leaf area, for Col and Sha, is overestimated

	Gen.	e	μ	ϕ	α_1	β_1	α_2	β_2
Mean	Bur	$1.38{\cdot}10^{-3}$	$2.51{\cdot}10^{+0}$	$1.24{\cdot}10^{+1}$	$1.33{\cdot}10^{+0}$	$2.60{\cdot}10^{+0}$	$3.45{\cdot}10^{+0}$	$4.42{\cdot}10^{+0}$
	Col	$1.60{\cdot}10^{-3}$	$2.85{\cdot}10^{+0}$	$1.50{\cdot}10^{+1}$	$1.45{\cdot}10^{+0}$	$2.32{\cdot}10^{+0}$	$2.66{\cdot}10^{+0}$	$3.29{\cdot}10^{+0}$
	Sha	$2.08{\cdot}10^{-3}$	$3.86{\cdot}10^{+0}$	$1.82{\cdot}10^{+1}$	$1.04{\cdot}10^{+0}$	$2.33{\cdot}10^{+0}$	$2.32{\cdot}10^{+0}$	$2.51{\cdot}10^{+0}$
	Tsu	$1.77{\cdot}10^{-3}$	$3.57{\cdot}10^{+0}$	$1.19{\cdot}10^{+1}$	$1.32{\cdot}10^{+0}$	$1.88{\cdot}10^{+0}$	$3.31{\cdot}10^{+0}$	$4.62{\cdot}10^{+0}$
Std	Bur	$2.49{\cdot}10^{-5}$	$6.37{\cdot}10^{-2}$	$2.62{\cdot}10^{-1}$	$3.57{\cdot}10^{-2}$	$8.67{\cdot}10^{-2}$	$2.36{\cdot}10^{-2}$	$5.13{\cdot}10^{-2}$
	Col	$1.07{\cdot}10^{-5}$	$2.32{\cdot}10^{-2}$	$2.48{\cdot}10^{-1}$	$2.99{\cdot}10^{-2}$	$5.17{\cdot}10^{-2}$	$3.71{\cdot}10^{-2}$	$4.47{\cdot}10^{-2}$
	Sha	$1.22{\cdot}10^{-5}$	$2.73{\cdot}10^{-2}$	$2.07{\cdot}10^{-1}$	$1.12{\cdot}10^{-2}$	$1.36{\cdot}10^{-2}$	$2.06{\cdot}10^{-2}$	$4.12{\cdot}10^{-2}$
	Tsu	$3.35{\cdot}10^{-5}$	$8.66{\cdot}10^{-2}$	$1.08{\cdot}10^{-1}$	$3.40{\cdot}10^{-2}$	$2.97{\cdot}10^{-2}$	$6.03{\cdot}10^{-2}$	$1.64{\cdot}10^{-1}$

TABLE 1

Mean and standard deviation for the posterior distributions of the different estimated parameters of the GreenLab model.

at the end of the growth. This can also be seen from modelling efficiencies in Table 2, for which the 5th leaf is the only one with values below 0.8. Some of the first leaves also exhibit values between 0.8 and 0.9, because our model imposes the same dynamics for the 1st and 2nd leaves on one hand and the 3rd and 4th leaves on the other. Priors for σ were chosen such that $\mathbb{E}[\sigma] = 1 \cdot 10^{-1}$ and $\mathbb{V}[\sigma] = (2\mathbb{E}[\sigma])^2$. The parameters of the inverse Gamma distribution $\sigma^2 \sim \mathcal{IG}(\alpha, \beta)$ were deduced accordingly. The standard deviations of the observation noise were estimated to expected values, all of them being comprised between $6 \cdot 10^{-2}$ and $8 \cdot 10^{-2}$ as detailed in Table 3. Overall, results for functional parameters, hidden states and noise parameters using these 4 real data sets are very satisfactory, even though the modelling of the

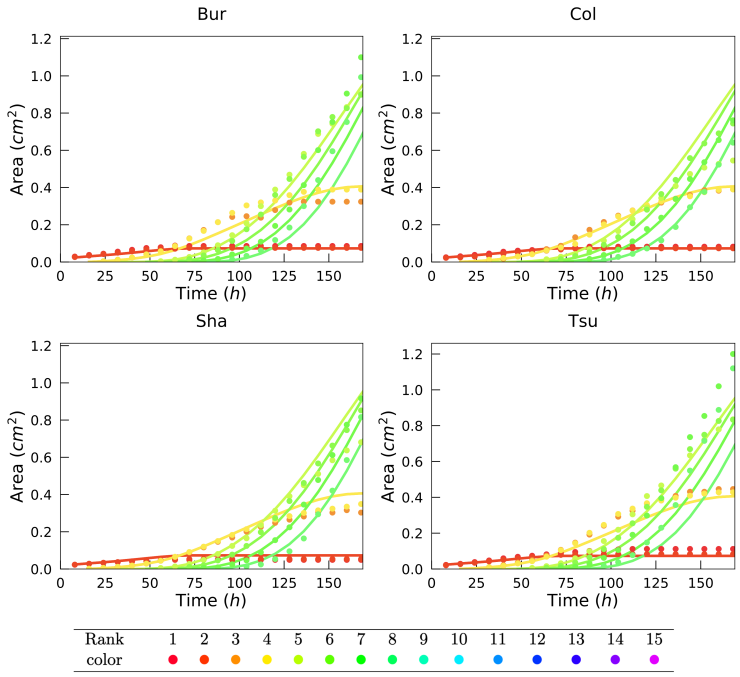

FIG 3. *Areas of the first 8 leaves for genotypes Bur (top left), Col (top right), Sha (bottom left) and Tsu (bottom right); observations (filled circles), and estimated states (lines).*

	Gen.	1	2	3	4	5	6	7	8
EF	Bur	$8.69{\cdot}10^{-1}$	$9.43{\cdot}10^{-1}$	$8.32{\cdot}10^{-1}$	$9.91{\cdot}10^{-1}$	$9.66{\cdot}10^{-1}$	$9.71{\cdot}10^{-1}$	$9.41{\cdot}10^{-1}$	$9.12{\cdot}10^{-1}$
	Col	$8.95{\cdot}10^{-1}$	$8.89{\cdot}10^{-1}$	$9.90{\cdot}10^{-1}$	$9.91{\cdot}10^{-1}$	$5.63{\cdot}10^{-1}$	$9.67{\cdot}10^{-1}$	$9.89{\cdot}10^{-1}$	$9.95{\cdot}10^{-1}$
	Sha	$9.13{\cdot}10^{-1}$	$8.95{\cdot}10^{-1}$	$9.70{\cdot}10^{-1}$	$9.95{\cdot}10^{-1}$	$7.12{\cdot}10^{-1}$	$9.78{\cdot}10^{-1}$	$9.60{\cdot}10^{-1}$	$9.20{\cdot}10^{-1}$
	Tsu	$7.52{\cdot}10^{-1}$	$8.79{\cdot}10^{-1}$	$9.63{\cdot}10^{-1}$	$9.35{\cdot}10^{-1}$	$8.43{\cdot}10^{-1}$	$9.29{\cdot}10^{-1}$	$9.41{\cdot}10^{-1}$	$9.31{\cdot}10^{-1}$

TABLE 2
Modelling efficiency $\mathrm{EF}(\hat{x}_{1:N}, y_{1:N})$ *for the first 8 leaves.*

Gen.	$\mathbb{E}[\sigma]$	α	β	$\mathbb{E}[\sigma\vert\dots]$	$\hat{\alpha}$	$\hat{\beta}$
Bur	$1.00{\cdot}10^{-1}$	$2.07{\cdot}10^{+0}$	$1.07{\cdot}10^{-2}$	$7.13{\cdot}10^{-2}$	$6.71{\cdot}10^{+1}$	$3.36{\cdot}10^{-1}$
Col	$1.00{\cdot}10^{-1}$	$2.07{\cdot}10^{+0}$	$1.07{\cdot}10^{-2}$	$6.08{\cdot}10^{-2}$	$6.51{\cdot}10^{+1}$	$2.36{\cdot}10^{-1}$
Sha	$1.00{\cdot}10^{-1}$	$2.07{\cdot}10^{+0}$	$1.07{\cdot}10^{-2}$	$6.48{\cdot}10^{-2}$	$6.76{\cdot}10^{+1}$	$2.80{\cdot}10^{-1}$
Tsu	$1.00{\cdot}10^{-1}$	$2.07{\cdot}10^{+0}$	$1.07{\cdot}10^{-2}$	$7.89{\cdot}10^{-2}$	$6.71{\cdot}10^{+1}$	$4.11{\cdot}10^{-1}$

TABLE 3
Prior $(\sigma^0, \alpha^0, \beta^0)$ and estimated $(\hat{\sigma}, \hat{\alpha}, \hat{\beta})$ values of the observation noise parameters for the different genotypes.

5th leaf dynamics could probably be improved.

In plant growth models, functional parameters are supposed to be characteristic of genotypes (Yin, Struik and Kropff, 2004; Hammer et al., 2006; Letort et al., 2008). A potential application of our parameterization study is thus to compare the different Arabidopsis genotypes based on the estimated parameters. Note that these genotypes are ecotypes, that is to say variants characteristic of ecological regions: Columbia originates from the USA (MO), Burren from Ireland, Tsushida from Japan, and Shahdara from Tadjikistan. Such detailed comparison is beyond the scope of this paper, but we can give a typical illustration for the parameter e, that is to say the leaf mass per surface area, (also referred to as LMA in ecological studies). It was shown that leaf traits associated with high leaf mass per surface area are characteristic of plants capable to adapt to dry conditions (Wright et al., 2004). In Table 1, we can see that e is indeed the largest for Shahdara, adapted to cold-arid conditions of Pamir-Alay mountains in Tadjikistan, compared to the other ecotypes of more temperate regions. Another known mechanism of adaptation to dry conditions is a longer functioning of leaves (Wright et al., 2004), which is also observed for Shahdara with its flatter sink dynamics (α and β lower values in the Gamma distribution of the demand function 2).

7.2. *LNAS: synthetic data.* The generation of synthetic data for $\tilde{q}^g_{50:10:150}$ and $\tilde{q}^r_{50:10:150}$ is described in Web Appendix 6. We again used Algorithm 2 with $M = 20,000$ iterations, this time with a PMMH-UKF sampler at Step 4, to estimate the 3 most influential parameters of the model (determined using a Sobol sensitivity analysis) $\theta \doteq (\mu, \gamma^0, \mu^a)$, hidden states $x_{1:N} \doteq (q^g_{1:N}, q^r_{1:N}, q^{\text{det}}_{1:N}, q^{\text{sto}}_{1:N}, \gamma^{\text{det}}_{1:N}, \gamma^{\text{sto}}_{1:N})$, and observation and process noise parameters (σ^g, σ^r) and $(\sigma^q, \sigma^\gamma)$. Values used for data simulation are denoted with a superscript $*$. Based on biological knowledge (Damay and Le Gouis, 1993), (Chen, 2014), priors were chosen for functional parameters as $\mu \sim \mathcal{N}(3.6, 0.3^2)$, $\gamma^0 \sim \mathcal{N}(0.8, 0.1^2)$, and $\mu^a \sim \mathcal{N}(600, 50^2)$. For $\Diamond \in \{g, r, q, \gamma\}$, the mean of the prior distributions $\mathbb{E}[\sigma_\Diamond]$ was chosen at least twice higher than $\sigma^*_\Diamond$ for a better state space exploration, the vari-

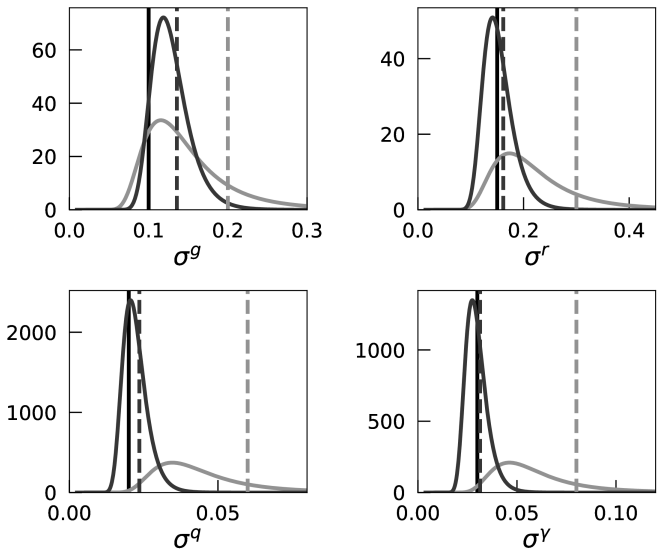

FIG 4. *Estimation of the observation and process noise parameters with the true value (solid black line), the prior distribution (light gray curve), the prior mean (dashed light gray line), the posterior distribution (dark gray curve) and the posterior mean (dashed dark gray line).*

ance was set such that $\mathbb{V}[\sigma_\Diamond] = (2\mathbb{E}[\sigma_\Diamond])^2$ and the parameters $(\alpha_\Diamond, \beta_\Diamond)$ for the inverse Gamma prior were deduced accordingly. Values for each noise are displayed in Figure 4. The results of the first estimation for the observed states q^g and q^r are very satisfactory, with $\mathrm{EF}(\hat{q}^g_{1:N}, q^g_{1:N}) = 0.997$ and $\mathrm{EF}(\hat{q}^r_{1:N}, q^r_{1:N}) = 0.994$. This first PMMH algorithm allows a precise estimation of the standard deviation for the observation noises using Equation 10, as we obtained $\mathbb{E}[\sigma_g|\dots] = 1.36{\cdot}10^{-1}$ and $\mathbb{E}[\sigma_r|\dots] = 1.62{\cdot}10^{-1}$.

These adequate estimates of θ, σ_g and σ_r are used for a second PMMH algorithm aimed at refining the estimates on parameters, hidden states and process standard deviations σ_q and σ_γ. For the second PMMH run, priors for θ were chosen as the posteriors obtained with the first PMMH run, and observation standard deviations were fixed at their estimated values. This two-step strategy was adopted as the estimation of observation noise parameters is accurate and provide values close to the truth for a refined estimation, which is more efficient than updating all noise parameters at once. The joint estimation of θ and $x_{1:N}$ yield precise estimates. In particular, all the hidden states have modelling efficiencies higher than 0.98. This ensures the quality of process noise parameter estimates, inferred from hidden states using Equation 10, ultimately yielding $\mathbb{E}[\sigma_q|\dots] = 2.34{\cdot}10^{-2}$

and $\mathbb{E}[\sigma_\gamma|\dots] = 3.12{\cdot}10^{-2}$. All results for noise parameters are displayed in Figure 4. In conclusion, only σ_g is slightly overestimated, all other noises have distributions whose mean is very close to the true value and with low standard deviations. It is particularly interesting to obtain such results for process noise parameters, since hidden states that are not observable are usually hard to estimate. This was notably allowed thanks to the use of the PMMH sampler for the joint estimation of functional parameters and hidden states as well as the Bayesian update of noise parameters. Overall, we obtained very good results for all model variables involved.

7.3. *LNAS: real data.* We now turn to the case of real data for the LNAS model. The data used for this study were obtained by the French institute for sugar beet research (ITB, Paris, France) in 2010, details of the experimental protocols can be found in (Baey et al., 2014). The biomasses of green leaves and roots, q^g and q^r, were collected on 50 plants at 14 dates such that $\mathcal{T}_g = \mathcal{T}_r = \{54, 68, 76, 83, 90, 98, 104, 110, 118, 125, 132, 139, 145, 160\}$. We estimated the same functional parameters and hidden states as in the case of synthetic data and apply the same estimation strategy; only the number of MCMC iterations was increased to $M = 50,000$ as real data needed more iterations for convergence. For functional parameters, the same priors as in Section 7.2 were used. For noise parameters, we chose $\mathbb{E}[\sigma_g] = \mathbb{E}[\sigma_r] = 2.00{\cdot}10^{-1}$, $\mathbb{E}[\sigma_q] = \mathbb{E}[\sigma_\gamma] = 5.00{\cdot}10^{-2}$ and again $\mathbb{V}[\sigma_\diamondsuit] = (2\mathbb{E}[\sigma_\diamondsuit])^2$ for $\diamondsuit \in \{g, r, q, \gamma\}$. From Figure 5, hidden states seem to be well estimated despite the observation noise. The estimation of noise parameters yields credible values:

$$\begin{array}{ll} \mathbb{E}[\sigma_g|\dots] = 1.29{\cdot}10^{-1}, & \\ \mathbb{E}[\sigma_r|\dots] = 1.16{\cdot}10^{-1}, & \end{array} \quad \text{and} \quad \begin{array}{ll} \mathbb{E}[\sigma_q|\dots] = 1.80{\cdot}10^{-2}, & \\ \mathbb{E}[\sigma_\gamma|\dots] = 2.23{\cdot}10^{-2}. & \end{array}$$

Figures of prior and posterior distributions for noise parameters can be viewed in Web Appendix 8.

Compared to GreenLab, the LNAS model is less detailed from a mechanistic point of view. It is thus less adapted to genotypic analysis. However, it has shown good performances in crop yield prediction in a simpler deterministic version (Baey et al., 2014). The improved statistical estimation proposed here with proper uncertainty evaluation opens new perspectives in risk analysis, while providing the model with more flexibility. This flexibility (allowed by the process noise in production and allocation equations) is also interesting for further biological analysis. The estimated noises at each time step η_n^q and η_n^γ can be analyzed in correlation with environmental variables, notably stresses, to see how they impact production and allocation: For example, if η_n^γ was negative for a significantly long period of time and if we

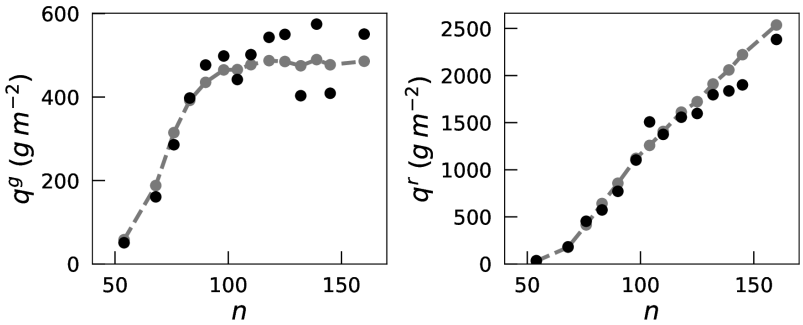

FIG 5. *Estimation of the hidden states related to the observation noises q^g (left) and q^r (right): observations are displayed as filled black circles, and corresponding estimated hidden states as filled gray circles.*

observed a water stress during this period, we could hypothesize that crops adapt to water stress by changing their allocation strategy, and improve the model accordingly. No such correlation was however highlighted in this study.

8. Discussion. We considered two plant growth models within the generic framework of GSSMs, namely GreenLab for *A. thaliana* and LNAS for sugar beet. GreenLab includes observation noises but no process noises, and hidden states are a deterministic function of functional parameters, as is the case of many plant growth models. For its part, LNAS includes both observation and process noises, which makes the estimation of hidden states much more challenging. We proposed a full Bayesian estimation method based on PMCMC methods for the joint estimation of functional parameters, hidden states, and observation and process noise parameters. It is very general since it can handle the two main types of models considered, falling back to a simple model simulation instead of an SMC method within the MCMC algorithm in the absence of process noises. The estimation of noise parameters is based on the Bayesian framework and an adequate choice of prior distributions for a straightforward update of noise parameters. It can handle several noise models, provided that a prior distribution for noise parameters can be chosen for an analytical update of the full conditional distributions.

We illustrated this strategy on several cases. First on a real data set for *A. thaliana* using the GreenLab model, where estimates for functional parameters and observation noise parameters allowed to explain the data for individuals belonging to different genotypes very well. Second, on both synthetic and real data for sugar beet using the LNAS model, with process noise. On synthetic data, we notably showed how the overall procedure could

yield very precise estimates of all model variables, but more importantly hidden states and, as a consequence, process noise parameters.

Great emphasis has been placed on the appropriateness of a full Bayesian approach in the context of plant growth models where important a priori information may often be used. Not only may it be available thanks to expert knowledge, it also proves to yield a flexible statistical framework for the update of noise parameters. To the best of our knowledge, it is the first time that the estimation of hidden states, process and observation noise parameters thanks to PMCMC methods is undertaken.

PMCMC algorithms are of high interest in applications dealing with complex state space models, large amounts of data and with highly multimodal posterior distributions. One of their drawbacks is their numerical cost. Some extensions have recently been proposed, such as interacting PMCMC algorithms (Mingas, Bottolo and Bouganis, 2017) or by introducing new latent variables (Fearnhead and Meligkotsidou, 2016) for enhanced mixing rates. Their parallelization on CPU, GPU or field programmable gate arrays (FPGAs) is a topic of primary importance given their computing cost.

For complex models whose simulation is costly and with many parameters to estimate, Hamiltonian Monte Carlo (HMC) methods (Neal, 2011), (Hoffman and Gelman, 2014) could be used for an enhanced exploration of the state space. Other strategies based on meta-modelling like Gaussian process approximation of simulations (Overstall and Woods, 2013) could allow computationally efficient MCMC inference and could also be tested.

The full Bayesian estimation procedure described here could also be adapted to hierarchical models for populations of plants. While a topic of major importance, few works have been undertaken in the plant community (Illian, Møller and Waagepetersen, 2009).

References.

Andrieu, C., Doucet, A. and Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society* **72** 269-342.

Andrieu, C. and Roberts, G. O. (2009). The Pseudo-Marginal Approach for Efficient Monte Carlo Computations. *The Annals of Statistics* **37** 697-725.

Andrieu, C. and Thoms, J. (2008). A tutorial on adaptive MCMC. *Statistics and Computing* **18** 343–373.

Baey, C., Didier, A., Lemaire, S., Maupas, F. and Cournède, P.-H. (2014). Parametrization of five classical plant growth models applied to sugar beet and comparison of their predictive capacity on root yield and total biomass. *Ecological Modelling* **290** 11 - 20.

Beaumont, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics* **164** 1139–1160.

Chen, Y. (2014). Bayesian inference in plant growth models for prediction and uncertainty assessment, PhD thesis, École Centrale Paris.

Chen, Y. and Cournède, P. (2014). Data assimilation to reduce uncertainty of crop model prediction with convolution particle filtering. *Ecological Modelling* **290** 165-177.

Cournède, P.-H., Letort, V., Mathieu, A., Kang, M. Z., Lemaire, S., Trevezas, S., Houllier, F. and De Reffye, P. (2011). Some parameter estimation issues in functional-structural plant modelling. *Mathematical Modelling of Natural Phenomena* **6** 133–159.

Damay, N. and Le Gouis, J. (1993). Radiation use efficiency of sugar beet in Northern France. *European Journal of Agronomy* **2** 179 - 184.

de Reffye, P., Hu, B., Kang, M., Letort, V. and Jaeger, M. (2020). Two decades of research with the GreenLab model in Agronomy. *Annals of Botany*.

DeJong, T. M., Da Silva, D., Vos, J. and Escobar-Gutiérrez, A. J. (2011). Using functional–structural plant models to study, understand and integrate plant development and ecophysiology. *Annals of botany* **108** 987–989.

Doucet, A., De Freitas, N. and Gordon, N. (2001). *Sequential Monte Carlo methods in practice*. Springer-Verlag New York, New York.

Doucet, A., Pitt, M. K., Deligiannidis, G. and Kohn, R. (2015). Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika* **102** 295–313.

Cournède, P. H., Chen, Y., Wu, Q., Baey, C. and Bayol, B. (2013). Development and Evaluation of Plant Growth Models: Methodology and Implementation in the PYGMALION platform. *Mathematical Modelling of Natural Phenomena* **8** 112–130.

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research: Oceans* **99** 10143–10162.

Fearnhead, P. (2011). MCMC for State-Space Models. In *Handbook of Markov chain Monte Carlo* (S. Brooks, A. Gelman, G. Jones and X.-L. Meng, eds.) 513–529. CRC.

Fearnhead, P. and Meligkotsidou, L. (2016). Augmentation schemes for particle MCMC. *Statistics and Computing* **26** 1293–1306.

Fransson, P., Franklin, O., Lindroos, O., Nilsson, U. and Bränström, Å. (2020). A simulation-based approach to a near-optimal thinning strategy: allowing harvesting times to be determined for individual trees. *Can. J. of Forest Research* **50** 320–331.

Haario, H., Saksman, E. and Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli* 223–242.

Hammer, G., Cooper, M., Tardieu, F., Welch, S., Walsh, B., Van Eeuwijk, F., Chapman, S. and Podlich, D. (2006). Models for navigating biological complexity in breeding improved crop plants. *Trends in Plant Science* **11** 587–593.

Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn Sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* **15** 1593–1623.

Illian, J. B., Møller, J. and Waagepetersen, R. P. (2009). Hierarchical spatial point process analysis for a plant community with high biodiversity. *J. Env. and Ecol. Stat.* **16** 389–405.

Jones, J. W., Hoogenboom, G., Porter, C. H., Boote, K. J., Batchelor, W. D., Hunt, L., Wilkens, P. W., Singh, U., Gijsman, A. J. and Ritchie, J. T. (2003). The DSSAT cropping system model. *European journal of agronomy* **18** 235–265.

Jones, J. W., He, J., Boote, K. J., Wilkens, P., Porter, C. H. and Hu, Z. (2015). *Estimating DSSAT Cropping System Cultivar-Specific Parameters Using Bayesian Techniques* In *Methods of Introducing System Models into Agricultural Research* 13, 365-393. John Wiley & Sons, Ltd.

Julier, S. J. and Uhlmann, J. K. (1997). New extension of the Kalman filter to nonlinear

systems. *SPIE Proceedings* **3068** 182–193.

Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME – Journal of Basic Engineering* **82** 35–45.

Keating, B. A., Carberry, P. S., Hammer, G. L., Probert, M. E., Robertson, M. J., Holzworth, D., Huth, N. I., Hargreaves, J. N., Meinke, H., Hochman, Z. et al. (2003). An overview of APSIM, a model designed for farming systems simulation. *European journal of agronomy* **18** 267–288.

Lehmann, N., Finger, R., Klein, T., Calanca, P. and Walter, A. (2013). Adapting crop management practices to climate change: Modeling optimal solutions at the field scale. *Agricultural Systems* **117** 55–65.

Letort, V., Mahe, P., Cournède, P. H., De Reffye, P. and Courtois, B. (2008). Quantitative genetics and functional-structural plant growth models: simulation of quantitative trait loci detection for model parameters and application to potential yield optimization. *Annals of Botany* **101** 1243–1254.

Liu, J., Wong, W. and Kong, A. (1994). Covariance structure and convergence rate of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* **81** 27–40.

Marcelis, L., Heuvelink, E. and Goudriaan, J. (1998). Modelling of biomass production and yield of horticultural crops: a review. *Scientia Horticulturae* **74** 83–111.

Mingas, G., Bottolo, L. and Bouganis, C.-S. (2017). Particle MCMC algorithms and architectures for accelerating inference in state-space models. *International Journal of Approximate Reasoning* **83** 413 - 433.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In *Handbook of Markov chain Monte Carlo* (S. Brooks, A. Gelman, G. Jones and X.-L. Meng, eds.) 2. CRC press.

Overstall, A. M. and Woods, D. C. (2013). A Strategy for Bayesian Inference for Computationally Expensive Models with Application to the Estimation of Stem Cell Properties. *Biometrics* **69** 458-468.

Pluchinotta, I., Pagano, A., Giordano, R. and Tsoukiàs, A. (2018). A system dynamics model for supporting decision-makers in irrigation water management. *Journal of environmental management* **223** 815–824.

Pretzsch, H., Biber, P. and Ďurskỳ, J. (2002). The single tree-based stand simulator SILVA: construction, application and evaluation. *Forest Ecol. and Man.* **162** 3–21.

Qi, R., Ma, Y., Hu, B., De Reffye, P. and Cournède, P. H. (2010). Optimization of source-sink dynamics in plant growth for ideotype breeding: a case study on maize. *Computers and Electronics in Agriculture* **71** 96–105.

Quilot-Turion, B., Ould-Sidi, M.-M., Kadrani, A., Hilgert, N., Génard, M. and Lescourret, F. (2012). Optimization of parameters of the Virtual Fruit model to design peach genotype for sustainable production systems. *European Journal of Agronomy* **42** 34–48.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* **77** 257–286.

Roberts, G. O. and Sahu, S. K. (1997). Updating Schemes, Correlation Structure, Blocking and Parameterization for the Gibbs Sampler. *J. Royal Stat. Soc.: Series B* **59** 291–317.

Sherlock, C., Thiery, A. H. and Lee, A. (2017). Pseudo-marginal Metropolis–Hastings sampling using averages of unbiased estimators. *Biometrika* **104** 727-734.

Smith, M. R., Rao, I. M. and Merchant, A. (2018). Source-sink relationships in crop plants and their influence on yield development and nutritional quality. *Frontiers in Plant Science* **9** 1889.

Tisné, S., Serrand, Y., Bach, L., Gilbault, E., Ben Ameur, R., Balasse, H.,

Voisin, R., Bouchez, D., Guerche, P., Chareyron, G., Da Rugna, J., Camilleri, C. and Loudet, O. (2013). Phenoscope: an automated large-scale phenotyping platform offering high spatial homogeneity. *The Plant Journal* **74** 534–544.

Viaud, G. (2018). Statistical methods for the genotypic differentiation of plants using growth models, PhD thesis, Université Paris-Saclay.

Viaud, G., Loudet, O. and Cournède, P.-H. (2017). Leaf Segmentation and Tracking in Arabidopsis thaliana Combined to an Organ-Scale Plant Model for Genotypic Differentiation. *Frontiers in Plant Science* **7** 2057.

Wallach, D. (2006). *Evaluating crop models* In *Working with Dynamic Crop Models: Evaluation, Analysis, Parameterization, and Applications* 2, 11–54. Elsevier.

White, A. C., Rogers, A., Rees, M. and Osborne, C. P. (2016). How can we make plants grow faster? A source–sink perspective on growth rate. *Journal of Experimental Botany* **67** 31–45.

Wilhelm, W. and McMaster, G. S. (1995). Importance of the phyllochron in studying development and growth in grasses. *Crop Science* **35** 1–3.

Wright, I. J., Reich, P. B., Westoby, M., Ackerly, D. D., Baruch, Z., Bongers, F., Cavender-Bares, J., Chapin, T., Cornelissen, J. H., Diemer, M. et al. (2004). The worldwide leaf economics spectrum. *Nature* **428** 821–827.

Wu, L., Le Dimet, F.-X., De Reffye, P., Hu, B.-G., Cournède, P.-H. and Kang, M.-Z. (2012). An optimal control methodology for plant growth—Case study of a water supply problem of sunflower. *Mathematics and computers in simulation* **82** 909–923.

Yin, X., Struik, P. C. and Kropff, M. J. (2004). Role of crop physiology in predicting gene-to-phenotype relationships. *Trends in Plant Science* **9** 426-432.

Address of the First and Third authors,
E-mail: gautier.viaud@centralesupelec.fr
E-mail: paul-henry.cournede@centralesupelec.fr

Address of the Second author,
E-mail: yutingchen@lbl.gov