# Iterative convolution particle filtering for nonlinear parameter estimation and data assimilation with application to crop yield prediction

Yuting CHEN * Samis TREVEZAS * Paul-Henry COURNEDE *

**Abstract**

The complexity of plant growth models and the scarcity of experimental data make the application of conventional data assimilation techniques rather difficult. In this paper, we use the Convolution Particle Filter (CPF) and an iterative adaptation, the Iterative Convolution Particle Filter (ICPF) for nonlinear parameter estimation. Both methods provide prior distributions in the Bayesian framework for data assimilation. CPF is sequentially used to update state and parameter estimates in order to improve model prediction and to assess the predictive uncertainty. The predictive performances of the two methods are evaluated by an application to the LNAS sugar beet growth model with three sets of real measurements, one used for parameter estimation and the two others used to test the model predictive capacity, both with and without data assimilation. Despite the low accuracy and the scarcity of the early data used for assimilation, the CPF-based data assimilation approach with the prior distribution based on ICPF estimations showed promising predictive capacities and provided robust confidence intervals. The method can therefore be considered as a potential candidate for yield prediction applications in agriculture.

## 1 Introduction

Given a dynamic model and corresponding experimental observations, data assimilation techniques can be used to update state and parameter estimates during the system's evolution through time in order to improve model prediction.

Dynamic crop models are generally characterized by complex interacting processes and a large number of model parameters to describe plant growth and development. When combined to data assimilation methods, they can be considered as useful tools for yield prediction and decision support [15]. However, as a specificity of agricultural systems, experimental data acquisition tends to be costly (when direct field data collections are involved) or inaccurate (for indirect measurements such as satellite images), and generally irregular. Therefore, due to the nonlinear dynamics of the system equations, the restricted experimental data and the considerable uncertainty of the inputs, the parameterization of these models is generally regarded as a key issue which may affect strongly the quality of model prediction. Efficient and precise parameter estimation with proper assessment of the parameter uncertainty is thus crucial to assure reliable and satisfactory prediction performance.

In the past decade, data assimilation problems have been commonly reformulated and studied with a Bayesian probabilistic perspective, which allows the simultaneous estimation of model states and parameters ([16], [17], [25]) in the framework of generalized state-space models. However, the most frequently used approaches still rely on Kalman filter, such as the Extended Kalman Filter (EKF) [12] or the Ensemble Kalman Filter (EnKF) [13]. In the EKF approach, the idea is to approximate locally the nonlinear system dynamics in order to infer recursively the state variables, yet the local approximations may cause divergence [12]. As for the EnKF approach, the main constraint is the Gaussian assumption of all the involved probability distributions [24]. These approximations often lead to critical filter divergence when the experimental data are not frequent enough. Another alternative that can be used for Baysian inference is the Sequential Importance Sampling, also known as Particle Filtering ([14], [18]). Unlike the Kalman filter related approaches, the advantage of Monte-Carlo filtering methods has been well established since they aim at providing better approximation of the exact posterior distributions for data assimilation in nonlinear systems. However, their main drawbacks are potential sample degeneracy problems [1], which according to [10] could be alleviated by increasing the number of particles or resampling. Nevertheless, the resampling procedure conducted by inappropriate process noise can lead to sample impoverishment [14].

In this paper, we choose to adapt a new Bayesian assimilation approach, the Convolution Particle Filtering (CPF) [4], for sequential data assimilation, which allows to circumvent the potential side effects of resampling and proves to be very robust and easy to apply to plant growth models. With the purpose of improving the parametrization performance especially with stochastic

[1]Digiplante, Applied Mathematics and Systems Laboratory, Ecole Centrale Paris, 92295 Châtenay-Malabry, France

models and in the case of rare or irregular experimental data, a conditional iterative version of the Convolution Particle Filtering (ICPF) has as well been proposed [7].

The objective of this paper is to explore CPF-based parameter estimation methods with an emphasis on the sequential data assimilation for a plant growth model with constrained observations. An illustrative application is provided using the LNAS model of sugar beet growth [9] and three years of experimental data obtained in comparable but different situations. One dataset is used for parameter estimation and the two others are used to test model prediction, with assimilation of the early experimental measurements. The properties and performances of the CPF and ICPF estimation methods are investigated, specifically regarding their abilities to provide reliable a priori estimates for data assimilation and prediction purposes.

In Section 2, the principles of Convolution Particle Filtering are recalled and an iterative version for parameter estimation is introduced. In Section 3, we detail the test case by describing the LNAS model, the experimental datasets, the calibration and assimilation procedures. The results are presented and discussed in Section 4 before drawing the conclusions.

## 2 Methods

Plant growth models or crop models are generally written in a state-space form and considered as deterministic. In rarer cases, modelling and measurement noises are introduced to build a stochastic model. In both situations, they can be used by the statistical framework of sequential data assimilation aiming to estimate the time evolving state variables. The major challenge for yield prediction is to develop an appropriate data assimilation method to overcome the uneven and irregular data problems and thus to integrate the available information to a complex nonlinear dynamic model.

In this paper, we consider cases in which no satisfying prior distributions are available for the considered parameters, so that a first calibration step is performed on a full dataset to provide the prior distributions. Afterwards, the obtained prior distributions are used for data assimilation configurations during which the early experimentation data are assimilated to improve model prediction for the measurements of later dates.

The proposed approach consists of three steps. In the first selection step, the least influential model parameters are screened using sensitivity analysis methods [5] and are thus fixed. In the second step, the other parameters are estimated with the CPF approach based on a given experimental dataset. The posterior distributions obtained can thus be used to assess the uncertainty of the estimates. Meanwhile, the modelling and measurement noises can as well be evaluated. An alternative approach that we propose for parameter estimation relies on the iterative CPF approach. The CPF is iterated until convergence using the posterior distribution of iteration $k$ as prior distribution for iteration $k+1$. However, the obtained final distributions are no longer representative of the uncertainty of the estimates, since they are influenced by the regularization effect and the empirical estimation of the modelling and measurement noises. Under these circumstances, a parametric bootstrap can be carried out to assess the uncertainty of the estimates. During the third assimilation phase, a new comparable experimental dataset (for instance same crop, but from a different year, or at different location, or of different genotype ...) with few early measurements is used. The CPF approach is anew implemented this time with the prior distributions provided by the previous calibration step. After a short recalibration period while model parameters and state variables are updated based on the available measurements, the particles (simulations) continue to propagate in order to predict the system evolution and to evaluate the related uncertainty.

In this section, we first recall the general state-space model framework used for parameter estimation and data assimilation in this study. Then the Convolution Particle Filter is presented followed by its iterative version that we proposed for parameter estimation.

**2.1 General State-Space Models:** The statistical framework for crop data assimilation is provided by the discrete nonlinear general state-space model with a state function and an observation function:

$$\begin{cases} X(t+1) = f\left(X(t), \Theta, \eta(t), t\right) \\ Y(t) = g\left(X(t), \Theta, \xi(t), t\right) \end{cases} \tag{2.1}$$

The evolution equation is embodied in the function $f$, which is time dependent. $X(t)$ represents the state variables at time $t$, $\Theta$ is a vector of parameters of dimension $p$ and the modelling noise is represented with the random variables $\eta(t)$ (corresponding to model imperfections or uncertainty in the model inputs). The observation equation incorporates observations on the state variables of interest. $Y(t)$ is the outcome vector which is related to the state variable vector $X(t)$ through the function $g$. $Y(t)$ consists of state variables that can be observed experimentally and usually differ from $X(t)$ (for instance, biomasses of some plant organs can be measured while the daily biomass production cannot). Measurement noises are denoted by $\xi(t)$. $(\eta(t))_t$ and $(\xi(t))_t$ are considered as sequences of independent and identically distributed random variables. Since experimental observations are usually limited due to high

costs, observations are only available at irregular times. Let $(t_1, t_2, ...t_N)$ be the $N$ measurement time steps. For all $n \in [1; N]$, we set: $X_n := X(t_n)$, $Y_n := Y(t_n)$ and $Y_{1:n} := (Y(t_1), Y(t_2), \ldots, Y(t_n))$.

**2.2 Convolution Particle Filter for Bayesian Parameter Estimation:** Particle filtering is a recursive Bayesian filter based on Monte Carlo simulations [1]. The basic idea is the recursive approximation of the filtering distribution by a time evolving weighted sample.

Inspired by the Post-Regularized Particle Filter [23], the objective of the Convolution Particle Filter [4] is to estimate jointly the parameters and the hidden states of the dynamic system by processing the data online. A possible way to incorporate the vector parameter $\Theta$ in the state equation is by considering $\Theta_n$ with a constant evolution. An augmented state vector $X_n^a = (X_n, \Theta_n)$ is thus defined which contains $X_n$ the true hidden state at time $t_n$ and $\Theta_n$ the vector of unknown parameters. In the following, if $X$ represents a random variable with values in $\mathcal{X}$, then for all $x \in \mathcal{X}$, $p(x)$ will denote the probability density of $X$ in $x$. The first-order hidden Markov model is characterized by the transition density $p(x_n^a|x_{n-1}^a)$ corresponding to the state equation, the observation density $p(y_n|x_n^a)$ corresponding to the observation equation and the initial density $p(x_0^a)$.

In the initialization step of our implementation, the parameters are initialized from either informative distributions ($p(x_0^a)$) or non-informative distributions for all the particles. Particle weights are assigned uniformly. Each filtering step is performed recurrently in two stages and occurs only at time steps when the observation is available [4]:

**Prediction**: The objective is to provide a kernel estimator of $p(x_{n+1}^a, y_{n+1}|y_{0:n})$ denoted by $\hat{p}(x_{n+1}^a, y_{n+1}|y_{0:n})$. $M$ particles $\{\tilde{x}_n^{a\,(i)}, i = 1, \ldots, M\}$ are sampled from the distribution with conditional density $\hat{p}(x_n^a|y_{0:n})$. The $M$ particles are propagated through the evolution model until the next available measurement to obtain the predicted states $\{\tilde{x}_{n+1^-}^{a\,(i)}, i = 1, \ldots, M\}$ . The updating scheme relies directly on Bayes' law. The particle weights are calculated based on the experimental measurements and the predictions, and then normalized. The empirical kernel approximation of the probability density of $(X_{n+1}^a, Y_{n+1})$ conditional to $Y_{0:n}$ can thus be deduced using the Parzen-Rosenblatt kernel $K^X_{h_M^X}$, with bandwidth parameter $h_M^X$:

$$\hat{p}(x_{n+1}^a, y_{n+1}|y_{0:n}) = \frac{1}{M}\sum_{i=1}^{M} K^X_{h_M^X}\left(x_{n+1}^a - \tilde{x}_{n+1^-}^{a\,(i)}\right) \cdot p\left(y_{n+1}|\tilde{x}_{n+1^-}^{a\,(i)}\right). \tag{2.2}$$

**Correction**: The *a posteriori* form of the estimation is deduced from Bayes' law and the kernel approximation for $p(x_{n+1}^a|y_{1:n+1})$ is given by:

$$\hat{p}(x_{n+1}^a|y_{1:n+1}) = \frac{1}{\sum_{i=1}^{M} p(y_{n+1}|\tilde{x}_{n+1^-}^{a\,(i)})} \cdot \sum_{i=1}^{M} K^X_{h_M^X}(x_{n+1}^a - \tilde{x}_{n+1^-}^{a\,(i)}) p(y_{n+1}|\tilde{x}_{n+1^-}^{a\,(i)}). \tag{2.3}$$

The part $p(y_{n+1}|\tilde{x}_{n+1^-}^{a\,(i)})/\sum_{i=1}^{M} p(y_{n+1}|\tilde{x}_{n+1^-}^{a\,(i)})$ can be considered as the normalized weight $\tilde{w}_{n+1}^{(i)}$ associated to the particle $\tilde{x}_{n+1^-}^{a\,(i)}$. In the case that the analytic form of the observation density $p(y_n|x_n)$ is unknown, an observation kernel can similarly be introduced [4]. The new set of particles $\left\{x_{n+1}^{a\,(i)},\ 1 \le i \le M\right\}$ are then sampled from $\hat{p}(x_{n+1}^a|y_{1:n+1})$.

**2.3 Iterative Convolution Particle Filter:** In the case of off-line estimation with a finite number of observations, in order to determine the prior distributions for data assimilation, an iterative version of CPF can hence be applied. At iteration $k$, the particles $x_0^{a(i)}$ are obtained as follows: the initial state vectors $\{\tilde{x}_0{}^{(i)}, i = 1, \ldots, M\}$ are selected in the same way as for the classical filtering process (sampled from $p(x_0)$), and the vectors of unknown parameters $\{\tilde{\Theta}_0^{(i)}, i = 1, \ldots, M\}$ are sampled from the multivariate Gaussian distribution defined by the mean and covariance matrix of $\{\tilde{\Theta}_N{}^{(i)}, i = 1, \ldots, M\}$ at iteration $k-1$.

Due to the stochastic nature of this method [6], a burn-in period of $K$ iterations is defined. Averaged estimators are built to decrease the fluctuations. If we denote $\hat{\Theta}^{(l)}$ and $\hat{x}_n^{(l)}$the estimates of the parameters and the hidden state variables at the $l$-th filtering iteration respectively, then for $l > K$ :

$$\bar{\hat{\Theta}}^{(l)} = \frac{1}{l-K}\sum_{j=K+1}^{l} \hat{\Theta}^{(j)} \text{ and } \bar{\hat{x}}_n^{(l)} = \frac{1}{l-K}\sum_{j=K+1}^{l} \hat{x}_n^{(j)}\ , \tag{2.4}$$

if we use a constant number of particles.

**2.4 Conditional ICPF for Uncertainty Assessment:** In order to estimate the noise parameters and to

evaluate the uncertainty related to the estimates of the ICPF approach, we partition $\Theta_n$: $\Theta_n = (\Theta_1, \Theta_2)$. $\Theta_1$ denotes the parameters from the deterministic part of the model (state equation and measurement equation) and $\Theta_2$ denotes those of the noise model (the parameters of the distributions of $\eta$ and $\xi$ in (2.1)). A conditional ICPF algorithm proposed by [7] is thus implemented.

In the first place, the estimation of the hidden states and of $\Theta_1$ is performed by considering that $\Theta_2$ is known. In practice, small initial variances for the noises seem helpful to ensure the convergence of the algorithm towards satisfactory estimation results for the hidden states and for $\Theta_1$. From this first estimation of $\Theta_1$ and of the hidden states, we can estimate the parameters of the distributions of the modelling and measurement noises $\Theta_2$ from the results. Conditionally to the new estimated $\Theta_2$, the ICPF approach is then carried out again to estimate $\Theta_1$ together with the hidden states. In this way, the algorithm can be iterated until the convergence of both $\Theta_1$ and $\Theta_2$, which is claimed by a standard stopping rule when the relative change in the estimates from three successive iterations is reasonably small [3].

However, since the posterior distributions of the parameters is no longer representative of the estimates' uncertainty because of the regularization effect from the successive iterations of the filtering process, parametric bootstrap [11] is implemented to calculate the related confidence intervals. New observation vectors are hence randomly generated with $\hat{\Theta}$ and the conditional ICPF algorithm estimation is performed for each of them. The uncertainty related to the estimation $\hat{\Theta}$ can thus be evaluated properly. Moreover, we highlight that the algorithmic uncertainty (linked to the stochastic algorithm) can also be assessed by applying the conditional ICPF approach to the same experimental data set a large number of times as presented by [8]. Tests based on simulated data are detailed in [7].

## 3 Application

In this section, we illustrate the application of the two CPF-based methods to a dynamic crop model based on real experimental data of sugar beet. The parameter estimation and data assimilation related problems are addressed. A simple stochastic model, the Log Normal Allocation and Senescence (LNAS) daily crop model is introduced for this purpose.

**3.1 LNAS Model of Plant Growth** The equations are specifically derived for the sugar beet, per unit surface area, with two kinds of organ compartments taken into account: foliage and root system.

**Biomass production:** $Q(t)$ is the biomass production on day $t$ per unit surface area ($g.m^{-2}$) which can be obtained by generalizing the Beer-Lambert law [21]: $\left(1 - e^{-\lambda \cdot Q_g(t)}\right)$ represents the fraction of intercepted radiation, with $\lambda$ ($g^{-1}.m^2$) a parameter and $Q_g(t)$ the total mass of green leaves on day $t$ (in $g.m^{-2}$). The biomass production of the whole plant is then deduced by multiplying the total amount of absorbed photosynthetically active radiation per unit surface area (PAR, in $MJ.m^{-2}$) and an energetic efficiency $\mu$ (in $g \cdot MJ^{-1} \cdot$):

$$Q(t) = \left(\mu \cdot PAR(t) \left(1 - e^{-\lambda Q_g(t)}\right)\right) \cdot (1 + \eta_Q(t)) \tag{3.5}$$

with the modelling noise $\eta_Q \sim \mathcal{N}(0, \sigma_Q^2)$.

**Allocation** for the foliage and root system compartments:

$$Q_f(t+1) = Q_f(t) + \gamma(t) \cdot Q(t) \tag{3.6}$$

$$Q_r(t+1) = Q_r(t) + (1 - \gamma(t)) \cdot Q(t) \tag{3.7}$$

$$\gamma(t) = (\gamma_0 + (\gamma_f - \gamma_0) \cdot G_a(\tau(t))) \cdot (1 + \eta_\gamma(t)) \tag{3.8}$$

with $\tau(t)$ the thermal time, which corresponds to the accumulated daily temperature since emergence day, $G_a$ the cumulative distribution function of a log-normal law parameterized by its median $\mu_a$ and standard deviation $s_a$, and the modelling noise (process noise) denoted by $\eta_\gamma(t) \sim \mathcal{N}(0, \sigma_\gamma^2)$.

**Senescence:** The senescent foliage mass $Q_s$ is a proportion of the accumulated foliage mass given by the cumulative distribution of a log-normal law of median $\mu_s$ and standard deviation $s_s$:

$$Q_s(t) = G_s(\tau(t) - \tau_{sen}) Q_f(t) \tag{3.9}$$

with $\tau_{sen}$ the thermal time at which the senescence process initiates. The green foliage mass $Q_g$ can be hence obtained easily:

$$Q_g(t) = Q_f(t) - Q_s(t) \tag{3.10}$$

**Observations:** The observation variables potentially available from field measurements are:

$$Y(t) = \begin{pmatrix} Q_g(t) \cdot (1 + \epsilon_g(t)) \\ Q_r(t) \cdot (1 + \epsilon_r(t)) \end{pmatrix} \tag{3.11}$$

with measurement noises: $\epsilon_g(t)) \sim \mathcal{N}(0, \sigma_g^2)$, and $\epsilon_r(t) \sim \mathcal{N}(0, \sigma_r^2)$.

**3.2 Experimental Data** The data used for this study were obtained by the French institute for sugar beet research (ITB, Paris, France) in 2006, 2008 and 2010 with slightly different cultivars and in different locations (details of the experimental protocols can

be found in [2] and [20]). For the test case, we choose 2010 as the dataset for calibration since more observation points are available compared to the other two datasets (of course, in real applications, the most aged datasets are supposed to be used for the calibration). Dry matter of root and leaves were collected on 50 plants at 12 different dates: $\mathcal{O}_{2010} = \{54, 68, 76, 83, 98, 104, 110, 118, 125, 132, 145, 160\}$, whereas for the two datasets used for assimilation, the same type of observations were made only at 7 different dates: $\mathcal{O}_{2006} = \{54, 59, 66, 88, 114, 142, 198\}, \mathcal{O}_{2008} = \{39, 60, 67, 75, 88, 122, 158\}$. For each plant, the green foliage mass denoted by $Q_g$ and the root compartment mass denoted by $Q_r$ were measured. The observation vector $Y_n$ is obtained by averaging each data on all the samples and extrapolated at $m^2$ level by multiplying by the observed density.

### 3.3 The Three Steps of the Analysis:

**3.3.1 Parameter Screening by Sensitivity Analysis:** When a model contains a large number of parameters, as it is often the case for plant growth models, parametric estimation from experimental data can be difficult leading to large estimates uncertainty. Therefore, sensitivity analysis is classically applied to select the parameters to be estimated among those identified as the most influential ones while those screened as the least influential ones can be fixed to any values in their domains. In the context of sensitivity analysis, this method is called "screening" or "factor fixing" [5].

With this objective, we use the algorithm proposed by [26] to compute Sobol's indices (first order and total order) of all the functional parameters, choosing as output a generalized least-square criteria.

As indicated by Fig. 1, we screen the parameters $s_a$, $\mu_{sen}$, $s_{sen}$ and fix them to their mean values of the variation interval, as their total order indexes are all below 0.02 . For the five other parameters, their total order effects cannot be neglected and thus should be estimated from experimental data.

**3.3.2 Parameter Estimation:** Based on the sensitivity analysis results, the unknown parameter vector for the deterministic part of the model is $\Theta_1 = (\mu, \lambda, \mu_a, \gamma_0, \gamma_f)$ and the unknown noise parameter vector is $\Theta_2 = (\sigma_Q, \sigma_\gamma, \sigma_g, \sigma_r)$. For the CPF approach, 500000 particles were initialized with non-informative prior distributions, while as for the conditional ICPF approach, 8000 particles were drawn from the same non-informative prior distributions. The conditional ICPF estimation process began with the estimation of $\Theta_1$ given $\Theta_2$, then $\Theta_2$ was estimated empirically based on the estimates of the hidden states. The estimation then proceeded with the new value of $\Theta_2$ and iterated. Finally, 3 repetitions of the conditional version of ICPF were implemented in our test, each of them contains about 200 filtering iterations. A parametric bootstrap was also achieved to evaluate the estimates' uncertainty. Standard deviations and confidence intervals were hence obtained from 200 bootstrap samples. The corresponding results are given in Table 1.

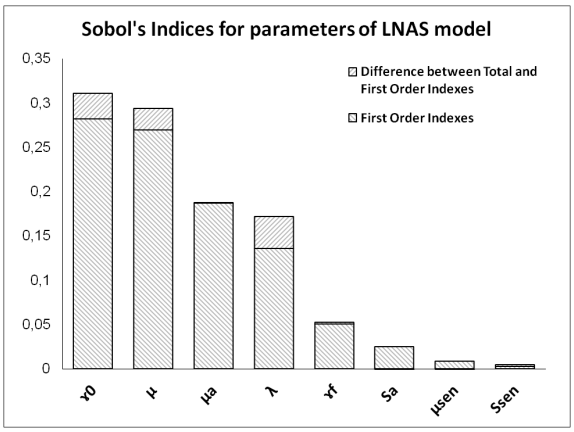



Figure 1: Comparison of the first and total order indexes for $\mu$, $\lambda$, $\gamma_0$, $\gamma_f$, $\mu_a$, $s_a$, $\mu_{sen}$ and $s_{sen}$.

| Parameter | ICPF | | CPF | |
|---|---|---|---|---|
| | Estimates | Std. | Estimates | Std. |
| $\mu$ | 3.55 | 0.16 | 3.50 | 0.01 |
| $\lambda$ | 56.6 | 3.9 | 57.7 | 7.7 |
| $\gamma_0$ | 0.925 | 0.091 | 0.864 | 0.104 |
| $\gamma_f$ | 0.104 | 0.027 | 0.099 | 0.013 |
| $\mu_a$ | 553.9 | 86.5 | 678.7 | 26.1 |
| $\sigma_Q$ | 0.011 | - | 0.021 | - |
| $\sigma_\gamma$ | 0.013 | - | 0.080 | - |
| $\sigma_g$ | 0.098 | - | 0.102 | - |
| $\sigma_r$ | 0.070 | - | 0.072 | - |

Table 1: Estimated values and approximated standard deviations for the ICPF and CPF estimation for 5 functional parameters and 4 noise parameters of LNAS model.

**3.3.3 Data Assimilation with CPF:** In the calibration step, the CPF and the conditional ICPF approaches were applied to the LNAS model in order to estimate jointly the unknown parameters and the hidden state variables based on the 2010 experimental dataset. On one hand, the posterior distribution of the CPF estimation, and on the other hand, the parametric bootstrap results of the ICPF estimation, were used as prior information of the CPF method in the assimilation step (we considered a normal prior). For both the 2006 and

2008 datasets, 500000 particles were simulated, all but the last two measurements were used to update the parameter and state estimates. For the 2006 dataset, after day 114 (resp. day 88 for the 2008 dataset), the propagation of particles through the stochastic dynamic model continued. The simulated values of the state variables $Q_g$ and $Q_r$ on day 142 and 198 (resp. day 122 and 158 for the 2008 dataset) of all the particles as well as their associated weight were used to build the posterior distributions of the prediction.

In order to provide reference values of the prediction without assimilation, Uncertainty Analysis (UA) is also performed. 500000 simulations were initialized in the same way as in the CPF approach, which indicates that samples of $\Theta_1$ were drawn from the distributions defined by the covariance matrix and the mean estimates given by the calibration phase. The independent simulations of these samples in the stochastic dynamic system can thus provide the distribution of the model outputs of interest.

## 4 Results and Discussion

Based on the parameter estimation results from the 2010 data, we compare the predictive capacity of the model for the 2006 and the 2008 experiments, with and without data assimilation, for the two last dates of observations ($t_{142}$ and $t_{198}$ for 2006 and $t_{122}$ and $t_{158}$ for 2008).

Although the three experiments are quite different (different locations, in different years, on slightly different cultivars) the two CPF-based approaches were able to provide fair predictions in most cases and managed to reduce the prediction errors compared to the results obtained with UA (without data assimilation), see Fig. 2. It is particularly spectacular for the leaf mass ($Q_g$) prediction.

Table 2 illustrates the prediction results of the three methods. The average estimates are given with their corresponding standard errors. The prediction results are clearly improved both in terms of mean prediction and confidence intervals for the two CPF-based approaches compared to UA. In fact, as suggested by Fig. 2, the 95% confidence interval provided by the UA can be considered non-reliable since it does not always contain the real measurement values in several cases. Remarks have been made that the confidence intervals proposed by the two CPF-based approaches are narrower than those of UA, which is evident if the assimilation did help to reduce the uncertainty.

Concerning the comparison between the two CPF-based methods, the predictions obtained from the conditional ICPF estimates are closer to the real values compared to the predictions based on the CPF estimates in most cases, except for the root yield prediction $Q_r$ in 2006 (Table 2 and Table 3). Fig. 3 presents a comparison of the posterior distribution of the prediction of $Q_g$ in 2008 obtained after data assimilation based on the estimates obtained with CPF and ICPF. Although

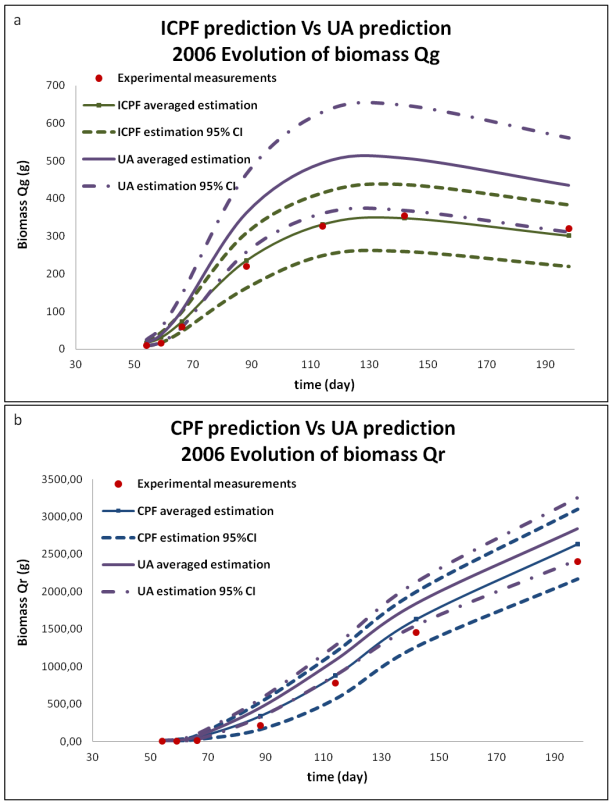



Figure 2: Comparison of the predictions for 2006. a: The predictions of $Q_g$ given by Data Assimilation based on ICPF estimates and the predictions given by UA. b: The predictions of $Q_r$ given by Data Assimilation based on CPF estimates and the predictions given by UA.

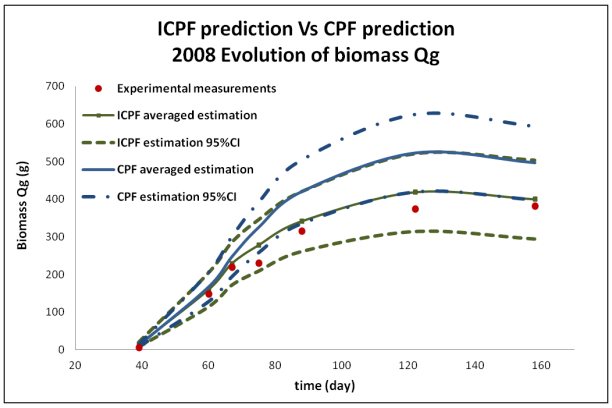



Figure 3: Comparison of the predictions of $Q_g$ in 2008 given by Data Assimilation based on CPF and ICPF estimates.

| | Real Data 2006 | ICPF estimates (relative error in %) | Std. | UA estimates (relative error in %) | Std. | CPF estimates (relative error in %) | Std. |
|---|---|---|---|---|---|---|---|
| $Q_b\,(t_{142})$ | 355.2 | 348.1 (2.01%) | 44.7 | 507.8 (42.96%) | 69.7 | 480.6 (35.30%) | 52.5 |
| $Q_b\,(t_{198})$ | 320.6 | 301.3 (6.04%) | 41.1 | 435.7 (35.90%) | 62.5 | 406.4 (26.76%) | 44.1 |
| $Q_r\,(t_{142})$ | 1459.2 | 1716.2 (17.61%) | 144.1 | 1930.7 (32.31%) | 163.9 | 1630.1 (11.71% ) | 183.8 |
| $Q_r\,(t_{198})$ | 2400.0 | 2644.3 ( 10.18%) | 217.4 | 2942.9 (22.62%) | 243.9 | 2634.9 (9.79%) | 232.7 |

Table 2: Comparison of model prediction with and without data assimilation based on the 2006 dataset.

| | Real Data 2008 | ICPF estimates (relative error in %) | Std. | UA estimates (relative error in %) | Std. | CPF estimates (relative error in %) | Std. |
|---|---|---|---|---|---|---|---|
| $Q_b\,(t_{122})$ | 373.5 | 417.8 (11.86%) | 51.5 | 527.1 (41.14%) | 73.2 | 522.0 (39.76%) | 51.7 |
| $Q_b\,(t_{158})$ | 380.6 | 399.2 (4.88%) | 52.2 | 502.9 (32.12%) | 71.6 | 496.1 (30.34%) | 49.0 |
| $Q_r\,(t_{122})$ | 1559.1 | 1531.1 (1.80%) | 127.8 | 1656.4 (6.24%) | 141.3 | 1476.8 (5.28% ) | 115.5 |
| $Q_r\,(t_{158})$ | 2327.7 | 2192.9 ( 5.79%) | 180.8 | 2352.4 (1.1%) | 196.0 | 2164.9 (6.99%) | 159.5 |

Table 3: Comparison of model prediction with and without data assimilation based on the 2008 dataset.

the predictions based on CPF estimates generally have narrower confidence intervals, they failed to contain all the real observation values as in the case of prediction for 2008 $Q_g$. It is probably related to the fact that the covariance matrix used to initialize the particles in the two cases were quite different. As illustrated by Table 1, ICPF gave more important evaluation of the uncertainties compared to CPF. This result indicates that prior information is crucial to quantify the predictive uncertainty. Since the obtained posterior distributions based on the CPF estimates seem to have underestimated the predictive uncertainty, it implies that the assessment of the parameter uncertainty provided by the CPF approach in the calibration step might not be adequate. Moreover, another possible explanation could be that the level of uncertainty obtained from the estimation in 2010 could not characterize properly the level of uncertainty in 2006 and in 2008 since the experimental conditions are slightly different.

Another interesting point concerns the noise variances. Although the modelling noises are quite low (around 1%), which should imply a good adaptation of the model, we have noticed that the observation noises are rather important (10% and 7%, cf. Table 1). Therefore, the prediction gain with data assimilation remains rather limited since the observations may not be reliable enough to update parameters and state variables. Several tests showed that the results were quite sensitive to the level of the observation noises and these may also vary a lot in different experimental situations, so that their proper evaluation is a key issue which requires further studies.

## 5 Conclusion

Sequential data assimilation techniques offer the possibility to improve model predictive capacities while assessing different sources of uncertainty. However, in the context of plant growth models, traditional approaches are commonly not able to cope with the model complexity combined to the scarcity of experimental data and their general lack of accuracy. This explains why despite the various data assimilation approaches and their proven benefits for uncertainty analysis and model prediction, only few applications to plant growth model and crop yield prediction have been published ([22], [19]).

In this context, among the existing approaches convolution particle filtering appears to be one of the most suitable ones, principally due to two of its features: no approximation needs to be made for nonlinear state-space models and being capable of preserving the variability of updated state variables and parameters while confronting irregular limited data, both of which are desirable when dealing with the plant growth sequential data assimilation problems.

In this paper, we illustrated the implementation of CPF-based methods both for parameter estimation and data assimilation problems while facing real data of sugar beet with the LNAS model. The methods can explicitly account for different sources of uncertainty during the calibration process, and the estimates provided by the traditional CPF method or its iterative version (ICPF) can be used as a priori distributions for the data assimilation step sequentially performed by CPF. State variables and parameters are assimilated to improve model prediction and the predictive uncertainty. Although the experimental conditions of the 3 datasets are quite different (different locations, different years, slightly different cultivars), the CPF-based approaches were able to provide fair predictions after a short period of adjustment in most cases and managed to reduce significantly the prediction errors. The estimates provided by the conditional ICPF estimation method showed better performances during the prediction phase, both in

terms of mean prediction error and robustness of the provided confidence interval. This method can hence be regarded as a potential candidate for yield prediction applications in agriculture, since it appears to be flexible enough to be adapted to various contexts with different types of models and measurements.

## References


[1] M. Arulampalam, S. Maskell, and T. Gordon, N.and Clapp, *A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking*, IEEE Trans. Signal Proces., 50 (2002), pp. 174–188.

[2] C. Baey, A. Didier, S. Li, S. Lemaire, F. Maupas, and P.-H. Cournède, *Evaluation of the predictive capacity of five plant growth models for sugar beet*, in 4th international symposium on Plant Growth and Applications(PMA12), Shanghai, China, IEEE, 2012.

[3] J. G. Booth and J. P. Hobert, *Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm*, Journal of the Royal Statistical Society Series B, Royal Statistical Society, 61 (1999), pp. 265–285.

[4] F. Campillo and V. Rossi, *Convolution Particle Filter for Parameter Estimation in General State-Space Models*, IEEE Transactions in Aerospace and Electronics., 45 (2009), pp. 1063–1072.

[5] F. Campolongo, J. Cariboni, and A. Saltelli, *An effective screening design for sensitivity analysis of large models*, Environmental Modelling and Software, 22 (2007), pp. 1509–1518.

[6] O. Cappé, E. Moulines, and T. Rydén, *Inference in Hidden Markov Models*, 2005.

[7] Y. Chen, B. Bayol, C. Loi, S. Trevezas, and P.-H. Cournède, *Filtrage par noyaux de convolution itératif*, in Actes des 44èmes Journées de Statistique (JdS'12), Brussels, Belgium, 2012.

[8] Y. Chen and P.-H. Cournède, *Assessment of parameter uncertainty in plant growth model identification*, in IEEE 4th International Symposium on Plant growth Modeling, simulation, visualization and their Applications (PMA12), Shanghai, China, 2012.

[9] P.-H. Cournède, Y. Chen, Q. Wu, C. Baey, and B. Bayol, *Development and evaluation of plant growth models: Methodology and implementation in the pygmalion platform*, Submitted, (2013).

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

[11] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall/CRC Monographs on Statistics and Applied Probability, 1994.

[12] G. Evensen, *Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics*, J. Geophys. Res., (1994), pp. 10143–10162.

[13] ——, *Data assimilation: The ensemble Kalman Filter*, Springer, 2006.

[14] N. Gordon, D. Salmond, and A. Smith, *Novel approach to nonlinear/non-gaussian bayesian state estimation*, Proc. Inst. Electr. Eng., Part F 140 (1993), pp. 107–113.

[15] V. Houlès, B. Mary, M. Guérif, D. Makowski, and E. Justes, *Evaluation of the ability of the crop model STICS to recommend nitrogen fertilisation rates according to agro-environmental criteria*, Agronomie, 24 (2004), pp. 339–349.

[16] A. Jazwinski, *Stochastic Processes and Filtering Theory*, Academic Press, New York, 1970.

[17] N. Kantas, A. Doucet, S. S. Singh, and J. M. Maciejowski, *An overview of sequential monte carlo methods for parameter estimation in general state-space models*, in Proceedings of the 15th IFAC Symposium on System Identification, 2009.

[18] G. Kitagawa, *Monte carlo filter and smoother for non-gaussian nonlinear state space models*, Journal of Computational and Graphical Statistics, 5 (1996), pp. 1–25.

[19] M. Launay and M. Guérif, *Assimilating remote sensing data into a crop model to improve predictive performance for spatial applications*, Agric. Ecosyst. Environ., 111 (2005), pp. 321–339.

[20] S. Lemaire, F. Maupas, P.-H. Cournède, and P. de Reffye, *A morphogenetic crop model for sugarbeet (beta vulgaris l.).*, in International Symposium on Crop Modeling and Decision Support: ISCMDS, Nanjing, China, 2008.

[21] J. Monteith, *Climate and the efficiency of crop production in britain*, Proceedings of the Royal Society of London B, 281 (1977), pp. 277–294.

[22] C. Naud, D. Makowski, and M.-H. Jeuffroy, *Is it useful to combine measurements taken during the growing season with a dynamic model to predict the nitrogen status of winter wheat?*, European Journal of Agronomy, 28 (2008), pp. 291 – 300.

[23] N. Oudjane and C. Musso, *Regularized particle schemes applied to the tracking problem*, in International Radar Symposium, Munich, 1998.

[24] R. H. Reichle, P. W. Jeffrey, D. K. Randal, and R. H. Paul, *Extended versus ensemble kalman filtering for land data assimilation*, J.Hydrometeorol., 3(6) (2002), pp. 728–740.

[25] P. van Leeuwen and G. Evensen, *Data assimilation and inverse methods in terms of a probabilistic formulation*, Monthly Weather Review, 124 (1996), pp. 2898–2913.

[26] Q. Wu, P.-H. Cournède, and A. Mathieu, *An efficient computational method for global sensitivity analysis and its application to tree growth modelling*, Reliability Engineering & System Safety, 107 (2012), pp. 35–43.